THEFT ESTIMATION METHODOLOGY: EXECUTIVE SUMMARY

Introduction

Energy theft is a global problem for electricity and natural gas distribution. There have been estimates in the past as to the scale of the problem. However, there is no reliable estimate for energy losses due to theft today in Great Britain (GB). This is making it challenging to justify further investment into detection methods and calls for a reliable estimating method.

RECCo commissioned Capgemini Invent to develop a "Theft Estimation Methodology" (TEM) and, utilising this methodology, calculate the estimated value of GB Energy Theft. This document sets out the methodology and results.

Method

The TEM project looked to develop a model to provide an estimate of the amount of theft there could be across GB. A data gathering exercise was undertaken and a methodology was subsequently defined which made optimal use of the data obtained. A key data source was from the Theft Risk Assessment Service (TRAS) which contained confirmed and suspected thefts. Data was also obtained from other sources including Crimestoppers, Elexon, Xoserve and non-industry data such as that published by the Office of National Statistics (ONS).

Based on the methodology, a model was built which utilised Machine Learning to derive insight from the data obtained. The model was trained and tested on an agreed subset of the population before running it on a sample across the network. The outcome of this classification prediction and averaged losses provided a range of energy theft for both Gas and Electricity.

Estimate

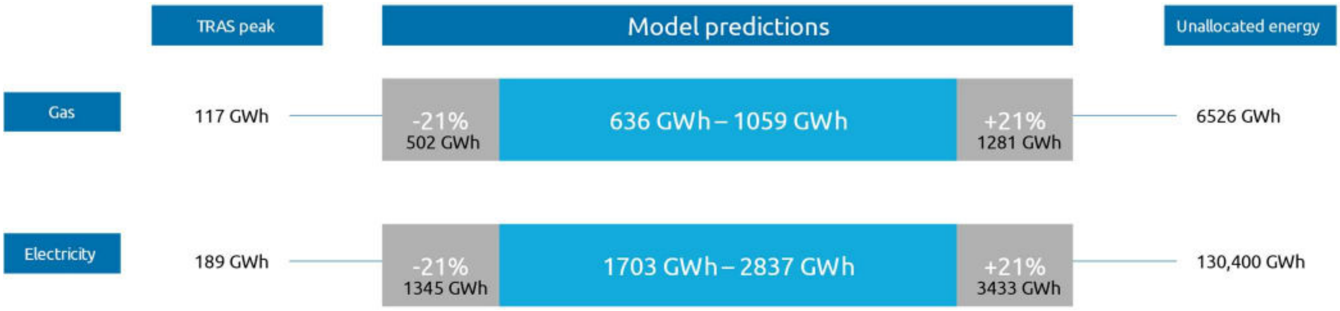
The range of losses due to energy theft derived by the model were as follows:

- Gas losses: 636 GWh to 1059 GWh per year
- Electricity losses: 1703 GWh to 2837 GWh per year

This does need to be treated with caution as it is an estimate developed on best efforts within the data available at this time. The classification model had an accuracy rating of 68% - 93% which when averaged could provide a further range of ±21%.

We have a minimum value from the confirmed theft cases using the TRAS data and we have the upper limit from the comparison to the unallocated energy. The ranges developed are within these parameters, as demonstrated in the following diagram.

Figure 1: Overview of gas and electricity theft loss estimates within the constraints of peak TRAS losses and unallocated energy volumes



To put this into context, in 2019, 323.8 TWh of electricity was generated within the UK[1]. The losses predicted here are just 0.5 – 0.9% of this despite nearly 40% being unallocated. The demand for Gas in 2019 was 859.8 TWh. The losses predicted here make up just 0.1% of this whilst overall losses are around 0.8%.

Despite these losses making up such small proportions of total energy in the UK, taken at OFGEM's capped retail prices as of December 2022, the Theft losses in monetary terms are **£737m - £1233m for electricity and £93m - £155m for gas**. The current costs for these losses end up with the paying consumer. This could potentially add £29 to £48 per year to the average household energy bill.

In terms of carbon equivalent emissions[2], the electricity losses estimated equate to 397,000 $kgCO_2e$ to 661,400 $kgCO_2e$ and carbon equivalent emissions for the gas losses estimated equate to 117,000 $kgCO_2e$ to 194,700 $kgCO_2e$.

The current model is based on the data available and new data is gradually becoming available, as networks and devices increase their instrumentation, that will support a more accurate prediction. Additionally, this could also be used to identify theft in a timelier manner. This includes data from smart meters such as accurate consumption data, being combined with low voltage feeder meter data, to identify mismatches in energy consumption within an area. In addition, the tamper alerts coming from smart meters could be used, in augmentation with other items, to signal theft may be occurring. The ability to act on the data provided by these new data sources would allow the limited resources to be more effectively utilised to reduce theft, increase safety, and ultimately reduce the cost to the paying consumer.

[1] UK Energy in Brief 2021 (publishing.service.gov.uk)
[2] Greenhouse gas reporting: conversion factors 2020 - GOV.UK (www.gov.uk)

THEFT ESTIMATION METHODOLOGY

Contents

1. METHODOLOGY

This section of the report will provide an overview of the methodology to estimate a value of energy theft within GB. It will discuss the data collection process and the challenges within this, the types of analysis worked through and why these choices have been made. It will also highlight the approach to validation of the model and output.

The method that has been developed with considerations to the data that was available, and the aim of the analysis is to both provide some useful insights from the data regarding some anticipated bias as well as to determine useful and appropriate inputs to the development of the model. The selection of the random forest algorithm is anticipated to work best for the types of data being used. The estimation produced is a product of the limitations of the data available at this time. There are expectations that as further data becomes available, the model options could be refined.

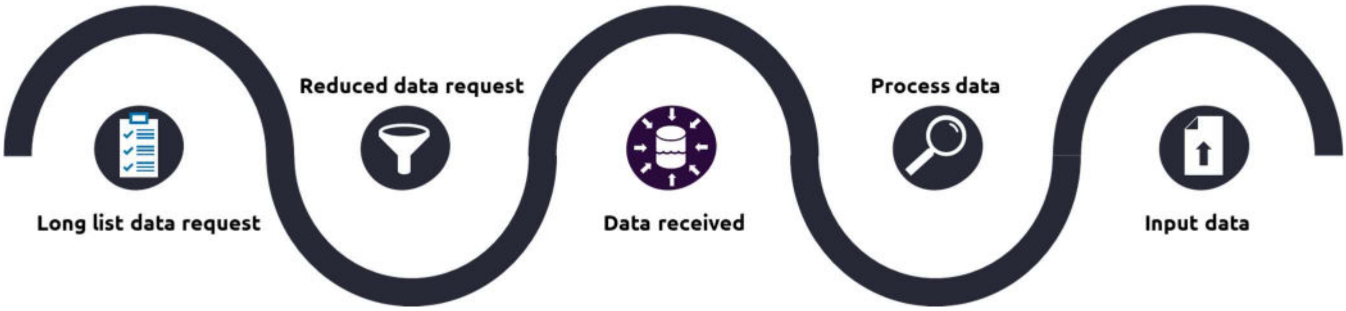
DATA ACQUISITION
ANALYSE
VALIDATION

1.1. QUESTION POSED

RECCo is obligated to provide an estimate of energy theft within the GB to enable decisions to be made on the investment towards detecting theft. This estimate will be provided along with the methodology and the limitations within which it has been reached. Historic attempts to estimate a precise value of theft have been met with varying scepticism, this has, in part, undermined the efforts to reduce energy theft. Therefore, this methodology sets out to produce an approximate theft range (and confidence), rather than a precise value to enable the industry to move forward with appropriate initiatives to improve. This estimate range, is to be backed up by the testing of individual hypothesis' around theories of energy theft to either prove or disprove; "myth busting."

1.2. DATA ACQUISITION

The methodology developed within this paper was dependent on the data that has been made available for the analysis. A long process was followed, from the initial request for data through to collating enough data to enable the project to proceed.



At the outset, an initial data request was made to a variety of data owners. Much of this data was not easily obtainable and the methodology element of the project was put on hold. Table 1 summarises the original request and sources along with the status when the decision was made to pause the development of a methodology. The pause was initiated in Autumn 2021 and the project was not able to be resumed until September 2022 when it was deemed there was a minimum level of data available for a methodology to be developed.

Those with a red status had no provision to share data agreed in principle or commercial discussion progressing. Those with an amber status had an agreement in principle to access the data but a formal agreement was not in place and discussions were ongoing.

Table 1: Initial Data Request

Data Requested	Industry Participant	Status
a. Electricity Settlement data b. Performance Assurance data c. Technical loss calculations	Elexon	🟥
a. Gas Settlement data b. Performance Assurance data c. Technical loss calculations	Xoserve	🟥
Suppliers		
a. Smart meter data & Traditional meter data b. Consumption data or Load information c. Meter & unmetered d. Customer data e. Account data f. Theft Investigations data g. Property data h. Reports or summary	Energy Suppliers	🟨
Distribution Network Operators/Gas Transporters		
a. Substation or feeder level data b. Confirmed and Suspected Theft in Conveyance c. Supplier data - Smart & Traditional meters d. Reports or aggregated information around theft	UKPN	🟨
	Northern Power Grid	🟨
	Cadent Gas	🟥

Property data	Energy Savings Trust, Ordnance Survey	🟥
TRAS data a. Supply data b. Account data c. Consumption data d. Meter data e. Outcome data f. Property data	Experian	🟧
ETTOS data	Crimestoppers	🟧
Cannabis market trend data, Bitcoin	Police Publications, Home Office	🟧
Theft detection methods prior studies, Theft calculator manuals	RECCo	🟧
Water Utilities	Water Utilities	
ECOES	RECCo	

Having limited success with the initial data request, a reduced list of data sources was drafted as shown in Table 2. This data was successfully obtained, enabling a revised methodology to be explored based on these attainable sources. Where necessary, this data was requested from the data supplier and subsequently provided with the required agreements put in place, particularly to abide by GDPR regulations. Data provided by Elexon, National Grid, Xoserve and the Office for National Statistics (ONS) were open-source datasets downloaded from each respective website.

Table 2: Shortened Data Request

Source	Data	Status
Experian	Theft Risk Assessment Service (TRAS Data)	🟩
Xoserve	• Reconciliation by month • Theft of Gas Reporting	🟩
National Grid Data	• Gas Total Shrinkage • Gas Total Assessed • Demand Total • Temperature	🟩
REC Performance Assurance Team	• Confirmed Thefts Post TRAS	🟩
Elexon Data	• GSP Group Take Corrected 3ySFRF • ADR components • GSP Aggregated Metered Volume • P315 + P0276 + P0277	🟩
Crimestoppers	• Reports of Energy Theft	🟩
ONS (see Chapter 13 for source references)	• Deprivation data • Crime data • Fuel poverty • Housing • Rural/Urban • Population density	🟩

	- Spatial data	🟩

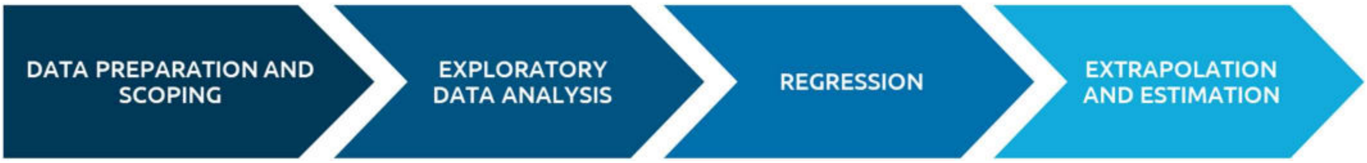
Additional data sources were identified during discussions on the hypothesis and their feasibility was explored. This included:

- Published hygiene data for food outlets, which although only for a subset, aligned with a hypothesis
- Business classification for all commercial properties, available publicly for single look ups, it is more difficult to acquire for the entire population.

1.3. ANALYSE

An exploratory approach was taken to developing the model as it was not known if the data available would be suitable. To determine the right model to build, a series of analyses took place to understand the data better and to identify the key inputs to the model.

In practice, the steps worked through are outlined below:


DATA PREPARATION AND SCOPING

EXPLORATORY DATA ANALYSIS

REGRESSION

EXTRAPOLATION AND ESTIMATION

DATA PREPARATION AND SCOPING

Several disparate datasets were identified to be used for the estimation, so these were to be joined together. A conceptual data model was developed to map these connections and pipelines were built within the database to join these together. A core part of this was joining TRAS data to additional data sets by post code/LSOA/LDZ/GSP to add deprivation, crime, and housing data.

Summary statistics of the datasets were run to provide an understanding of the size and quality of the data and ensure they could be used for the intended purposes. Invalid and missing values were identified, and steps taken to extrapolate where suitable.

EXPLORATORY DATA ANALYSIS (EDA)

Exploratory analysis including visualising the data was used to review theft investigation outcomes with potential predictors in the data, looking for correlations and other interesting insights. Hypothesis testing was also used to understand potential bias in the data. Hypotheses were developed following discussions with the theft SME's regarding the anecdotal behaviours of those who steal energy and the investigations that are carried out. The data needed to validate the hypotheses was identified and visualisation techniques have been utilised for many of the tests. The outcomes were played back and discussed with the SME's.

Spatial analysis was used to explore impact of location, supplier, DNO/ Gas network over time for the theft identification and by source. Identification of 'reliable regions' which were areas that were deemed to have investigation data which was 'good' to train and test the model.

PREDICTIVE MODELLING

A predictive model is needed to produce an estimate of theft based on patterns and trends in the existing theft data. This model can take many forms, and so the first step was to identify an appropriate model to implement. A scenario construct was identified to build fitting models for the multiple groups within the data. A classification model was built against the scenarios and trained and tested using the ‘reliable regions’ identified, followed by an iterative process to optimise the model performance. The performance metrics of the models were then scrutinised to determine suitability to select the best method for predictions and the impact of the variables on theft analysed.

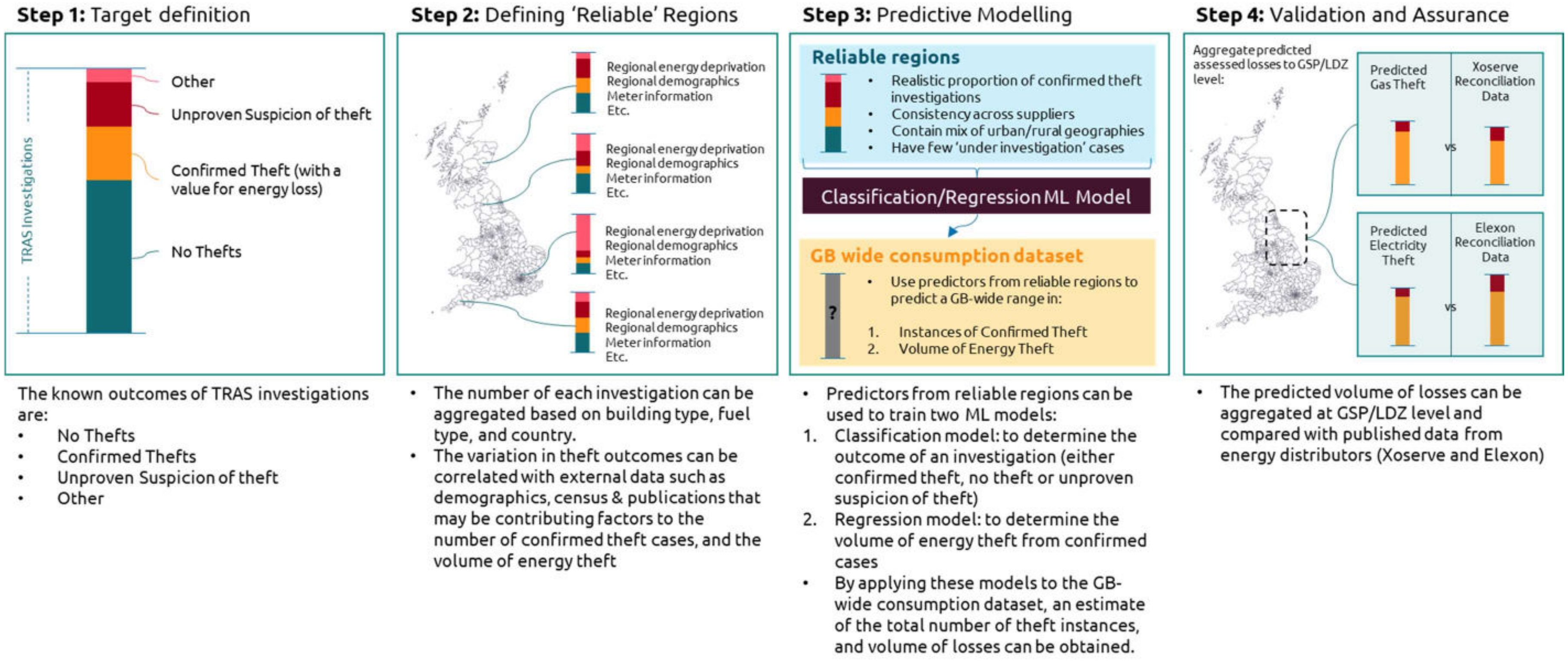
EXTRAPOLATION AND ESTIMATION

A population weighted sample of the TRAS consumption dataset was taken to run the classification to identify instances of theft and regression model to provide an assessed loss value. These were then extrapolated across the full network and converted to revenue values to provide the estimated ranges of theft.

1.3.1. HIGH LEVEL APPROACH

Figure 2 below shows the revised methodology developed based on the available data sources. For reference, the initial methodology is provided in Appendix B (Chapter 9) and an explanation of the data limitations preventing this method from being undertaken are outlined in section 1.5.1.

Figure 2 Refined theft estimation methodology



Step 1: Target definition
TRAS Investigations
Other
Unproven Suspicion of theft
Confirmed Theft (with a value for energy loss)
No Thefts
Step 2: Defining 'Reliable' Regions
Regional energy deprivation
Regional demographics
Meter information
Etc.
Regional energy deprivation
Regional demographics
Meter information
Etc.
Regional energy deprivation
Regional demographics
Meter information
Etc.
Regional energy deprivation
Regional demographics
Meter information
Etc.
Step 3: Predictive Modelling
Reliable regions
Realistic proportion of confirmed theft investigations
Consistency across suppliers
Contain mix of urban/rural geographies
Have few 'under investigation' cases
Classification/Regression ML Model
GB wide consumption dataset
?
Use predictors from reliable regions to predict a GB-wide range in:
1. Instances of Confirmed Theft
2. Volume of Energy Theft
Step 4: Validation and Assurance
Aggregate predicted assessed losses to GSP/LDZ level:
Predicted Gas Theft
vs
Xoserve Reconciliation Data
Predicted Electricity Theft
vs
Elexon Reconciliation Data

The known outcomes of TRAS investigations are:

- No Thefts
- Confirmed Thefts
- Unproven Suspicion of theft
- Other

- The number of each investigation can be aggregated based on building type, fuel type, and country.
- The variation in theft outcomes can be correlated with external data such as demographics, census & publications that may be contributing factors to the number of confirmed theft cases, and the volume of energy theft

- Predictors from reliable regions can be used to train two ML models:

1. Classification model: to determine the outcome of an investigation (either confirmed theft, no theft or unproven suspicion of theft)
2. Regression model: to determine the volume of energy theft from confirmed cases

- By applying these models to the GB-wide consumption dataset, an estimate of the total number of theft instances, and volume of losses can be obtained.

- The predicted volume of losses can be aggregated at GSP/LDZ level and compared with published data from energy distributors (Xoserve and Elexon)

1.4. VALIDATION

The methodology produced has been a result of the time constraints of when an estimation is required by and also the data availability and quality. As with any methodology, it is prudent to ensure it is fit for purpose. This has been achieved through validation of hypothesis testing with RECCo's Theft SMEs, the statistical testing of the model, and a review of the draft methodology ahead of publication.

1.5. DESIGN DECISIONS

1.5.1. DATA DRIVEN

It is acknowledged, there were several changes to the original methodology outlined in Appendix B (Chapter 9). The changes were required due to the lack of key data that would have fed into the total energy balance equation.

ENERGY BILLED

The supply of energy, particularly for gas was only available at LDZ level which was at a more aggregate level than initially anticipated. This meant there would not be the confidence in being able to map volumes of confirmed theft fully to that regional input to explain unidentified energy levels. Electricity was available at GSP level which was also too high to be confident in the output. Rather than use this at a low level, the data surrounding the unallocated energy will be used to validate the final output, i.e. the estimate theft volumes.

TECHNICAL LOSSES

It was assumed there would be data coverage of technical losses across the electricity network. On initiation of the TEM project, it was confirmed this data was not in existence within the data available to us and so an 'expected' level of technical losses would need to be relied upon. This again reduced the confidence of addressing the 'unidentified energy gap' with the 'confirmed theft' volumes at a more granular level.

1.5.2. CHOICE OF PREDICTIVE MODEL

A number of possible Machine Learning (ML) algorithms were considered for this analysis, and it was decided to use a Random Forest algorithm to balance accuracy with computational efficiency.

At a high level, a Random Forest runs multiple decision trees in parallel, each using a random subset of the training data (the data which is being used to train the model). Each individual decision tree generates an outcome, and the average of these is taken to produce the final prediction. A Random Forest algorithm can take a classification or regression form and, in the context of this project, can be used to estimate:

i) The outcome of a theft investigation
ii) The volume of theft stolen

A more thorough explanation of Random Forests is given in Chapter 4.4. Random forests were chosen over other methods because they:

- Use ensemble learning (the concept of using the average outcome of multiple decision trees) which performs better that using a single logit/linear regression model,
- Are easier to tune and less prone to overfitting compared with XGBoost, and
- Enable the relative significance of predictors to be extracted which is not available with Neural Networks

1.5.3. LIMITATIONS

The methodology being taken forward for testing has been developed in line with the data available and limitations of that data along with the recognition that a finite time is available to develop the model. It is also recognised there are additional data sources which RECCo do not currently have access to which could provide further opportunities to refine an estimation model.

Due to the short timeframe the theft data has been available for, there is limited time series analysis that can be performed. The short duration of available data has also been impacted by considerable number of ‘one off’ events, such as COVID restrictions, which prevent useful insight being generated. This creates a further dimension which could be explored in the future as a more stable data set is developed.

1.6. METHOD SUMMARY

To develop a model to estimate the energy theft, it has been determined that Machine Learning (ML) will be able to support deriving the estimate. ML is a good way to derive a model given the data available. From it we can infer estimates that will be as good as the data will allow. In addition, the estimates can be combined with additional data parameters such as the actual confirmed theft and unallocated energy after settlement to validate the estimate that has been produced.

Data quality issues were expected, particularly due to the lack of validation at the outset of TRAS and therefore validation checks were carried out to identify issues. With this knowledge, the methodology was then adapted to accommodate the data that was available for use within the ML algorithm. Hypothesis testing was also used to both, understand the data better and to identify the variables that needed to be included to build the algorithm.

Data provided from TRAS was used as a basis to build out an enriched dataset for use with the model. Known issues were ‘cleaned’ from the data and it was accepted that this would not be a true view of the total population as there are known to be unregistered assets drawing power from the networks, for example, when single properties convert to multiple properties without registering for new MPAN/MPRN. Working purely at an MPAN/MPRN level meant that other energy theft such as conveyancing would be a gap in the estimates.

A two-step model was explored, with the first to classify if theft was ‘confirmed’, ‘unproven’ or ‘no theft’, and then a regression to quantify the assessed losses if it was deemed to be ‘confirmed’. The classification model worked to a reasonable level however, the regression prediction was no better than taking an average, so that route was taken.

The exploration of the method was a timeboxed exercise and the estimation produced takes into consideration the limitations of the data available and its granularity.

2. DATA OVERVIEW

This chapter provides details on the setup of the data ahead of the analysis and how the various data sources were brought together. It also contains a summary of how the data was used and issues identified and how these were dealt with to enable the analysis to continue.

The key datasets forming the base of the datasets to be used with the machine learning algorithm are the TRAS consumption data and the TRAS Theft data. The inclusion of the RPA data within algorithm was ruled out due to the corrupted data within the file received.

There are still challenges with the TRAS data and considerations have needed to be made in how the data is used to enable the appropriate datasets to be formed ahead for use with the machine learning algorithm. To build an enriched dataset, additional sources of data were matched with the consumption data and the theft data using location data including common spatial data, particularly LSOAs. In addition, there were data sources which were only available to LDZ or GSP level including the Xoserve, National Grid and Elexon data which was being used to compare unallocated energy to the theft estimations.

2.1. DATA COLLECTION

When the Theft estimation work was initiated, requests for information were made to a variety of contacts as listed in Table 1: Initial Data Request. Although the sharing of data was agreed in principle, the formal arrangements for this to proceed were not forthcoming and the project was paused until arrangements to access the data were put in place. Over the proceeding months, access to a selection of data from a revised list (Table 3) was achieved. A project was initiated to set up an Azure database on behalf of RECCo to securely store the data sets and enable the theft estimation project to be resumed.

Table 3: Data source summary list and corresponding object

Source	Data	Object
Experian	Theft Risk Assessment Service (TRAS Data)	Theft Investigations TRAS Consumption Data Meter Points Accounts
Xoserve	• Reconciliation by month • Theft of Gas Reporting	Unallocated Energy
National Grid Data	• Gas Total Shrinkage • Gas Total Assessed • Demand Total • Temperature	Unallocated Energy
REC Performance Assurance Team	• Confirmed Thefts Post TRAS	Theft Investigations
Elexon Data	• GSP Group Take Corrected 3ySFRF • ADR components • GSP Aggregated Metered Volume	Unallocated Energy

	- P315 + P0276 + P0277	🟦
Crimestoppers	- Reports of Energy Theft	Theft Investigations
ONS	- Deprivation data - Crime data - Fuel poverty - Housing - Rural/Urban - Population density - Spatial data	ONS Predictors

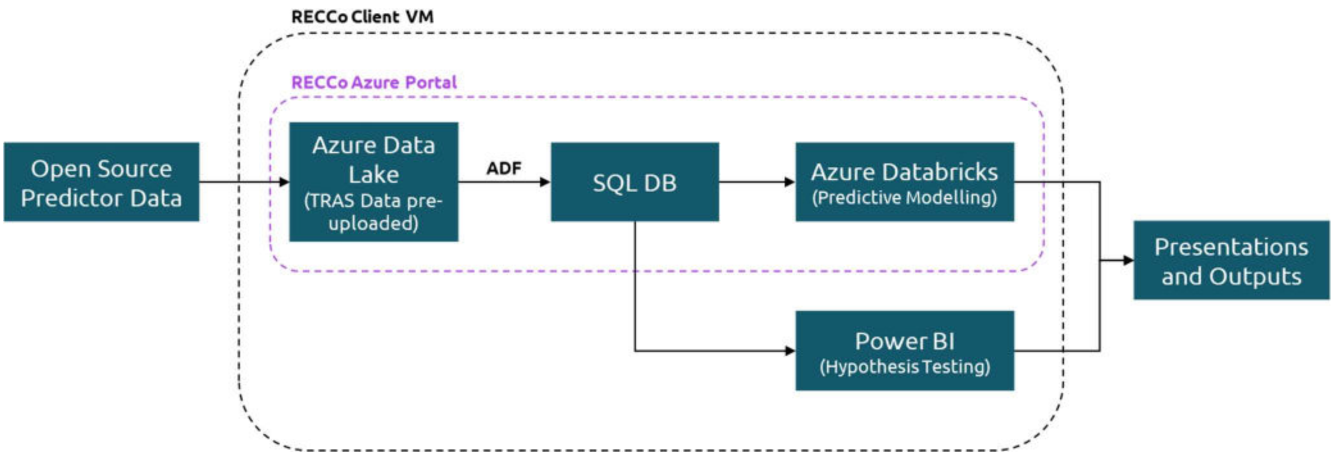
There was a disclaimer on resumption of the project that the data quality could not be ensured. For example, TRAS data was shared in its unvalidated format and so was the quality it had originally been received from the supplier. This was similar for the RPA data. Officially published sets were better quality but some required discussion with the data providers to ensure accurate interpretation of the data-sets.

2.2. DATA SET UP

The data was assessed and uploaded to the data lake before being fed through into usable tables. The technical detail of how the data was transformed has been provided to RECCo separately in 'TEM Data Overview'. The high-level data flows are provided below. There were some challenges around the quality of the data and a high-level overview of these has been provided.

The overall data handling pipeline is outlined in Figure 3 below.

Figure 3: Data handling pipeline


RECCo Client VM
RECCo Azure Portal
Open Source Predictor Data
Azure Data Lake (TRAS Data pre-uploaded)
ADF
SQL DB
Azure Databricks (Predictive Modelling)
Power BI (Hypothesis Testing)
Presentations and Outputs

The data handling processes were carried out via a RECCo Virtual machine to ensure that GDPR and Data Privacy protocols were adhered to, given the sensitive nature of the personal data. The majority of data platforms utilised within the project were hosted via the Azure Portal, specifically Azure Data lake, SQL Database and Azure Databricks. In addition, Power BI was in-stalled on the virtual machine to help derive data insights as part of the hypothesis testing process.

Access to the data via the Azure Portal was administered by RECCo, with access managed via Azure Active Directories (AAD). In addition to access being limited via ADD, it was also limited to individuals with whitelisted IP addresses.

Raw open data sources as identified above as our predictor data were uploaded to the Azure Data Lake in a variety of formats, and Azure Data Factory (ADF) was used to extract, transform and load this data, along with the pre-uploaded TRAS data, into the SQL Server Database so that it could be queried. Azure Databricks and Power BI connected directly to the SQL Database and imported data directly to perform the respective modelling and hypothesis testing tasks. Outputs from these processes were then downloaded and saved to be shared with the wider RECCo team and stakeholders.

2.3. DATA USE

There are 2 key data sets from which the predictor data is joined and then used to create the regression and model datasets. These are the TRAS consumption data and the TRAS Theft data. The data was then enriched with public data for their location such as deprivation data, fuel poverty and crime rates. Here is an overview of how the data from within the datasets are used.

TRAS CONSUMPTION DATA

TRAS consumption data is in the following tables in the raw data schema

- Commercial consumption: `raw_data.tras_comm_cons_data` (50 million rows)
- Residential consumption: `warehouse.RESI_CONS_DATA` (130 million rows)

This dataset serves two purposes:

- For obtaining metadata for meters and supply accounts that are involved in theft investigation. This metadata includes meter type, meter location and normal payment method, which are used as predictors in the ML model.
- For creating a list of meter points to use as the prediction dataset for the ML.

A detailed breakdown for the considerations taken to extract this data can be found in Appendix A (Chapter 8).

TRAS THEFT DATA

TRAS theft data was uploaded to the following tables in the raw data schema

- Commercial theft: `raw_data.tras_comm_tfto_data` (150,000 rows)
- Residential theft: `raw_data.tras_resi_tfto_data` (900,000 rows)

These tables are used to identify the instances of theft investigation. A detailed breakdown for the considerations taken to extract this data can also be found in Appendix A (Chapter 8).

XOSERVE DATA:

Monthly reconciliation data allocated against the month of gas used with the intention of using this for comparison of theft estimations against unallocated gas.

NATIONAL GRID DATA:

Shrinkage data had been expected to be used to define the unallocated energy however, it was not available at a granular level to build this up.

REC PERFORMACE ASSURANCE DATA:

Theft submissions from energy suppliers as part of the incentive schemes since September 2021 were intended to be used to train the model however, due to data quality, this data was only used for exploratory analysis.

ELEXON DATA:

At the project outset, it was intended to use Elexon data to identify unallocated electricity volumes at granular regional levels, however this was not possible with the data provided. After revising the methodology, the electricity data required to validate theft predictions at a high level was possible with the National Grid data. Therefore, the Elexon data was not required for this project.

ONS PREDICTOR DATA:

Officially published statistics which could be used as predictors of energy theft were identified and used to build an enriched data set along with the TRAS theft and consumption data. This enriched dataset was then used for the hypothesis testing as well as within the machine learning algorithm.

2.4. DATA CHALLENGES

It was anticipated the datasets would have data issues due to being collated from multiple suppliers. Quality checks were performed on the datasets and some further issues were also uncovered during the joining of data. This summary is intended to provide some insight into the nature of the data that was being handled and the limitations which come along with it.

TRAS DATA FROM EXPERIAN

It is recognised there was no validation in the early years of TRAS and therefore the data did contain errors such as theft end dates occurring before start dates. In addition, there were genuine behaviours causing data to look erroneous when it could be correct, such as the same meter could be removed from one property and installed elsewhere. Finds from the quality checks were discussed with theft SMEs to uncover if there were genuine data issues or behaviours which could explain the data.

Genuine poor data quality, some likely to be attributed to lack of validation checks during early submissions to TRAS:

Gaps in required fields within TRAS consumption data

- Approx. 1 in 6 meter type was missing
- Approx. 1 in 5 meter location was missing
- These records with missing fields were included in the dataset and their missing data field categorised as “unknown”

Challenges within TRAS Theft data

- Duplications of supplier ID across multiple dates within assessed loss field
- Cases of no theft with assessed losses
- Duplications of supplier investigation ID within which should be unique against different MPAN/MPRN

Any future collection process should validate the data on receipt to prevent a poor-quality dataset developing.

Explainable data issues:

- Theft durations with a cut off at 1 or 2 years, could be dependent on supplier policy – potentially leading to underestimation of volumes.
- MPAN/MPRN with multiple postcodes, could be genuine due to changes of postcodes for new developments however there could also be data quality issues with incorrect postcode or MPAN/MPRN entered. Each combination of MPAN/MPRN, supply postcode was used to identify a new number as excluding them risked excluding genuine meter points. This increased the meter points from the expected 57 million to 60 million.
- Meter Serial numbers with multiple Meter ID's, could be genuine as multiple manufacturers could issue the same serial number
- MPAN's not aligning to the 13 digit format within TRAS, MPANs can range from 6- 13 digits and there were a negligible volume with digit formats shorter than this.
- Cases of confirmed theft where no assess losses were provided, particular in gas this could be attributed to a fiscal theft. A bespoke tamper code was later provided to capture this type of theft.

RPA DATA

It was recognised the quality of the theft investigation data submitted to the REC was poor and guidance had been provided to suppliers to improve this. Due to the data issues stated below, the data was not compiled with the TRAS data for building the model, however it was able to be manipulated separately to the model to provide a brief overview of the data contained. This can be found in Chapter 3.

- The data from a particular supplier has been excluded due the unusable format.
- Corrupted format of data as MPAN/MPRN converted to scientific format which is a common issue when working with this data. Validation checks would prevent this data being received in future. TRAS consumption data was able to be used to match some of this data based on other fields in the data.
- Supplier Investigation ID was missing in 10% of cases.

XOSERVE DATA:

- Data containing no units making it difficult to interpret out of context.

NATIONAL GRID DATA:

- Data unavailable at GSP level as expected.

ONS PREDICTOR DATA:

- Inconsistent levels of granularity within the datasets. For example, crime data only available at Police Force Area level whilst Deprivation data available at LSOA level.
- Inconsistent data formats between England, Scotland and Wales meaning datasets could not be combined.

2.5. DATA SUMMARY

The data was provided by a range of third parties with differing levels of data quality. TRAS data was provided in an unprocessed format from Experian. Even within this dataset there were differing levels of quality since basic validation checks were not introduced until the latter years of TRAS being in operation. This has led to erroneous/invalid data within the system, particularly within the earlier years.

Some of the examples of unexpected data included:

- MPAN/MPRNs appearing in multiple postcodes
- Multiple meters attached to the same MPAN/MPRN
- The same meter serial number attached to multiple meters

Some of the work to validate this data included exploratory analysis and verification with RECCo's Subject Matter Experts (SMEs) to understand if there were circumstances where these could be genuine data inputs. This enabled rules to be built to extract valid data, a summary of which can be found in Appendix A.

The RPA (post-TRAS) data provided under TDIS, which contains raw investigation data compiled from multiple suppliers, was also received in an unprocessed format. The decision was taken not to include the RPA data within the model due to the data challenges however, it was used in the exploratory analysis to provide a wider picture of known theft.

For each investigation there could be multiple entries of updates however many of the additional columns were blank, but the updates appeared multiple times so the dataset had any 'old' updates removed.

Additional predictor information was used to enrich the dataset. This was mainly official national statistics joined at an LSOA level including deprivation, fuel poverty and other crime data It was used in both the hypothesis testing and within the machine learning algorithm.

In general, there were numerous errors within the raw data which were challenging to resolve. Where possible the data was cleaned to either remove erroneous data or the data was used in the most logical way possible.

There are some known gaps in the method. As TRAS has been utilised as the base of the model dataset, only MPAN/MPRN from their consumption files provided to them by suppliers have been used in the estimation. There are known to be unregistered assets, for example, when single properties convert to multiple properties. Theft in conveyancing is another gap.

3. EXPLORATORY DATA ANALYSIS

The purpose of this section is to provide a summary of the exploratory analysis undertaken to assess the data. Data was visualised interactively through a Power BI dashboard and discussed thoroughly with SMEs to ensure that the data was cleaned and interpreted correctly. It will cover the investigation outcome as well as the duration and volume of confirmed thefts.

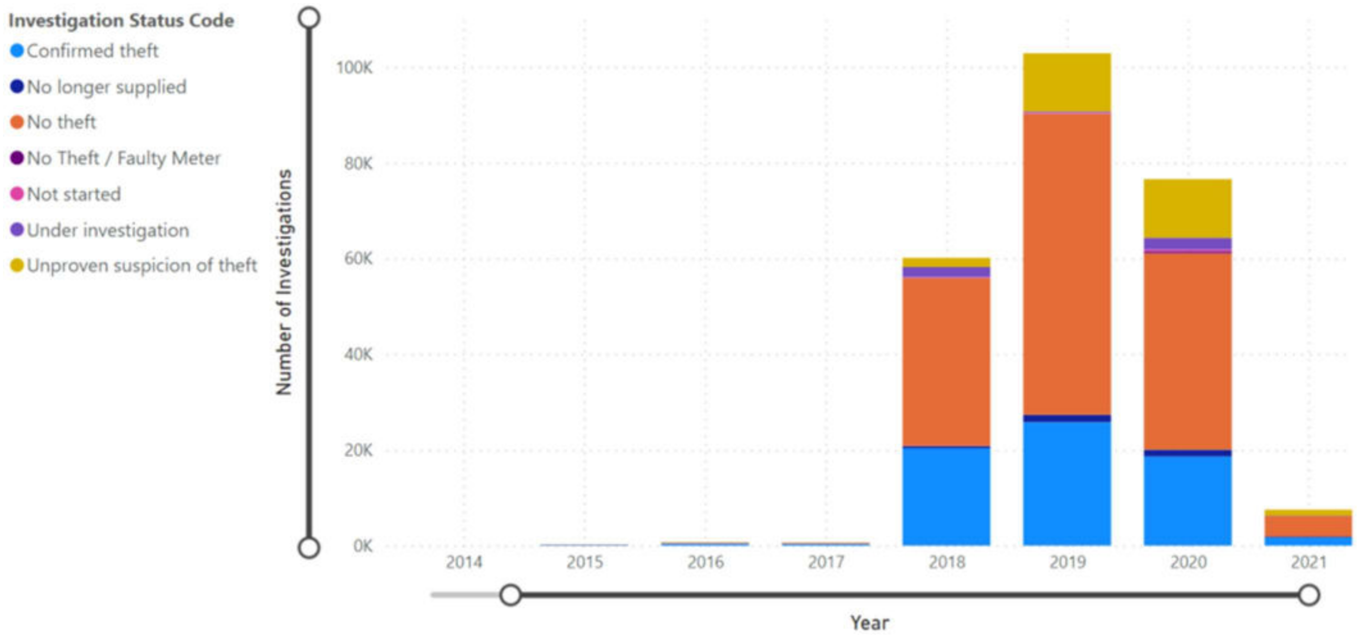
There are two core data sets containing the outcome of theft investigations. These are the TRAS theft data which captures investigation outcomes from 2014 until March 2021, and the RPA data collected post TRAS under TDIS which has collected theft investigation data since September 2021. Suppliers are incentivised to supply their data to TRAS and the RPA to receive a monetary compensation for carrying out the investigation. Incentives are received for confirmed thefts however, suppliers are requested to supply all investigation outcomes. This section will look at the overview of the volumes of investigation by outcome over time for both TRAS and RPA. It will then highlight some of the findings for the duration of theft and the volumes of energy lost due to theft from the TRAS data.

3.1. EXPLORATORY ANALYSIS FINDINGS

TRAS theft data overview

Figure 4 shows the total number of theft investigations (with valid Investigation ID’s) recorded in TRAS, split by the date in which the investigation closed. In total there are approximately 300k investigations, the majority of which were closed between 2018 and 2020. When interpreting this figure, it is important to note that the date in which an investigation closed does not necessarily mean that the full volume of theft associated with that investigation occurred within that year. Average theft durations were found to be just under 2 years, and so the distribution in the actual annual volume of energy theft would more likely shift towards the earlier years.

Figure 4: Number of theft investigations by outcome, split by investigation close date



Below summarizes the key findings from this analysis and outcomes of the discussions with SMEs:

- There were theft investigations with closure dates prior to 2013 and post 2022. Based on SME advice we have omitted these investigations from our analysis since they are likely to be entry errors.
- The number of closed investigations peak in 2018/19 and then drop off in 2020, likely due to COVID and the closure of TRAS. This unsteady timeframe has prevented any accurate time series analysis being performed.
- Overall, 23% of investigations resulted in confirmed theft, 50% resulted in No Theft, 10% were unproven suspicion of theft and 16% were under investigation.
- Unproven suspicions of theft can be considered as confirmed thefts, based on SME advice.
- Cases Under Investigation within TRAS are likely to be picked up in the RPA data.

RPA POST-TRAS DATA OVERVIEW

Figure 5 shows the total number of theft investigations (with valid Investigation ID's) recorded in the RPA post-TRAS data, split by the date in which the investigation closed. In total there are approximately 54k investigations and as expected, the majority of these were closed in 2021 and 2022 after TRAS ceased. The volumes represent partial years for data collection as some outcomes would have been reported to TRAS in early 2021. 2022 is not yet complete, it is also known that some suppliers submit their outcomes in time for the incentive to be paid which is once a year but not on a regular basis. Even with these caveats, from the data collected post TRAS, it could be concluded the number of investigations have decreased since the peak in 2019.

The RPA data was summarised during the exploratory analysis, however it was found to have issues with the corruption within the data. The formatting of the MPAN/MPRN was corrupted, meaning the true MPAN/MPRN could not be determined which meant it could not be used to map through to other data sets such as the TRAS consumption data to enrich the dataset. There were also missing Supplier investigation IDs in about 10% of the entries meaning they could not be clearly linked back to the original entries of the TRAS theft dataset. This is why although the summary of this data is included, the dataset was not taken forward to be used within the prediction modelling.

Figure 5: Number of investigations in the RPA post-TRAS data split by investigation close date and outcome

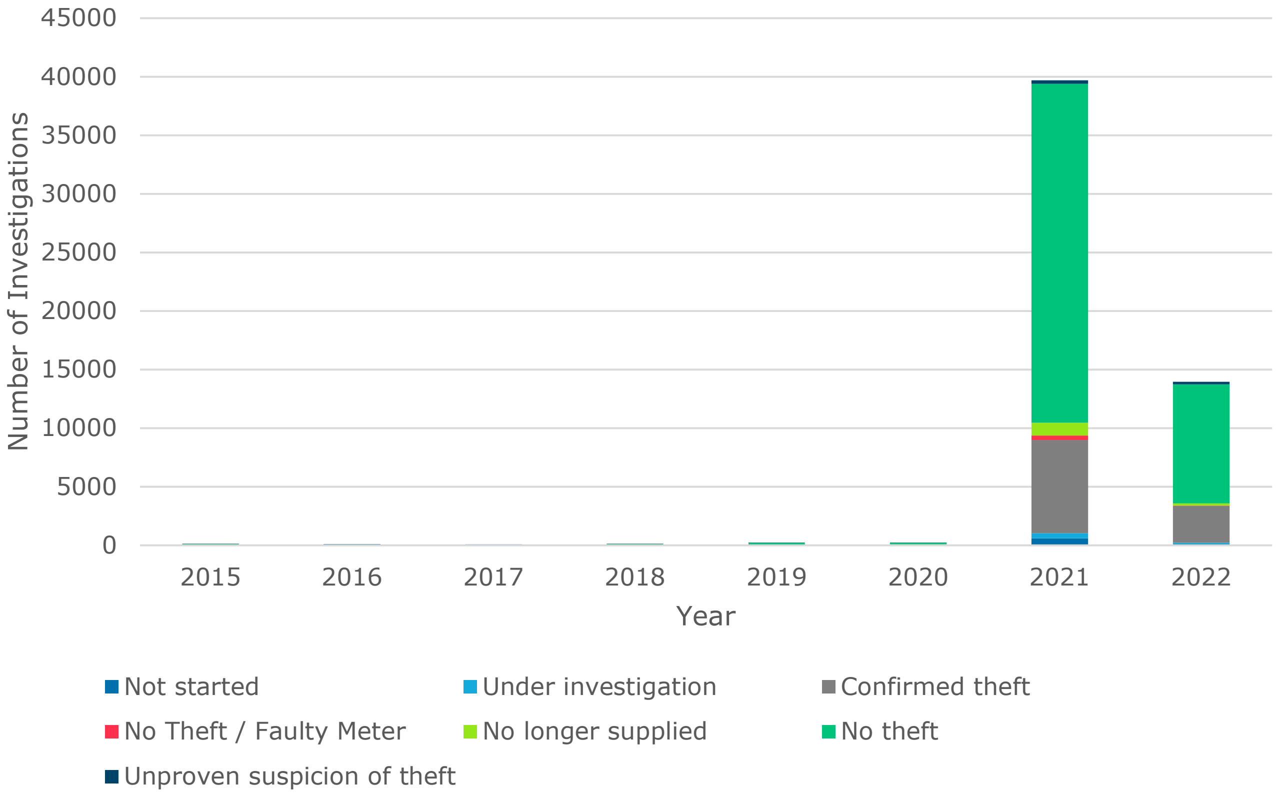


Table 4 summarises the proportion of investigations per outcome code in the RPA data. It was expected that a higher proportion of investigations would result in confirmed theft compared with the TRAS data, since there was less incentive for suppliers to report no thefts. However, the data shows that in both 2021 and 2022, the conversion rate of investigations to confirmed thefts remained at around 20%, which aligns with the TRAS data.

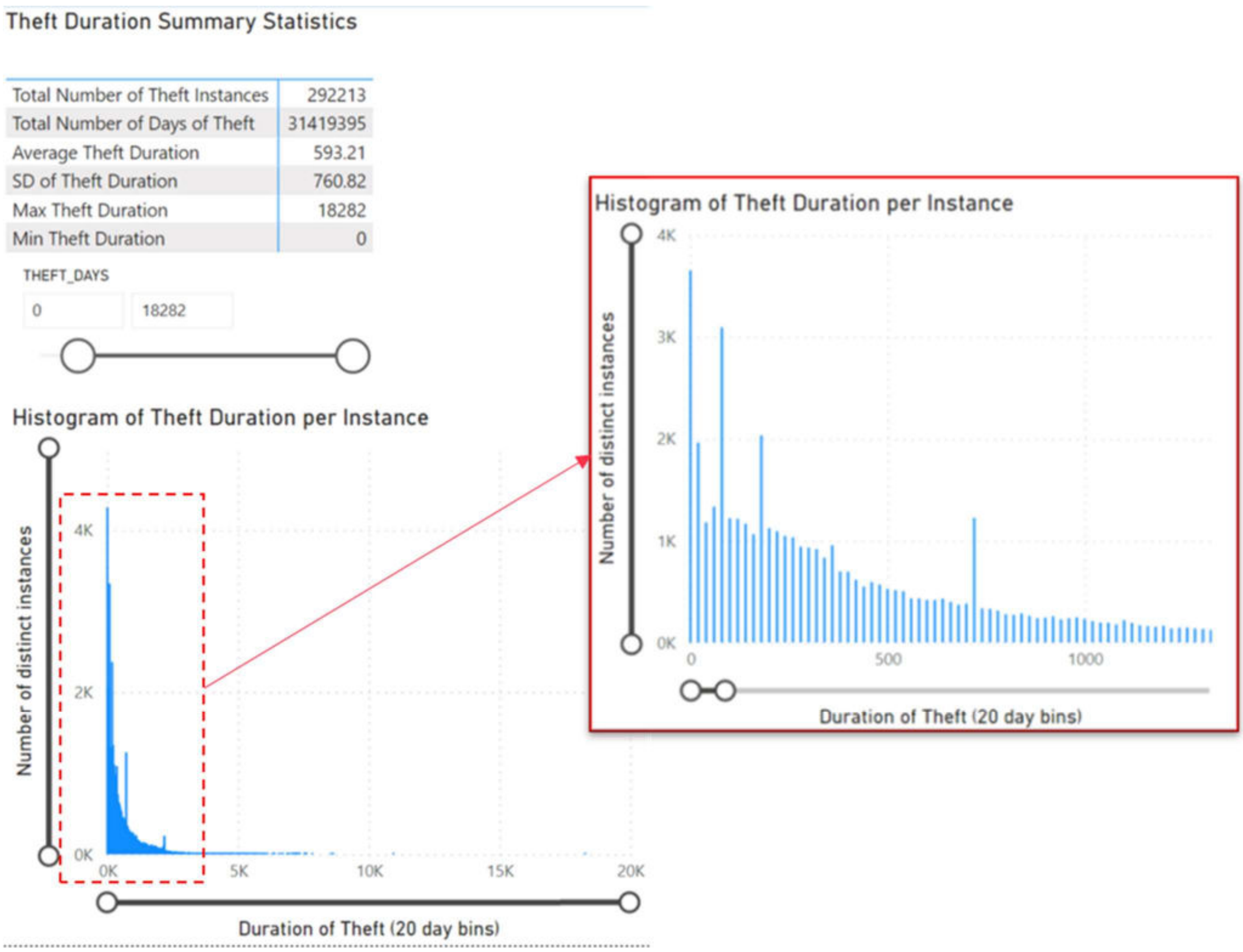
Table 4: Proportion of post-TRAS investigations per outcome status

Year	Not started	Under investigation	Confirmed theft	Faulty Meter	No longer supplied	No theft	Unproven suspicion	Total
2021	2%	1%	20%	1%	3%	73%	1%	100%
2022	1%	1%	23%	0%	1%	73%	2%	100%

TRAS-THEFT DURATION

Using the theft start dates and end dates, it was possible to explore the typical durations of theft as shown below in Figure 6. Again, through discussions with the SMEs it was possible to validate some of the unusual activity within the data. The histogram counts instances together in each 20-day banding. The peaks within the charts show where there are larger counts of theft occurring for that duration banding.

Figure 6: Summary of theft duration statistics from TRAS data



Theft Duration Summary Statistics

Total Number of Theft Instances	292213
Total Number of Days of Theft	31419395
Average Theft Duration	593.21
SD of Theft Duration	760.82
Max Theft Duration	18282
Min Theft Duration	0

THEFT_DAYS

0 18282

Histogram of Theft Duration per Instance
Number of distinct instances
4K
2K
0K
0K
5K
10K
15K
20K
Duration of Theft (20 day bins)
Histogram of Theft Duration per Instance
Number of distinct instances
4K
3K
2K
1K
0K
0
500
1000
Duration of Theft (20 day bins)

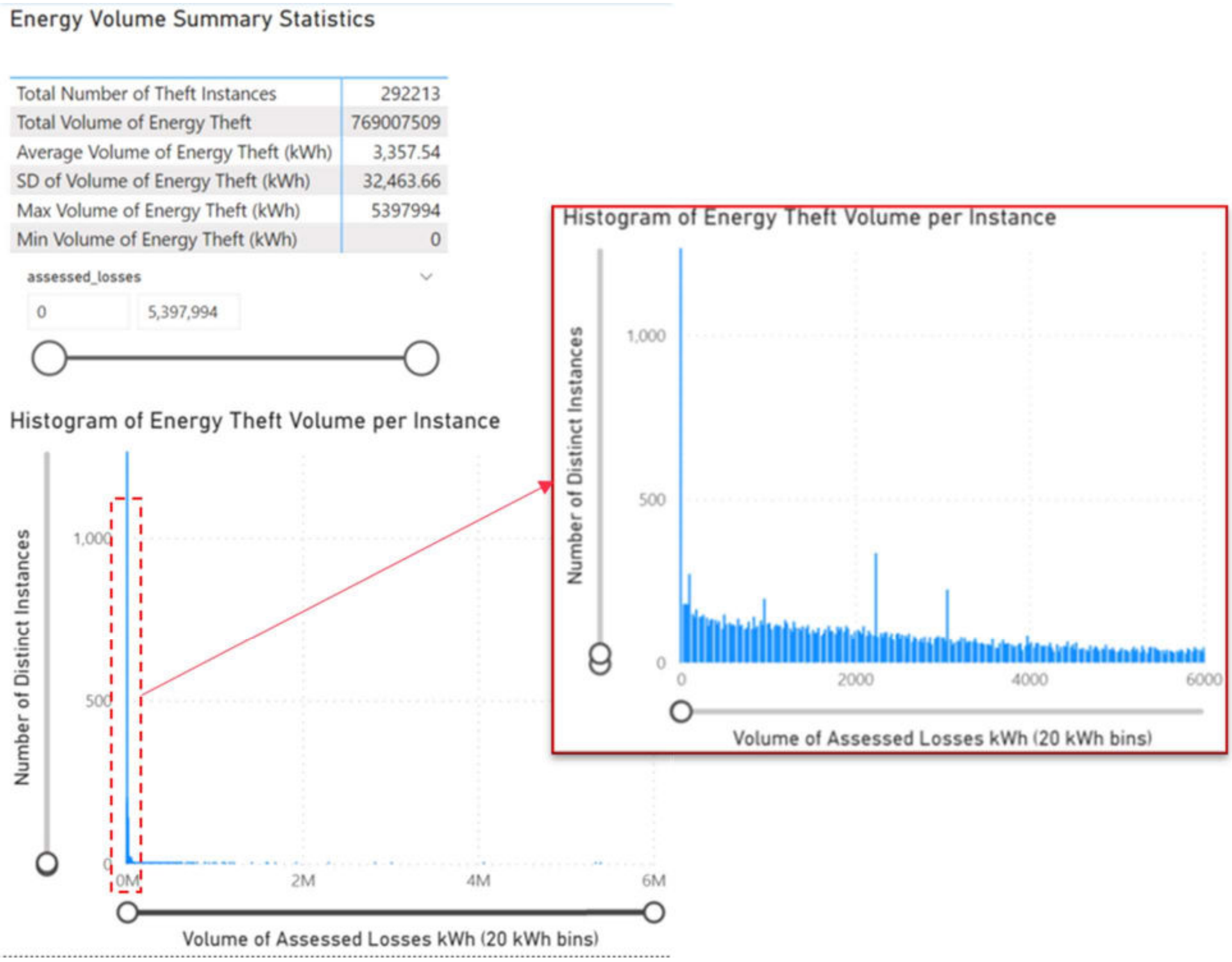
Below summarizes the key findings from this analysis and outcomes of the discussions with SMEs:

- Instances of negative theft days should be investigated on a case by case basis, or removed. These were likely due to input errors, such as the start and end dates of theft to be the wrong way round. TRAS validation checks were not in place when it was launched but did come in later to verify these errors.
- One instance of theft has a duration of 30 years which SME's advised to remove.
- There were clear spikes within the duration plot which were reviewed with the SMEs, and plausible reasons for these to remain within the data were found including:
    - Duration spikes at 90 and 180 days align with the growing cycle of cannabis.
    - Duration spikes which correspond to quarterly meter reads.
    - Duration spikes at 365 and 730 days align with suppliers having 1 or 2 year cut offs when recording theft duration.

TRAS THEFT – ENERGY VOLUME

A histogram showing the volume of energy theft was reviewed to understand the distribution of volumes and potential errors, and these are presented below in Figure 7. There were some very large losses recorded and these were explored individually. The histogram counts instances together for every 20kWh banding. Like with the durations, the peaks within the volumes indicate a higher number of instances within that volume banding.

Figure 7: Energy volume summary statistics from TRAS data



Energy Volume Summary Statistics

Total Number of Theft Instances	292213
Total Volume of Energy Theft	769007509
Average Volume of Energy Theft (kWh)	3,357.54
SD of Volume of Energy Theft (kWh)	32,463.66
Max Volume of Energy Theft (kWh)	5397994
Min Volume of Energy Theft (kWh)	0

assessed_losses
0
5,397,994
Histogram of Energy Theft Volume per Instance
Number of Distinct Instances
1,000
500
0
0M
2M
4M
6M
Volume of Assessed Losses kWh (20 kWh bins)
Histogram of Energy Theft Volume per Instance
Number of Distinct Instances
1,000
500
0
0
2000
4000
6000
Volume of Assessed Losses kWh (20 kWh bins)

Below summarises the key findings from this analysis and outcomes of the discussions with RECCo's SMEs:

- The largest volume of energy theft (20161118 kWh in 155 days) should be removed as it appears a date was entered in the assessed losses field.
- Other large entries (e.g. 4,728,881 kWh in 305 days) could be legitimate based on SME experience. This equates to approximately 15,000 kWh/ day which is a legitimate volume of energy theft for cannabis farms/bitcoin mining.
- There is a large spike in gas thefts with 0 assessed losses which are likely due to fiscal thefts where theft occurs without a recorded volume. A specific tamper code was brought in to TRAS to cater for these.
- The larger volumes of energy theft are typically associated with electricity and their investigations generated by Crimestoppers.
- Electricity counts for larger quantities of energy theft compared to gas.

3.2. EXPLORATORY DATA ANALYSIS SUMMARY

The exploratory analysis demonstrated there is some noise in the data, particularly the early years of the TRAS data which had clear data entry errors that were removed when validation checks were later introduced. Across the two data sets there is not a consistent time series available due to external factors such as covid, but also reporting factors such as changing ownership of the investigation submissions, however there is a decrease of investigations concluding in the years since the 2019 peak.

The ability to visualise the data and discuss the outputs with the SME’s has proved worthwhile. Without this working knowledge, it would be difficult to explain some of the unusual peaks, particularly within the durations where there were clear peaks at 3 months, 6 months, and 2 years. As discussed, these could be explained by identification of theft linked with cannabis growing cycle completion or linked with quarterly meter checks.

Exploration of some of the outliers, particularly the higher volumes stolen demonstrated that some of these data points were legitimate when interrogated, whilst others were obvious data entry errors. On the lower end, there were genuine reasons why many confirmed thefts had a zero recorded for assessed losses.

4. HYPOTHESIS TESTING

This section provides an overview of the hypothesis testing carried out. Hypotheses were defined to determine biases within the data and test some anecdotal information. They were also to support the identification of useful predictors that should be included within the predictive model. Data visualisation using Power BI followed by discussion with theft SME's were used to determine the outcome.

There were many myths and assumptions described during the discussion of energy theft so it was determined that alongside the building of the model, the project would test some hypothesis with the data available. This had the additional benefit of being able to rule out the inclusion of data sets if assumptions did not hold and there was no clear link to the data. This section provides the details of the hypothesis list that was generated, which ones were able to be tested with the data within the time frame and the ability to determine the outcome of the hypothesis.

4.1. HYPOTHESIS IDENFICATION

The early work in 2021 identified some hypotheses which were intended to be explored. These were reviewed in line with the data that was available for the resumption of this project to determine which ones to take forward. In addition, some additional hypotheses were added. Due to the time constrained nature of this project, the full list was prioritised with the SME's and those within the 'Must do' classification were taken forward for testing. The full list of revised hypotheses are presented in Table 5 below with those highlighted as 'Must Do' in green. The complete list of hypotheses, rationales and tests can be found in Appendix C. There is a summary of the outcomes in Table 7.

Table 5: List of established predictors taken forward to the predictive modelling stage

Ref	PRIMARY DATA	Priority
H1	Some suppliers are better at proactive investigations	Must do
H2	Those committing energy theft do not change supplier	Should do
H3	Theft can occur equally in the properties with or without smart meters	Would do
H4	Thefts may incorrectly be reported as faulty meters	Would do
H5a	Parties who have stolen in one instance are likely to do so in others	Could do
H5b	Theft may recur at the same property	Could do
H6	Thefts are more likely to be detected by suppliers in prepaid meters than credit meters	Must do
H7	Suppliers don't monitor credit meters enough where significant amount of theft occurs	Should do
H8	Potential thefts in remote locations are less likely to have been investigated	Must do
H9	Incidence of theft is likely to have increased due to the financial impact of Covid and lockdowns	Should do
H10	Cost of living impacts will increase theft	Should do
H11	Potential thefts in difficult to resource locations are less likely to have been investigated	Must do
H12	Crimestoppers campaign areas will have increased theft rates	Should do

H13	There has been a bias to investigations within deprived areas	Must do
H14	High deprivation is likely to have higher levels of theft	Must do
H15	Fuel poverty levels may show high association with the levels of theft	Must do
H16	Newly subdivided property areas have more unregistered meters (error /theft)	Would do
H17	Social housing areas are likely to have higher levels of theft	Could do
H18	Non energy crime is an indicator of energy theft	Would do
H19	Energy theft levels strongly relate to the levels of other types of crime	Must do
H20	There may be correlation with water theft for some theft types (e.g. cannabis farms)	Would do
H21	Theft is more likely to be detected in cannabis farms	Would do
H22	Theft is more likely in hospitality relating businesses, nursing homes, laundrettes	Should do
H23	Theft is more likely in poor quality businesses such as those with low hygiene levels	Should do
H24	Small/ medium business are more likely be involved in energy theft than large corporations	Should do

4.2. FINDINGS

Each hypothesis was tested via an interactive Power BI dashboard to quickly 'slice-and-dice' the data using different filters. Static snapshots showing the key findings from the 'Must Do' hypotheses are shown and described below.

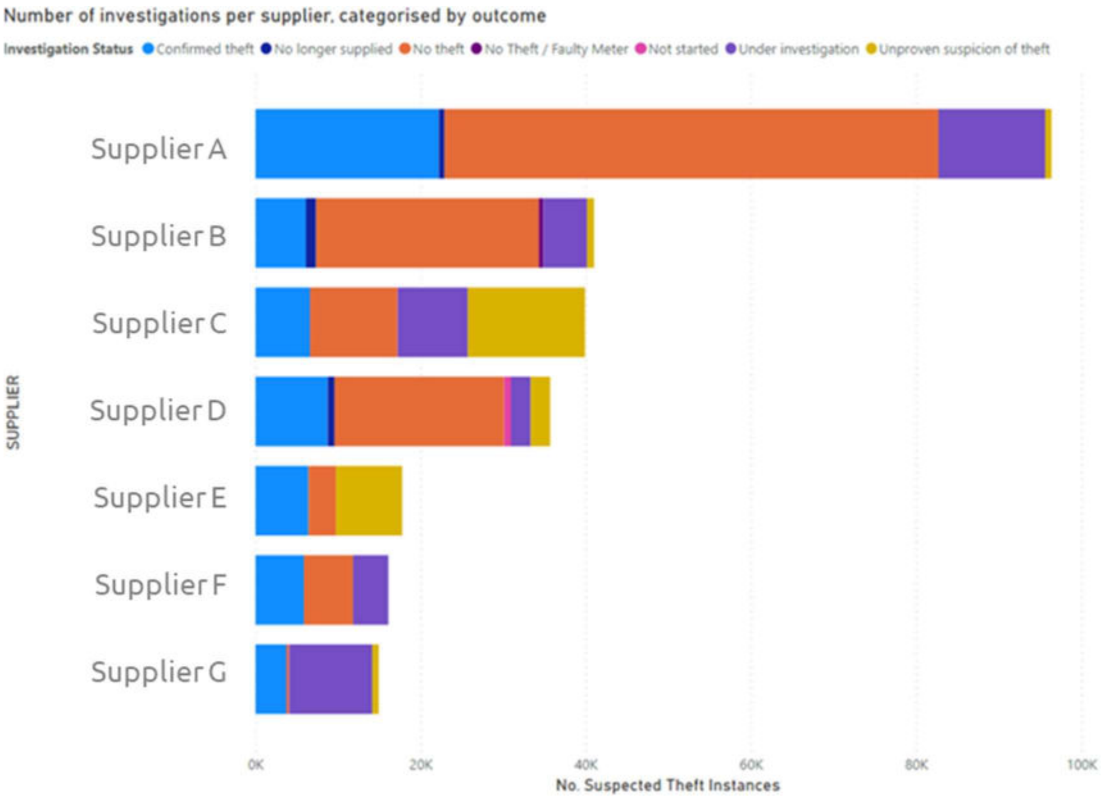
H1: SOME SUPPLIERS ARE BETTER AT PROACTIVE INVESTIGATIONS

There is a view that some suppliers are better at reporting and proactively investigating theft investigations compared to others. This is important to understand since suppliers who for example, only report confirmed cases or only investigate certain geographical areas, would not provide reliable data to inform the 'reliable regions' to feed into the training data for the model.

To test this hypothesis the TRAS data was used to explore the number of investigations, the relative number of investigations per GSP area, and the proportion of investigations by lead source. Since the purpose of this is to purely understand the differences between suppliers, each supplier below has been anonymised.

Figure 8 shows the number of investigations per supplier, categorised by outcome, as well as the key findings. It was expected that different suppliers will have a different number of investigations depending on the size and scale of the supplier, however this analysis is also useful to understand which suppliers have reasonable conversion rates of investigations to confirmed thefts.

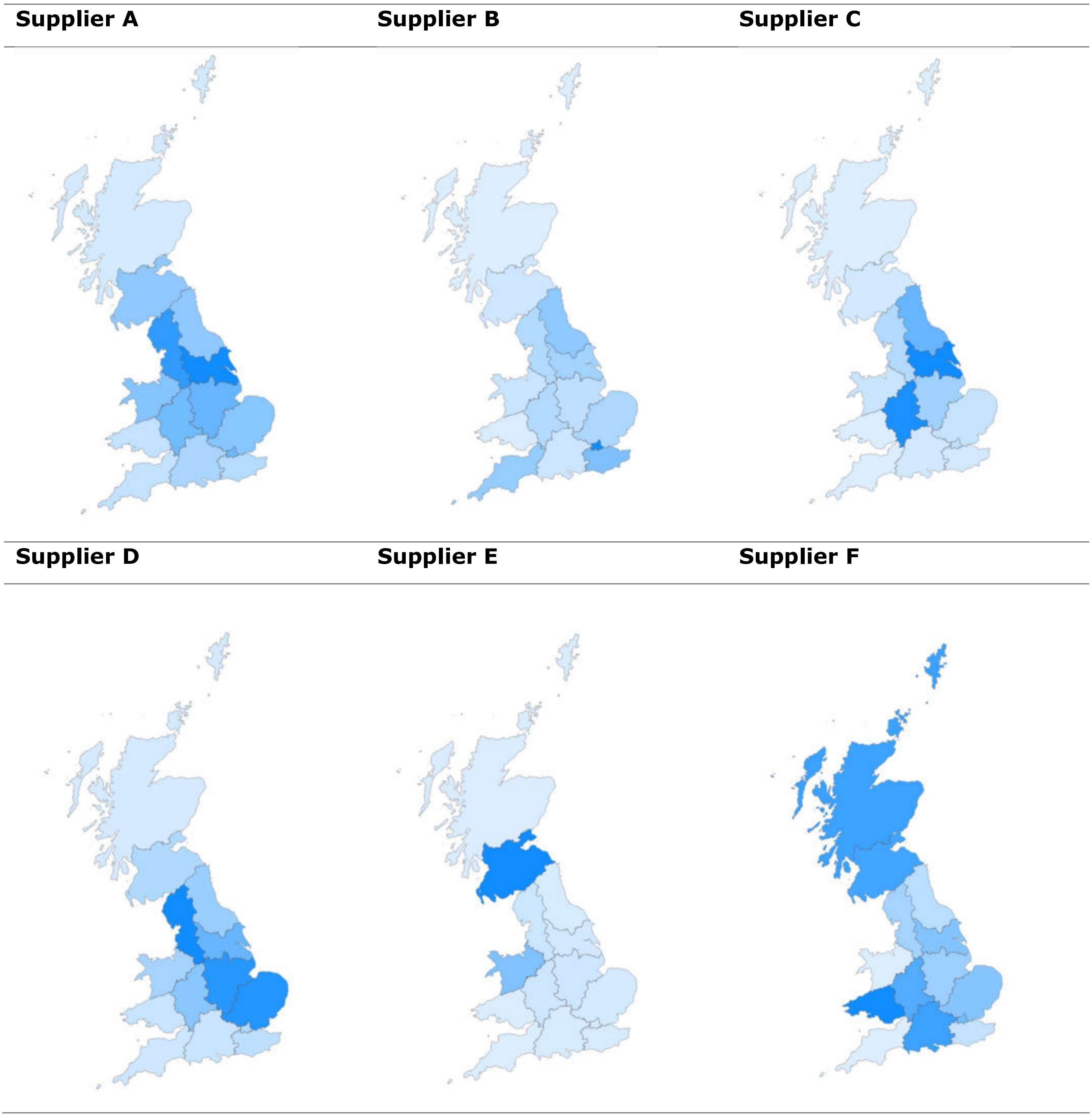
Figure 8: Number of investigations per supplier (anonymised), categorised by outcome



- Supplier A have 90k+ instances of suspected theft, more than double any other supplier
- Suppliers A, B and D have the largest proportion of no thefts
- Discussions with SME's suggested that a 20% rate of 'confirmed thefts' was reasonable to assume if suppliers had been accurately reporting investigations.

Figure 9 shows the relative number of investigations per GSP area, where darker blue indicates a higher number of investigations. There is a view that suppliers with a greater field force in a region have the means to analyse their data to a greater extent and therefore have been able to investigate theft more proactively. This is supported by the findings below. Some suppliers are focused on particular geographical areas, and this should be considered when selecting reliable regions.

Figure 9: Relative number of investigations per GSP region



Supplier A

Supplier B

Supplier C

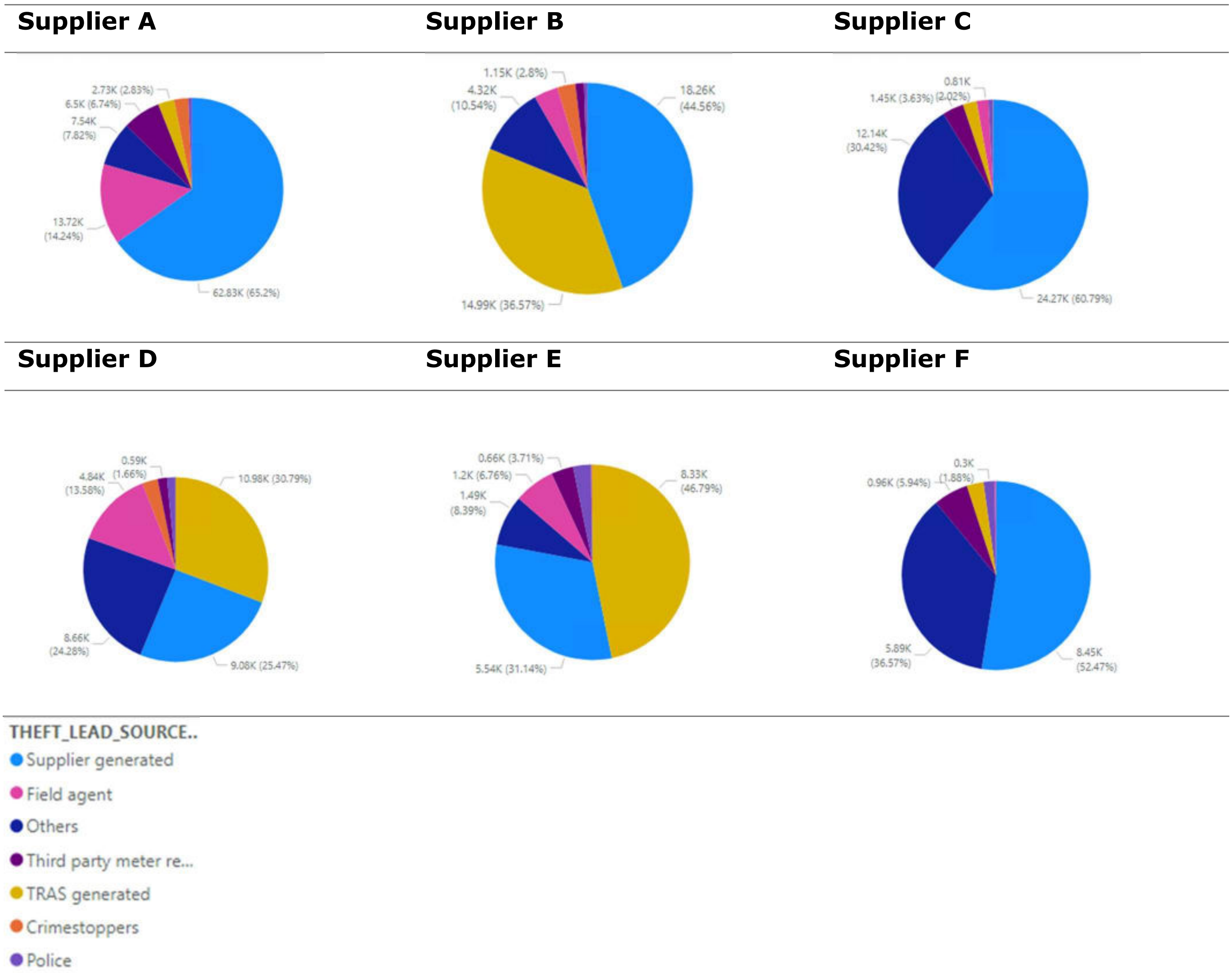
Supplier D

Supplier E

Supplier F

In addition, it is suspected that suppliers focus their attention to different lead sources for investigations. Figure 10 shows the proportion of investigations per lead source for each anonymised supplier and key findings are listed below.

Figure 10: Lead source of investigation by supplier (anonymized)



- In general, supplier generated theft formed the largest proportion of theft investigations
- Suppliers A and C are the most proactive in terms of investigating thefts, with considerably more supplier generated thefts than any other supplier.

Other key insights from the data exploration related to this hypothesis are as follows:

- Highest number of investigations in the Midlands, North and Southeast of England. Fewest investigations in Southwest England, Wales and Scotland.
- Confirmed theft conversion rate is 24%, and no Theft conversion rate is 50%, which aligns with SME experience.
- Most investigations are reactive (Supplier generated/ Field agents).
- Confirmed theft conversion rate for proactive investigations is much lower (15%) compared to reactive investigations (42%).
- Confirmed theft conversion rate for commercial properties is lower (12%) compared to residential (25%).
- Smaller suppliers tend to have higher confirmed theft conversion rates.

- Supplier A and Supplier B follow a more expected pattern of theft reporting.

The conclusions that can be drawn from all the above analyses are:

- Larger suppliers (A, B, C and D) appear to be more reliable in terms of the accuracy of their reporting of theft instances (both confirmed and no thefts).
- The spatial spread of theft by these suppliers is more focused around Mid, North and South-East England.
- Training dataset for predictive model should focus on these more reliable suppliers.

H6: THEFTS ARE MORE LIKELY TO BE DETECTED BY SUPPLIERS IN PREPAID METERS THAN CREDIT METERS

Discussions with SMEs indicated that there is likely to be a bias in investigations towards customers using pre-payment meters compared to credit meters. To investigate this, the number of investigations for each different payment method and investigation outcome (as recorded in TRAS) were analysed. These findings are presented in Figure 11 below. Table 6 shows the proportion of investigations per outcome, for each different payment method.

It should be noted that 50% of investigations within the TRAS dataset did not record the payment method, and so the data below relates to the remaining 50% of the data with valid entries.

The bar chart below shows that 66% of the valid investigations were related to pre-payment meters, supporting the hypothesis that there is a bias in investigations towards these customers. A further 2% of investigations were related to PAYG customers. Table 6 shows that of the pre-payment and PAYG investigations, 18% and 34% respectively resulted in confirmed thefts. This is comparable to those paying via Standing Order and transactional methods, but higher that those paying via Direct debit, and therefore this analysis does support the hypothesis.

The outcome of this hypothesis is assessed purely on the number of investigations, but further analysis could be undertaken to explore whether the volume of energy stolen varies by meter type.

Figure 11: Number of theft investigations by normal payment method and investigation status outcome

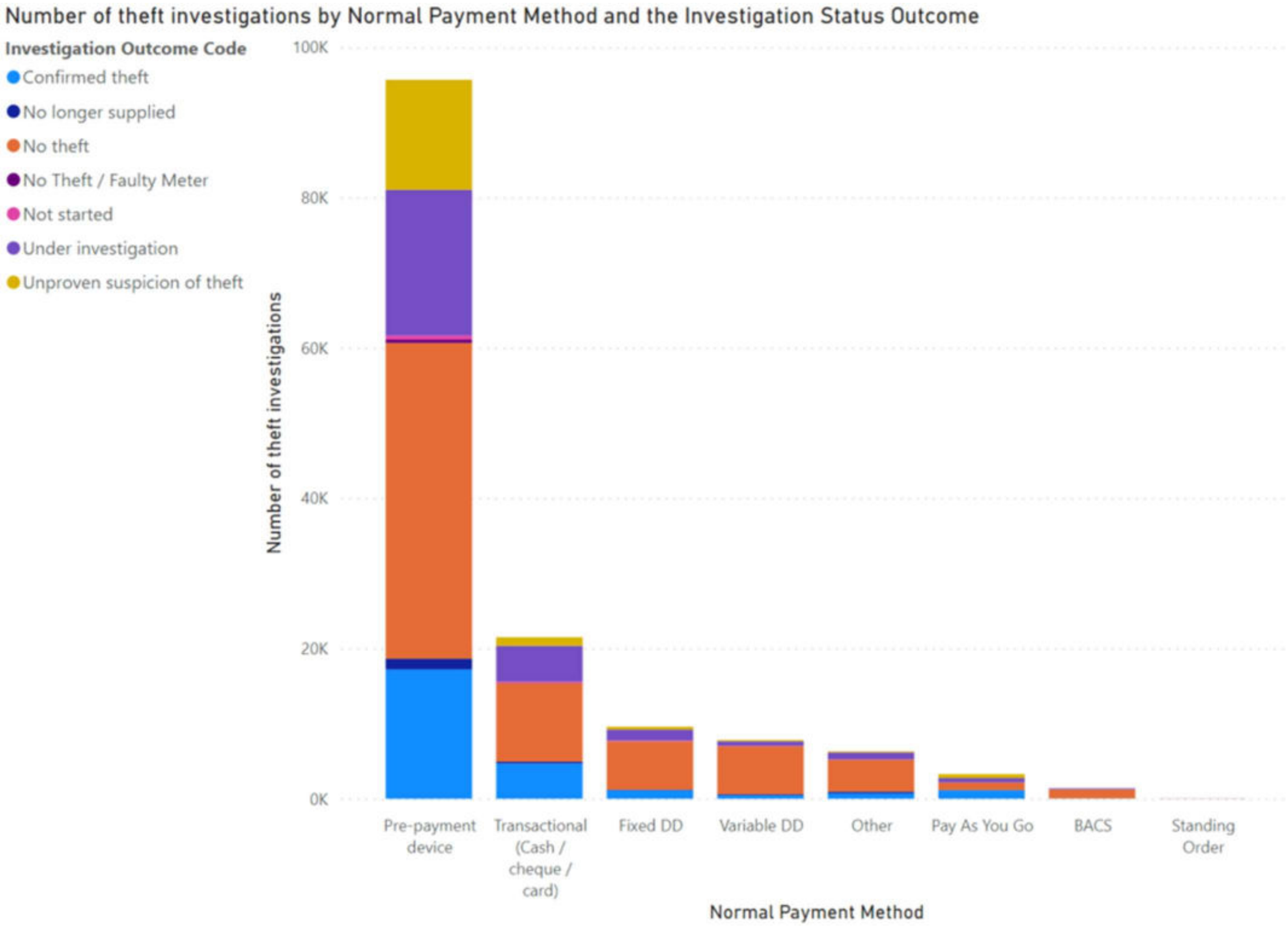
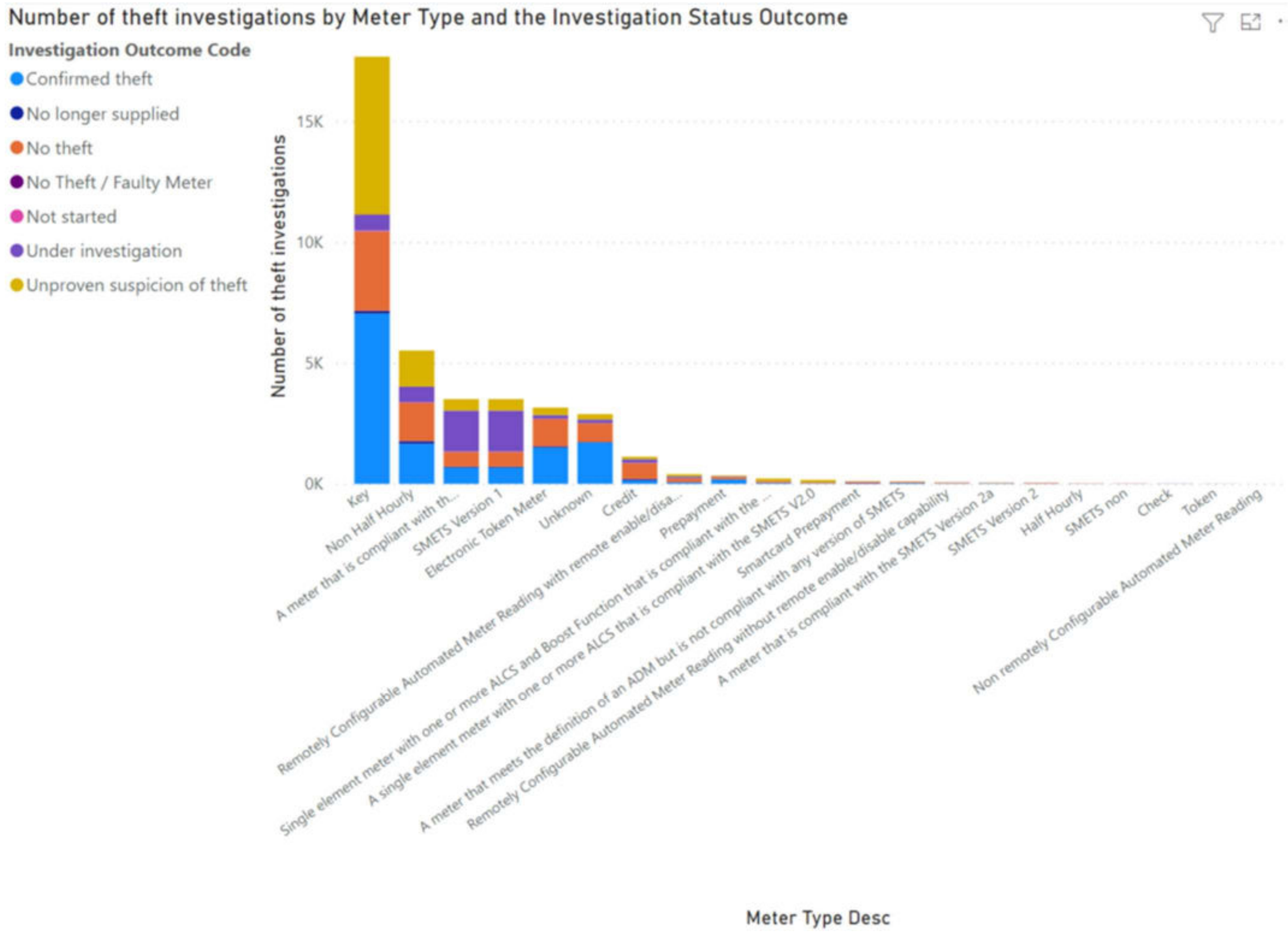


Table 6: Proportion of investigations per outcome, for each payment method

Payment Method	Confirmed theft	No longer supplied	No theft	Not started	Under investigation	Unproven suspicion of theft	Total
BACS	2%	2%	85%	0%	10%	1%	**100%**
Fixed DD	12%	1%	67%	1%	15%	4%	**100%**
Other	12%	3%	68%	0%	15%	2%	**100%**
Pay As You Go	34%	1%	30%	1%	17%	16%	**100%**
Pre-payment device	18%	1%	44%	1%	20%	15%	**100%**
Standing Order	20%	0%	61%	2%	16%	0%	**100%**
Transactional (Cash / cheque / card)	22%	1%	49%	1%	22%	5%	**100%**
Variable DD	6%	2%	82%	1%	8%	2%	**100%**

Of the 50% of the data with no valid payment method, the proportion of investigation outcomes based on the meter type was assessed, and this is shown below in Figure 12. Approximately 18k of these investigations were Key (pre-payment) meters of which 40% are recorded as confirmed thefts. These findings align with the above, that there is a bias in investigations towards pre-payment meters and that the conversion rate to confirmed thefts is higher than with other meter types.

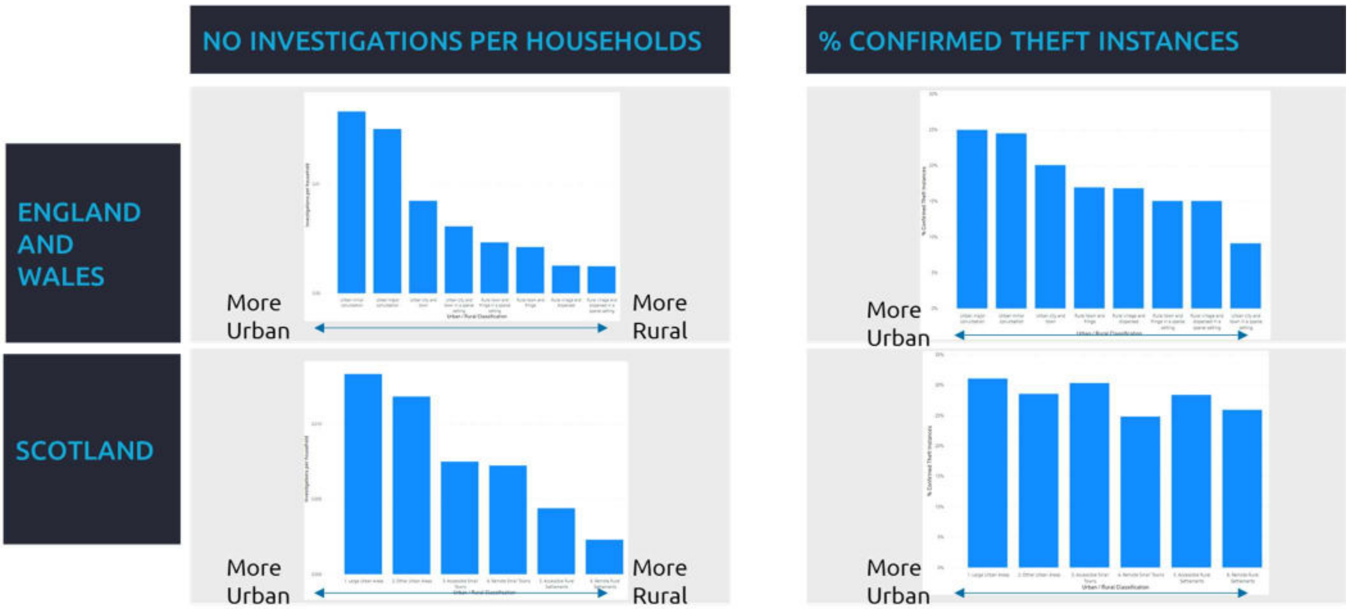
Figure 12: Number of theft investigations by meter type and status for the 50% of investigations with no valid normal payment type



H8: POTENTIAL THEFTS IN REMOTE LOCATIONS ARE LESS LIKELY TO HAVE BEEN INVESTIGATED

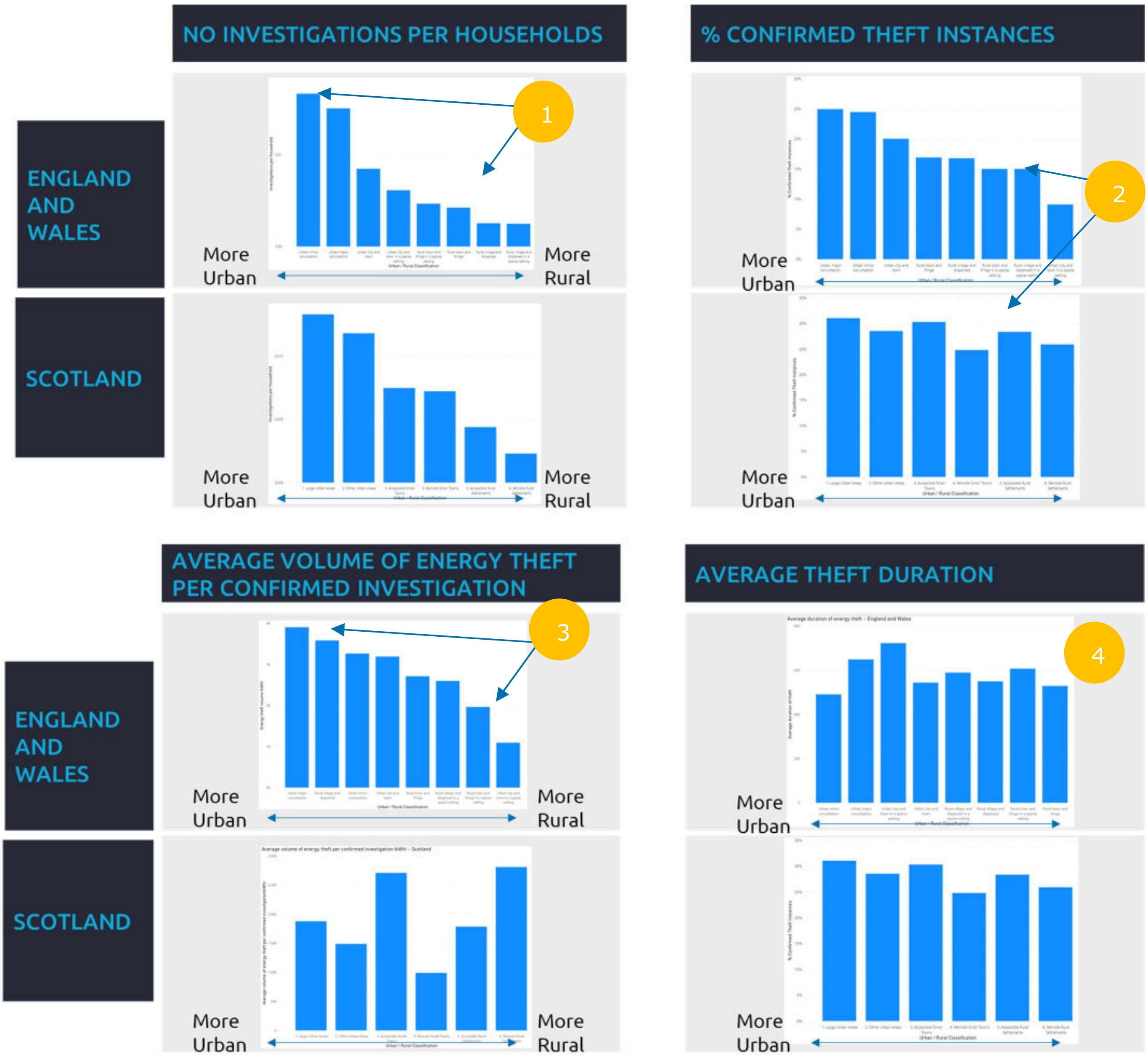
Using rural/urban classifications, the number of investigations and the percentage of confirmed theft instances within these areas were visualised to test if potential thefts within rural areas are less likely to be investigated. This is shown in

Figure 13 Investigations by urban and rural classification



below and the key findings are listed below, which correspond with the numbers on the diagram.

Figure 13 Investigations by urban and rural classification



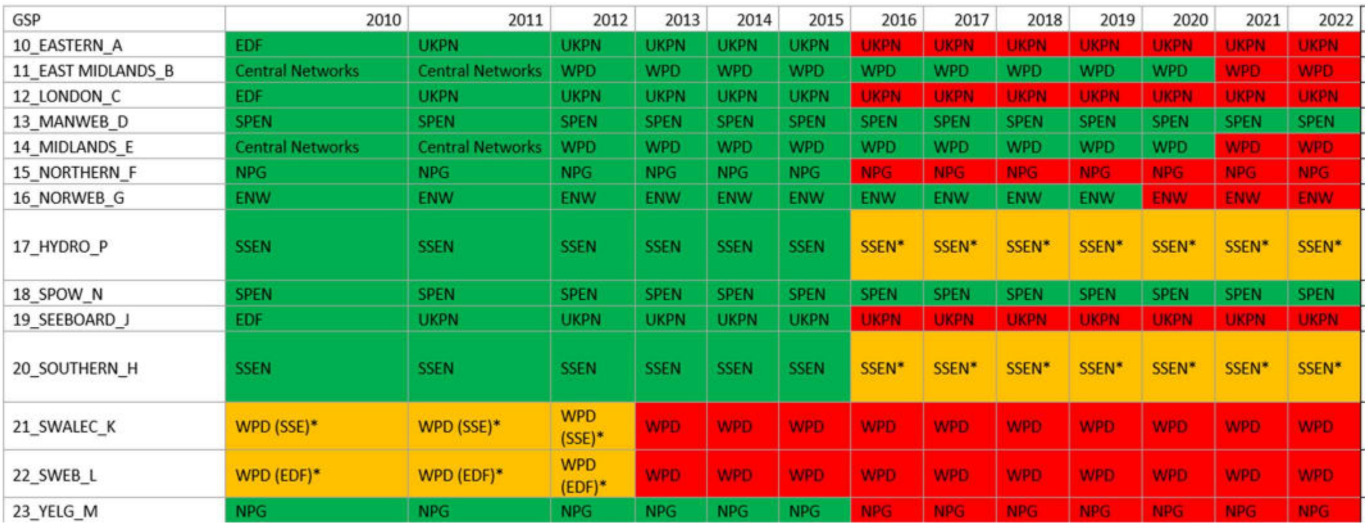
1. There are more investigations per household in urban regions than rural.
2. Scotland is achieving a more consistent confirmation of theft in all regions than England and Wales.
3. The average volume of energy theft per confirmed case is larger in urban areas in England and Wales, compared to rural. The distribution in Scotland is similar across urban and rural areas.
4. The average duration of energy theft per confirmed case is fairly consistent across all urban/rural areas in England, Wales and Scotland.

H11: POTENTIAL THEFTS IN DIFFICULT TO RESOURCE LOCATIONS ARE LESS LIKELY TO HAVE BEEN INVESTIGATED

There was a suggestion to explore investigations over time to determine if the availability of a workforce within an area influenced the investigations taking place. The intent was to evaluate changes in investigation volumes and outcome when DNO's withdrew investigation services within their regions, the data for which is shown in Figure 14. On assessment of the timings there were 2 clear observations

1. 2015/16 is a key transition year for withdrawal of DNO services, and as shown previously in Figure 4 (which the number of investigations split by investigation close date), there is low volume of investigations recorded within TRAS in this time period.
2. 2019/20/21 see withdrawals in 3 regions however figures are likely to be influenced by covid restrictions.

Figure 14: Year at which DNO's withdrew investigation services within their regions



GSP	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
10_EASTERN_A	EDF	UKPN	UKPN	UKPN	UKPN	UKPN	UKPN	UKPN	UKPN	UKPN	UKPN	UKPN	UKPN
11_EAST MIDLANDS_B	Central Networks	Central Networks	WPD	WPD	WPD	WPD	WPD	WPD	WPD	WPD	WPD	WPD	WPD
12_LONDON_C	EDF	UKPN	UKPN	UKPN	UKPN	UKPN	UKPN	UKPN	UKPN	UKPN	UKPN	UKPN	UKPN
13_MANWEB_D	SPEN	SPEN	SPEN	SPEN	SPEN	SPEN	SPEN	SPEN	SPEN	SPEN	SPEN	SPEN	SPEN
14_MIDLANDS_E	Central Networks	Central Networks	WPD	WPD	WPD	WPD	WPD	WPD	WPD	WPD	WPD	WPD	WPD
15_NORTHERN_F	NPG	NPG	NPG	NPG	NPG	NPG	NPG	NPG	NPG	NPG	NPG	NPG	NPG
16_NORWEB_G	ENW	ENW	ENW	ENW	ENW	ENW	ENW	ENW	ENW	ENW	ENW	ENW	ENW
17_HYDRO_P	SSEN	SSEN	SSEN	SSEN	SSEN	SSEN	SSEN*	SSEN*	SSEN*	SSEN*	SSEN*	SSEN*	SSEN*
18_SPOW_N	SPEN	SPEN	SPEN	SPEN	SPEN	SPEN	SPEN	SPEN	SPEN	SPEN	SPEN	SPEN	SPEN
19_SEEBOARD_J	EDF	UKPN	UKPN	UKPN	UKPN	UKPN	UKPN	UKPN	UKPN	UKPN	UKPN	UKPN	UKPN
20_SOUTHERN_H	SSEN	SSEN	SSEN	SSEN	SSEN	SSEN	SSEN*	SSEN*	SSEN*	SSEN*	SSEN*	SSEN*	SSEN*
21_SWALEC_K	WPD (SSE)*	WPD (SSE)*	WPD (SSE)*	WPD	WPD	WPD	WPD	WPD	WPD	WPD	WPD	WPD	WPD
22_SWEB_L	WPD (EDF)*	WPD (EDF)*	WPD (EDF)*	WPD	WPD	WPD	WPD	WPD	WPD	WPD	WPD	WPD	WPD
23_YELG_M	NPG	NPG	NPG	NPG	NPG	NPG	NPG	NPG	NPG	NPG	NPG	NPG	NPG

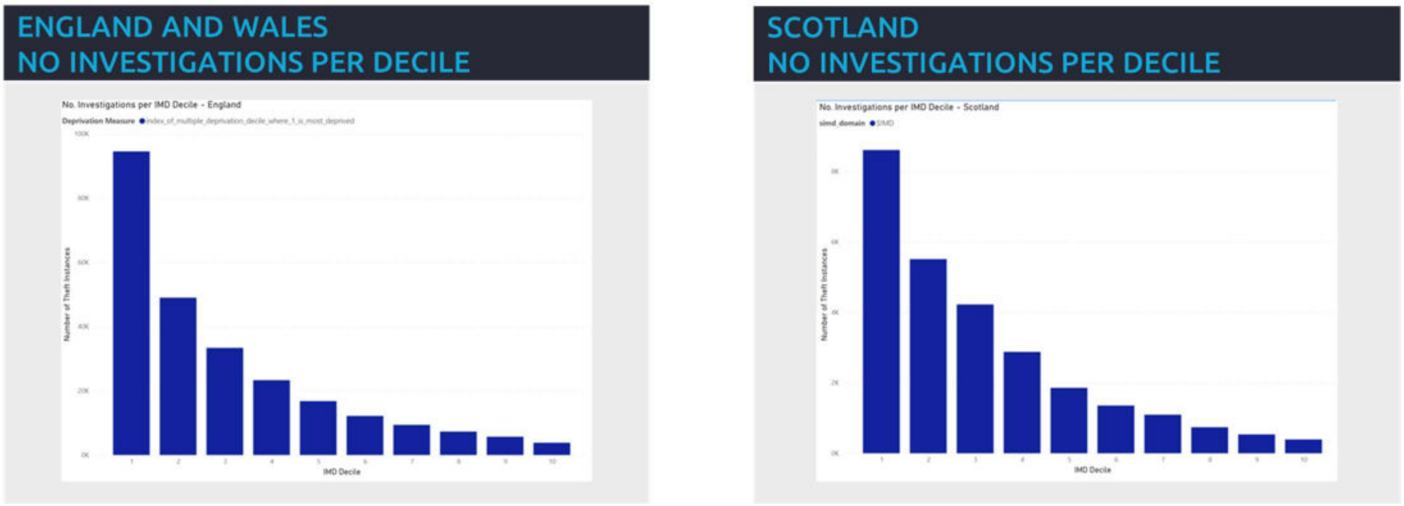
It was therefore not possible to draw a conclusion on the outcome of the availability of resources to investigate potential thefts.

H13: THERE HAS BEEN A BIAS TO INVESTIGATIONS WITHIN DEPRIVED AREAS

Investigations and success rates of confirmed thefts within the deprivation deciles have been compared to understand if there has been any bias in targeting the most deprived areas. Figure 15 below highlights there are more investigations occurring in the most deprived areas.

Output represents investigation per IMB decile (1 being most deprived areas).

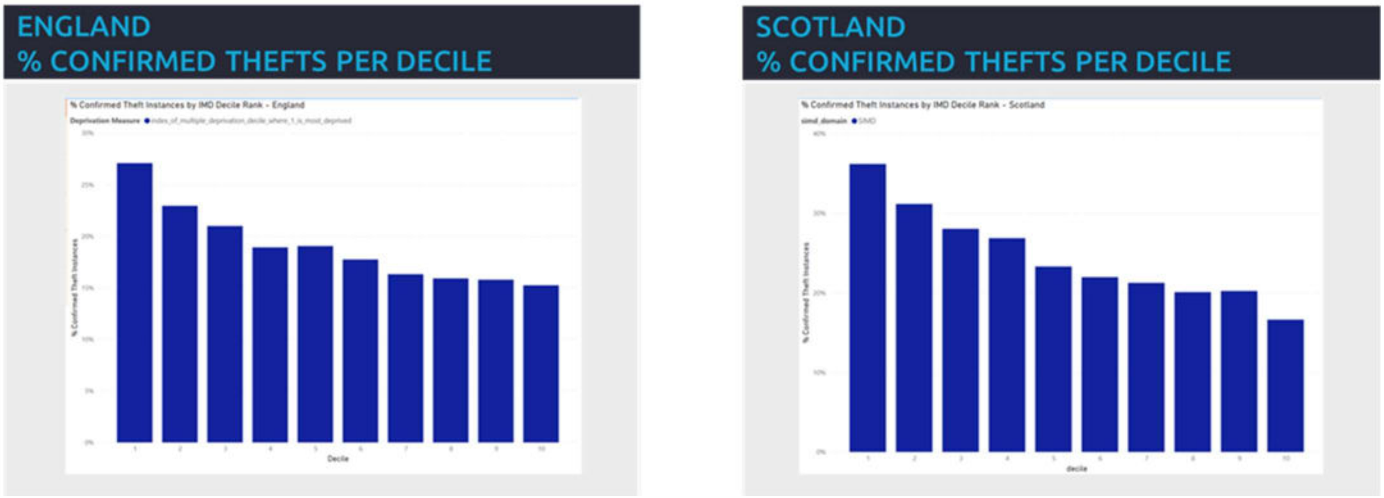
Figure 15: Number of investigations per IMD decile in England/Wales (left) and Scotland (right)



H14: HIGH DEPRIVATION ARE LIKELY TO HAVE HIGHER LEVELS OF THEFT

There is a skew within the deprivation deciles for rates of confirmed theft, however this is not as significant as it was within the volumes of investigations. Figure 16 represents percentage of confirmed theft cases per IMD decile (1 being most deprived areas).

Figure 16: Percentage of confirmed thefts per IMD decile in England/Wales (left) and Scotland (right)

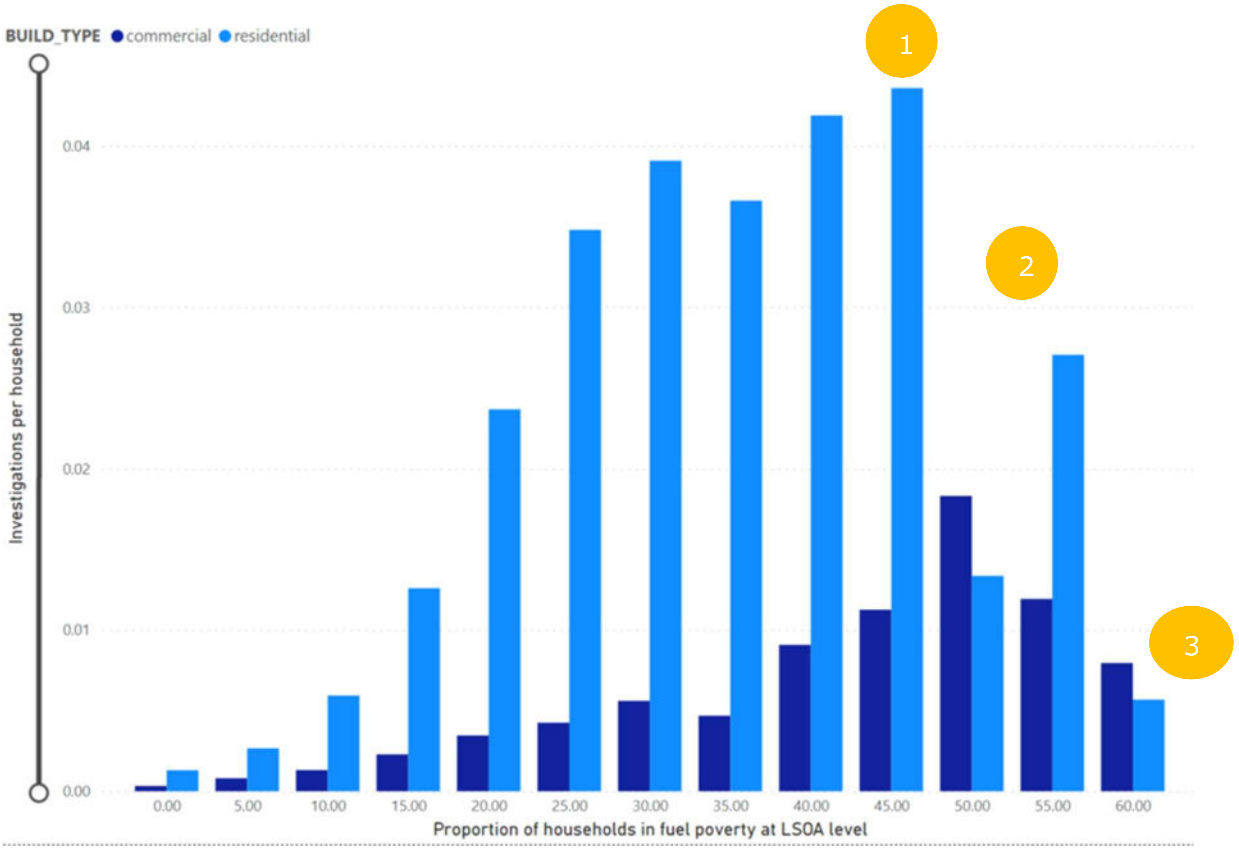


H15. FUEL POVERTY LEVELS MAY SHOW HIGH ASSOCIATION WITH THE LEVELS OF THEFT

Fuel poverty levels are reported in terms of the proportion of households in fuel poverty at LSOA level. For the upper bandings of these, there are fewer LSOAs within the bandings and so some of the conversion rates can be unreliable.

Figure 17 shows that the number of investigations per household does correlate with fuel poverty.

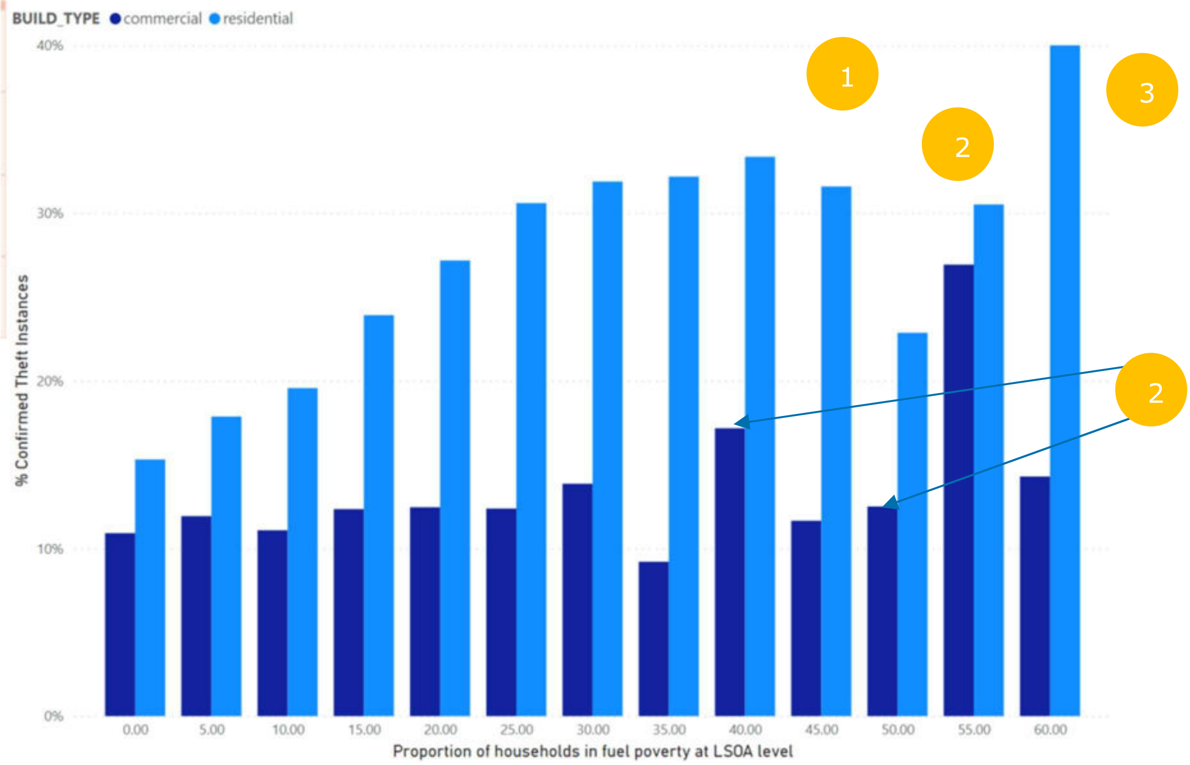
Figure 17: Number of investigations per household, split by fuel poverty band



1. For residential buildings, the highest rate of theft investigations occurred within LSOAs that had between 45-50% of households considered as fuel poor.
2. For commercial buildings, the highest rate of theft investigations occurred within LSOAs that had between 50-60% of households considered as fuel poor.
3. There a relatively few LSOAs with fuel poverty levels exceeding 50%, which explains the sharp drop in number of investigations per household beyond the 50% fuel poverty band.

Figure 18 below shows the percentage of confirmed theft investigations per fuel poverty band

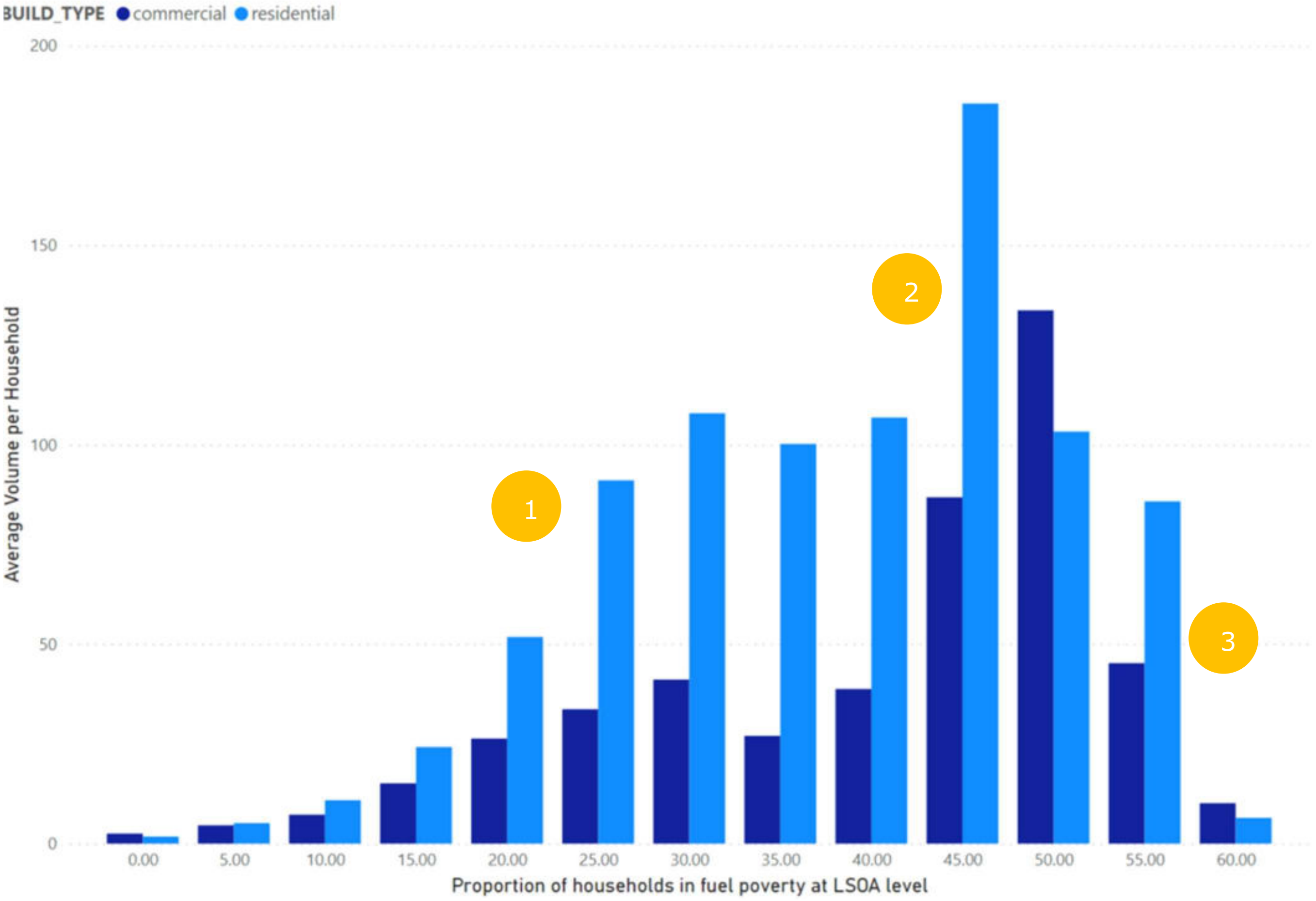
Figure 18: Percentage of confirmed thefts per household within each fuel poverty band



1. For residential buildings, the percentage of confirmed thefts increases within areas where fuel poverty levels are higher.
2. For commercial buildings, the percentage of confirmed thefts shows a weaker correlation with fuel poverty levels. There is a notable spike in confirmed theft rates within the 55-60% fuel poverty band.
3. As discussed before, the reason for the fluctuation in confirmed theft rates in the higher fuel poverty bands, is due to the lack of LSOA's that fall within these categories. These higher bands are easily skewed by the results of single investigations.

Figure 19 below shows the average volume of assessed losses per household fuel poverty banding

Figure 19: Average volume of assessed losses per household fuel poverty band



Findings:

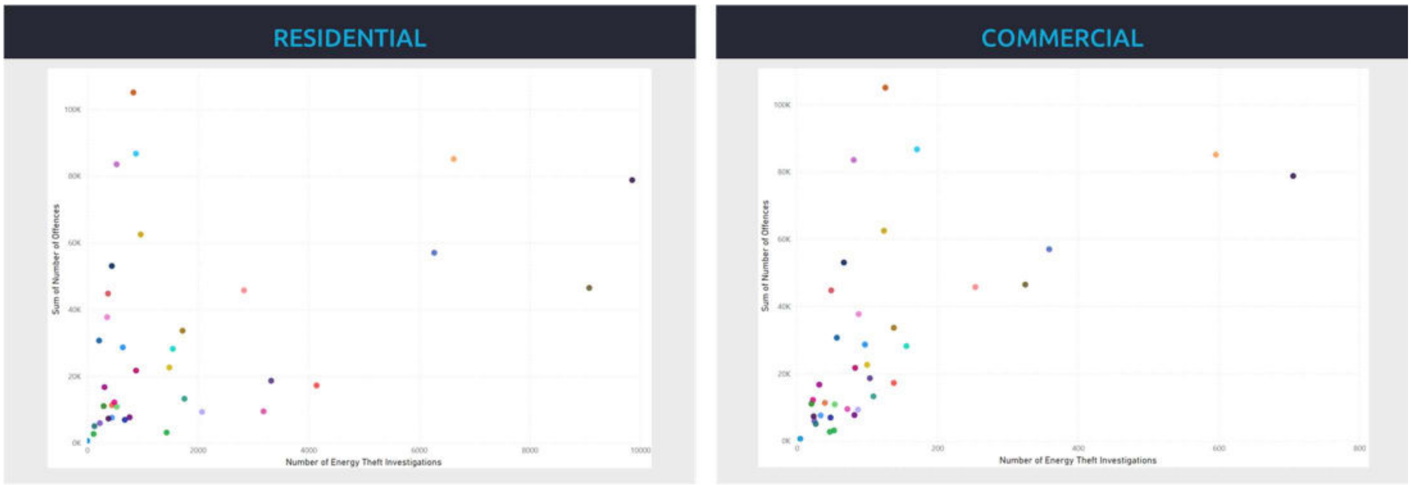
1. For residential buildings, there is greater volume of assess losses per household as the proportion of households in fuel poverty increases above 20%.
2. For commercial buildings, the volumes of assessed losses increase when the proportion of households in fuel poverty at LSOA level exceeds 45%.
3. There a relatively few LSOAs with fuel poverty levels exceeding 50%, which explains the sharp drop in number of investigations per household beyond the 50% fuel poverty band.

Overall, there is a greater level of theft with a greater volume amongst residential properties in areas with higher fuel poverty. For commercial properties, the correlation is not as strong but there is an increasing trend of cases.

H19: ENERGY THEFT LEVELS STRONGLY RELATE TO THE LEVELS OF OTHER TYPES OF CRIME

It was suggested through discussion with the SMEs that there are links between certain theft related crimes and energy theft. Data of crime types was available at Police Force Area level.

Figure 20: Scatter plot showing the relationship between confirmed instances of energy theft and other types of crime



Findings

1. There does not appear to be a clear correlation between crime levels and residential energy theft investigations.
2. There does appear to be a correlation between commercial energy theft investigations and the crime volumes within the policing areas.

The relationship with crime rates is mixed across residential and commercial properties. It is important to remember this is at a Police Force Area Level so this will hide some of the variance.

Table 7 below presents a summary of the outcomes of the hypotheses where a green reference box indicates the hypothesis was proven. An amber indicates the hypothesis was not able to be proven or there were mixed results.

Table 7: Summary of hypothesis outcomes

Ref	PRIMARY DATA	Priority	Outcome
H1	Some suppliers are better at proactive investigations	Must do	**True, suppliers have differing levels of proactive investigations**
H6	Thefts are more likely to be detected by suppliers in prepaid meters than credit meters	Must do	**True, outcomes show a much higher proportion of investigations where the meter is prepaid**
H8	Potential thefts in remote locations are less likely to have been investigated	Must do	**True - in Scotland it is less skewed resulting in a more consistent rate of confirmed thefts**
H9	Incidence of theft is likely to have increased due to the financial impact of Covid and lockdowns	Should do	**Unknown, unclear due to reducing volumes of investigations during covid and subsequently incomplete datasets with the transition of data capture**
H11	Potential thefts in difficult to resource locations are less likely to have been investigated	Must do	**There are too many external forces to reliably conclude this hypothesis**
H13	There has been a bias to investigations within deprived areas	Must do	**True, investigations are highly skewed to high deprivation areas**
H14	High deprivations are likely to have higher levels of theft	Must do	**True, there are higher rates of confirmed thefts in high deprivation areas but not as skewed as investigations**
H15	Fuel poverty levels may show high association with the levels of theft	Must do	**True, differing levels across residential and commercial however investigations are at similar levels across the deciles**
H19	Energy theft levels strongly relate to the levels of other types of crime	Must do	**Mixed, there is a higher correlation for commercial energy theft with levels of crime within an area**

4.3. PREDICTORS IDENTIFIED

The hypotheses tested a number of potential predictors to understand their connection with theft outcome and investigations. Through reviewing the outputs and discussions with SMEs, it was determined the predictors summarised in Table 8 below should be included within the model. Note that not all predictors were available for all countries in the same format and therefore some are omitted from certain model scenarios. Green and red indicates that the predictor was available and unavailable respectively.

Table 8: Summary of available predictors used for the Predictive Model

	Predictor Description	**England**	**Scotland**	**Wales**
1	Type of Meter (e.g. Credit, Prepayment)	🟩	🟩	🟩
2	Meter Location (e.g. Kitchen, Garage)	🟩	🟩	🟩
3	Energy Supplier	🟩	🟩	🟩
4	Normal Payment Method (e.g. pre-payment device, fixed DD)	🟩	🟩	🟩
5	Index of Multiple Deprivation Ranking	🟩	🟩	🟩
6	Population Density calculated at LSOA level	🟩	🟩	🟩
7	Urban/Rural classification	🟩	🟩	🟩
8	Percentage of bungalows within an LSOA, for England and Wales	🟩	🟥	🟩
9	Percentage of flats within an LSOA, for England and Wales	🟩	🟥	🟩
10	Percentage of terraced houses within an LSOA, for England and Wales	🟩	🟥	🟩
11	Percentage of semi-detached houses within an LSOA, for England and Wales	🟩	🟥	🟩
12	Percentage of detached houses within an LSOA, for England and Wales	🟩	🟥	🟩
13	Percentage of annexes within an LSOA, for England and Wales	🟩	🟥	🟩
14	Percentage of bungalows within an LSOA, for England and Wales	🟩	🟥	🟩
15	Percentage of houses within as LSOA, for Scotland	🟥	🟩	🟥
16	Percentage of semi-detached houses within as LSOA, for Scotland	🟥	🟩	🟥
17	Percentage of detached houses within as LSOA, for Scotland	🟥	🟩	🟥
18	Percentage of flats houses within as LSOA, for Scotland	🟥	🟩	🟥
19	Percentage of households in fuel poverty, within England	🟩	🟥	🟥
20	Crime Rate at Local Authority Level - Dishonest Use of Electricity	🟩	🟥	🟩
21	Crime Rate at Local Authority Level - Forgery or use of false drug prescription	🟩	🟥	🟩
22	Crime Rate at Local Authority Level - Fraud Forgery etc associated with vehicle or driver records	🟩	🟥	🟩
23	Crime Rate at Local Authority Level - Interfering with a motor vehicle	🟩	🟥	🟩
24	Crime Rate at Local Authority Level -Making, supplying or possessing articles for use in fraud	🟩	🟥	🟩
25	Crime Rate at Local Authority Level - Other drug offences	🟩	🟥	🟩
26	Crime Rate at Local Authority Level - Other forgery	🟩	🟥	🟩
27	Crime Rate at Local Authority Level - Possession of controlled drugs inc. cannabis	🟩	🟥	🟩
28	Crime Rate at Local Authority Level - Possession of controlled drugs exc. cannabis	🟩	🟥	🟩
29	Crime Rate at Local Authority Level - Possession of False documents	🟩	🟥	🟩
30	Crime Rate at Local Authority Level - Profiting from or concealing knowledge of the proceeds of crime	🟩	🟥	🟩
31	Crime Rate at Local Authority Level - Theft from automatic machine or meter	🟩	🟥	🟩

4.4. HYPOTHESIS CONCLUSIONS

Through the hypothesis testing, there were some clear conclusions able to be drawn whilst others were not able to be concluded on. Here are some of the interesting outcomes determined through this process:

1. Different suppliers have a differing mix of source lead associated to the investigations they carry out regarding energy theft.
2. There are more investigations carried out on those with prepaid meters than those with credit meters and there is a higher proportion of confirmed thefts for those with prepaid meters.
3. There are a greater proportion of investigations occurring in urban areas across GB. However, in England and Wales, more theft is confirmed in urban areas whilst in Scotland similar levels of theft are confirmed across the differing urban/rural classifications.
4. There are a greater number of investigations carried out within deprived areas and a greater proportion of those are confirmed to be theft than in non-deprived areas.
5. There is a greater number of investigations carried out in areas with higher fuel poverty, however only residential properties have a higher rate of confirmed theft in areas of higher fuel poverty.
6. For commercial properties, there is a correlation between investigations and crime levels by Police Force Area.

Conclusions 4 and 5 are particularly pertinent as this may be an indicator of those who are under financial pressure are more willing to take the risk in committing energy theft. Under a current cost of living crisis this is particularly pertinent.

5. PREDICTIVE MODELLING

This part of the report outlines the steps taken to build the model and then how we applied it for the predictive modelling to estimate a range in the potential volume of energy theft occurring within the wider GB population. The assessment of the model is discussed in terms of its statistical indicators and actions taken to ensure the best estimation with this data.

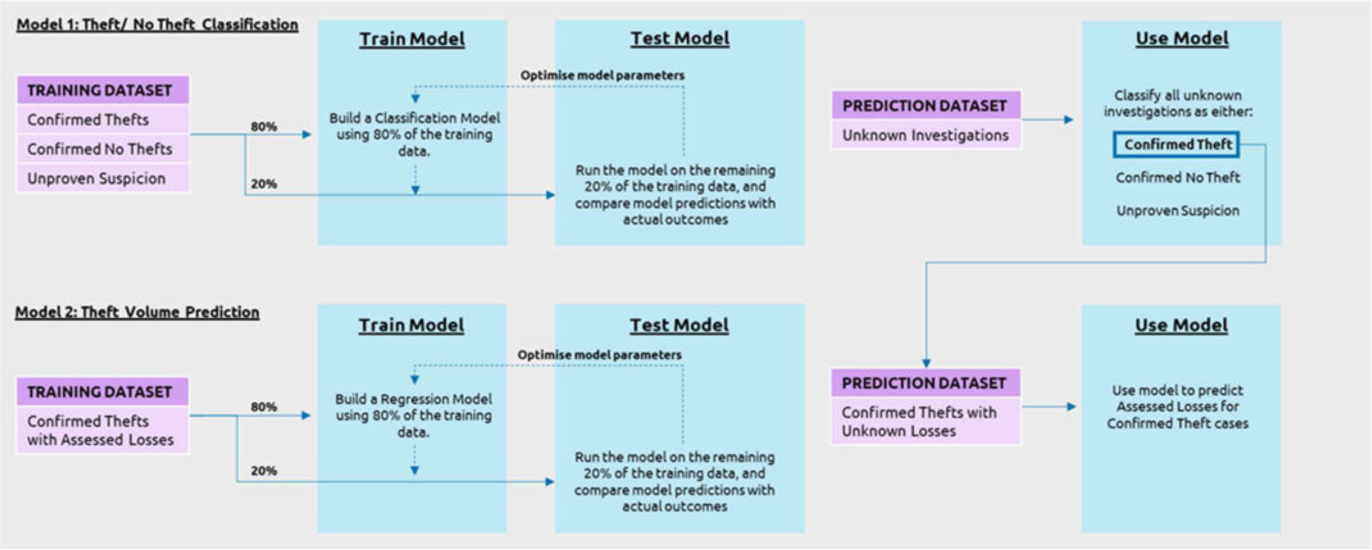
5.1. MODEL DEVELOPMENT

The approach taken here was two-fold:

1. Build and train a *classification model* that uses the predictors from known theft outcomes to classify other potential or unknown cases of theft as either 'Confirmed Thefts', 'No Thefts' or 'Unproven Suspicion of Thefts'.
2. Build and train a *regression model* that uses the predictors from 'Confirmed Theft' and 'Unproven Suspicion' instances with a valid associated theft value, to predict the volume of energy theft in 'Confirmed' or 'Unproven Suspicion' instances identified from the classification model.

Since these models are built using data with known outcomes, they both fall into the category of supervised machine learning. As such, whilst the models serve different purposes, their general structure is very similar. A high-level overview of this structure is shown in Figure 21, and each of the modelling steps are explained in more depth in the following pages.

Figure 21: High level Predictive Modelling Overview


Model 1: Theft/ No Theft Classification
TRAINING DATASET
Confirmed Thefts
Confirmed No Thefts
Unproven Suspicion
80%
20%
Train Model
Build a Classification Model using 80% of the training data.
Test Model
Optimise model parameters
Run the model on the remaining 20% of the training data, and compare model predictions with actual outcomes
PREDICTION DATASET
Unknown Investigations
Use Model
Classify all unknown investigations as either:
Confirmed Theft
Confirmed No Theft
Unproven Suspicion
Model 2: Theft Volume Prediction
TRAINING DATASET
Confirmed Thefts with Assessed Losses
80%
20%
Train Model
Build a Regression Model using 80% of the training data.
Test Model
Optimise model parameters
Run the model on the remaining 20% of the training data, and compare model predictions with actual outcomes
PREDICTION DATASET
Confirmed Thefts with Unknown Losses
Use Model
Use model to predict Assessed Losses for Confirmed Theft cases

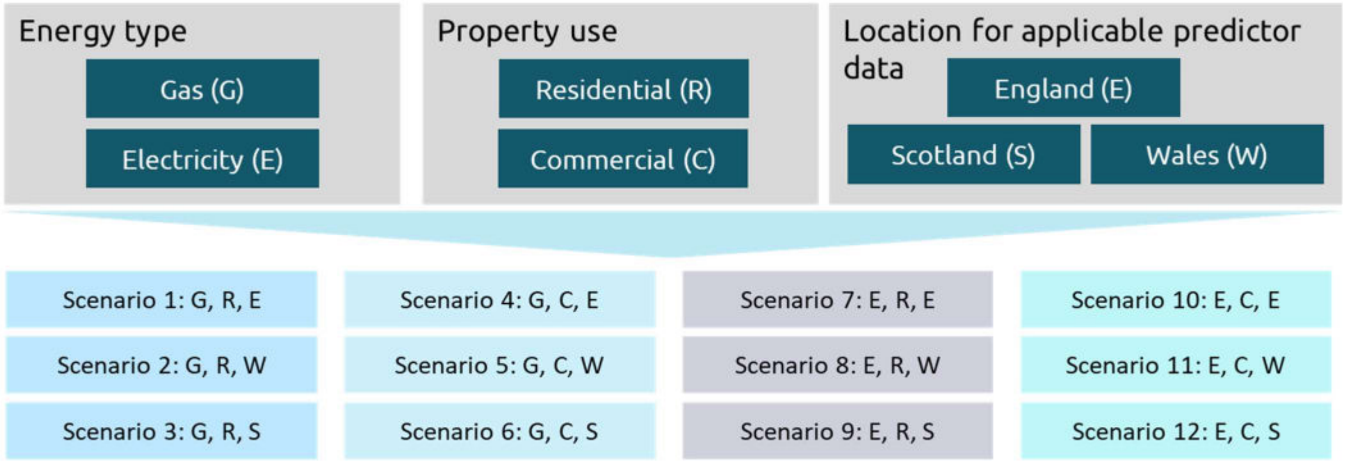
STEP 1: PRE-PROCESSING

The data from this analysis was imported directly into Azure Databricks from the SQL database using Java Database Connectivity (JDBC) within a Python notebook environment. The tables of relevance to this analysis include:

- Regression Table (SQL table name: regression.regression_2022_11_14): this contains predictors, investigation outcome and assessed losses (in the case of confirmed thefts).
- Predictor Table (SQL table name: regression.prediction_2022_11_14_thinned_to_5_percent): this contains predictor data for a 5% sample (~3 million records) of the GB-wide energy consumption data extracted from the TRAS database. Sampling was undertaken at an LSOA level to ensure the sample is representative of the wider GB population.

Through discussions with SMEs, combined with hypothesis testing, it was clear that each country (England, Scotland and Wales), building type (residential and commercial) and fuel type (electricity and gas) generated very different patterns of energy theft. It was therefore deemed appropriate to split the data into 12 distinct datasets and build a unique model to capture the specific nuances for each scenario. The 12 scenarios are outlined in Figure 22 below. In total, this resulted in the training and testing of 24 separate models; 12 classification models (1 per scenario) and 12 regression models (1 per scenario).

Figure 22: 12 distinct model scenarios used to train each classification and regression model

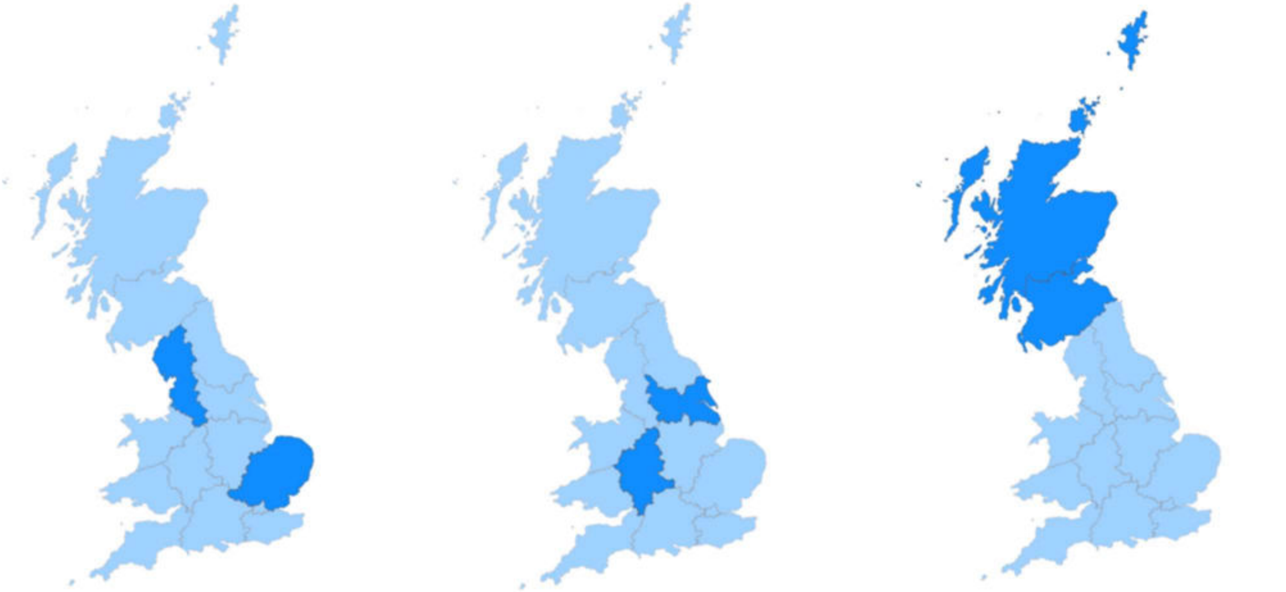


In addition, the accurate reporting of theft investigations was found to be dependent on the energy supplier and geographical location of the theft instance. To limit biases in the classification model, “reliable regions” were selected for each country. “Reliable regions” were defined at Grid Supply Point (GSP) level and aimed to capture regions that met the following criteria:

- The conversion rate of suspected investigations to confirmed cases was around 20% (in line with SME advice)
- The conversion rate was relatively consistent across all suppliers within the region
- There was a good balance of urban and rural geographies, and
- There were comparatively few “Under Investigation” cases

Based on this, the reliable regions selected are shown in Figure 23. Due to differences in the Census data available, the reliable regions used for England and Wales could not be used in Scotland. Scotland therefore uses regions that fall within its own boundary only.

Figure 23: Summary of reliable regions



Regions A and G were selected for electricity properties in England and Wales

Regions E and M were selected for gas properties in England and Wales

Regions P and N were selected for gas and electricity properties in Scotland

The use of reliable regions was applied to the Classification model only. For the Regression model, any confirmed theft case with a valid energy loss value (i.e. greater than zero and not NULL) was used to build the model. The reasoning behind this was because the quality of reporting a value for energy theft was assumed to be equal across all confirmed theft cases. Furthermore, by using all regions, this provided a larger training set to build the most robust regression model possible.

The factors influencing the building of each model were therefore:

- Classification Model: Fuel type, Building Type and Reliable Region
- Regression Model: Fuel Type and Building Type

As such, some models were applicable for multiple scenarios as shown in Table 9 below. This summarises the number of valid data points used to build each of the classification and regression models for each scenario.

Table 9: Summary of valid data points used to train and test each model scenario

Model Scenario	**Train/Test Data for Classification Model**			**Train/Test Data for Regression Model**
	No. Confirmed Thefts	No. No Thefts	No. Unproven suspicions of theft	No. Confirmed Theft cases with an associated value of theft
Scenario 1 (G, R, E)	1,622	4,999	552	5,126
Scenario 2 (G, R, W)				
Scenario 3 (G, R, S)	552	1710	103	
Scenario 4 (G, C, E)	164	2,136	26	557
Scenario 5 (G, C, W)				
Scenario 6 (G, C, S)	35	409	14	
Scenario 7 (E, R, E)	8,086	13,932	1,450	23,905
Scenario 8 (E, R, W)				

Scenario 9 (E, R, S)	5720	5394	3521	
Scenario 10 (E, C, E)	839	3070	215	2,188
Scenario 11 (E, C, W)				
Scenario 12 (E, C, S)	210	687	69	

STEP 2: MODEL SETUP

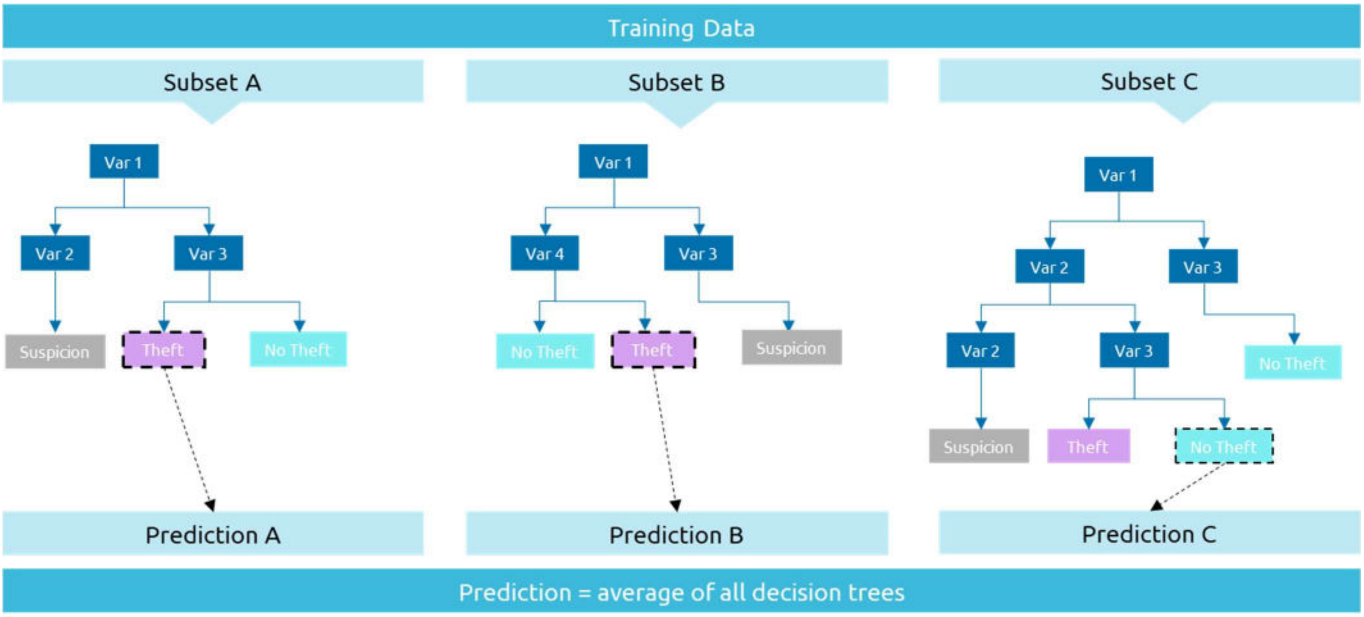
To setup each model, the categorical variables were first converted into a numeric data type by assigning each distinct value within a category field a unique integer value. This is necessary to enable the machine learning algorithm to ingest the data. Each dataset was then randomly split into a training and test dataset in the ratio of 80:20. The purpose of this is to enable the model to 'learn' from 80% of the data and then test its performance by comparing the predicted model outcome with the known outcome, using the remaining 20% of the data.

STEP 3: TRAIN MODEL

A Random Forest algorithm was chosen for this analysis. A Random Forest model[3] can take a classification or regression form, and so the algorithm was adapted to suit both our use cases.

In summary, a Random Forest Model builds multiple classification/regression decision trees, each of which predict an outcome based on a subset of the full training dataset. The final result is taken as an average of all the individual predictions. A simplified example is illustrated in the diagram below.

Figure 24: Random Forest algorithm overview



[3] Random forest - Wikipedia

The main benefits of adopting this approach include:

- A higher level of accuracy compared to other machine learning algorithms,
- Lower risk of over-fitting to the data due to the large number of trees and random sampling process,
- Robustness to outliers,
- Ability to handle a large number of predictors well,
- Efficiency when running on large datasets, and
- Ability to fine tune model parameters to optimise model performance.

As with all machine learning models, there are some limitations to consider. In the case of a Random Forest these include:

- Reduced transparency into the model details since the final outcome is averaged across all decision trees,
- Inability to extract the significance of each feature since this varies across each decision tree.

On balance, a Random Forest approach was considered to be the most suitable for our use case.

After training an initial model, a hyper-parameter tuning process was used to optimise performance. This was undertaken using an iterative 5-fold cross-validation approach, testing different combinations of the following to achieve the highest accuracy score when evaluating the model on the test dataset:

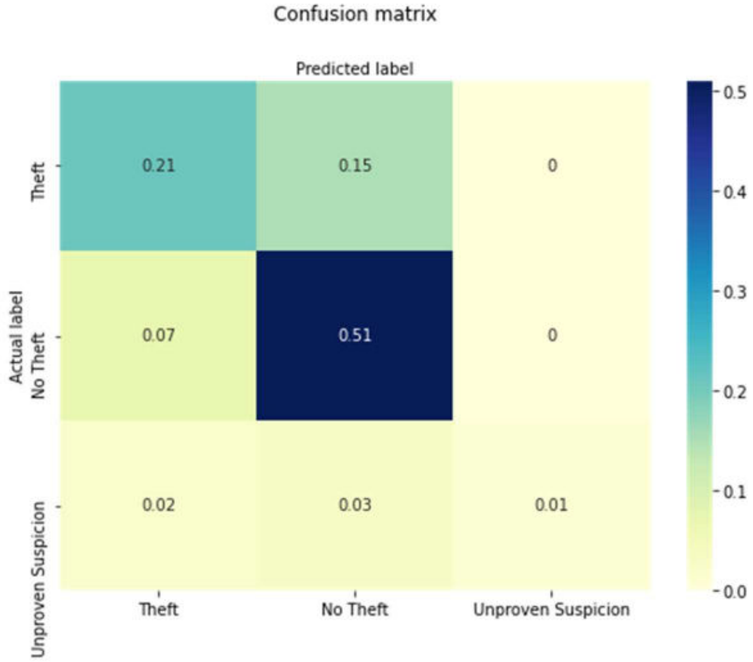
- The total number of decision trees
- The maximum number of features within each decision tree
- The maximum depth of each decision tree

STEP 4: TEST MODEL

All models were run using their respective test datasets and the predicted values were compared against the known results.

For the classification model, a confusion matrix was used to evaluate performance. A confusion matrix summarises the proportion of correctly and incorrectly classified investigations. The actual or 'true' outcome of the investigation is shown on the y axis, and the predicted outcome is shown on the x axis. An example of the confusion matrix for Scenario 7 and 8 is shown below in Figure 25, and a full summary of the confusion matrices for each model build can be found in Appendix D (Chapter 11). The overall accuracy score can be determined by summing the correctly predicted proportion of results, using the values on the diagonal from top left to bottom right. In total this shows that for Scenarios 7 and 8, 73.4% of the test data was correctly predicted using the model.

Figure 25: Example confusion matrix for scenarios 7 and 8 classification model



For the regression model, an R^2, Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) metric was calculated for each model. In the context of our analysis, these metrics are defined as:

- R^2 - a statistical measure of fit that indicates how much variation in the volumes of energy theft are explained by the predictor variables. Scores typically range between 0 and 1, where 1 indicates a perfectly performing model and 0 means it does not perform the prediction well. Negative numbers indicate it is no better than taking an average.
- MSE - the average squared difference between the estimated and actual value of energy theft. A smaller MSE indicates a better performing model.
- RMSE – the square root of the MSE. This is widely considered an excellent general purpose error metric for numerical predictions and as with the MSE, a smaller value indicates a better performing model.

Table 10 below summarises the performance metrics used to evaluate each scenario, for both the classification and regression models.

Table 10: Summary of model performance metrics

Model Scenario	Classification Model Performance Metrics	Regression Model Performance Metrics		
	Accuracy Score (%)	R^2	MSE (kWh)	RMSE (kWh)
Scenario 1 (G, R, E)	75.8	0.035	1,547,197,731	39,334
Scenario 2 (G, R, W)				
Scenario 3 (G, R, S)	76.2			
Scenario 4 (G, C, E)	90.0	-0.220	25,017,648,508	158,170
Scenario 5 (G, C, W)				
Scenario 6 (G, C, S)	93.0			
Scenario 7 (E, R, E)	73.4	-0.018	69,643,830,330	263,901
Scenario 8 (E, R, W)				

Scenario 9 (E, R, S)	68.0			
Scenario 10 (E, C, E)	79.5	-0.005	61,565,139,027	248,123
Scenario 11 (E, C, W)				
Scenario 12 (E, C, S)	76.4			

Overall, the accuracy scores for the classification models are reasonable. Based on the confusion matrices, the models do tend to overfit to the number of 'No Theft' cases given the bias towards this investigation outcome in the training datasets. However, since the training dataset is likely to contain a higher proportion of 'confirmed' and 'unproven suspicion' cases of theft than what would likely occur in reality, this is an acceptable trade-off, and the model seems appropriate to use on the wider GB population dataset.

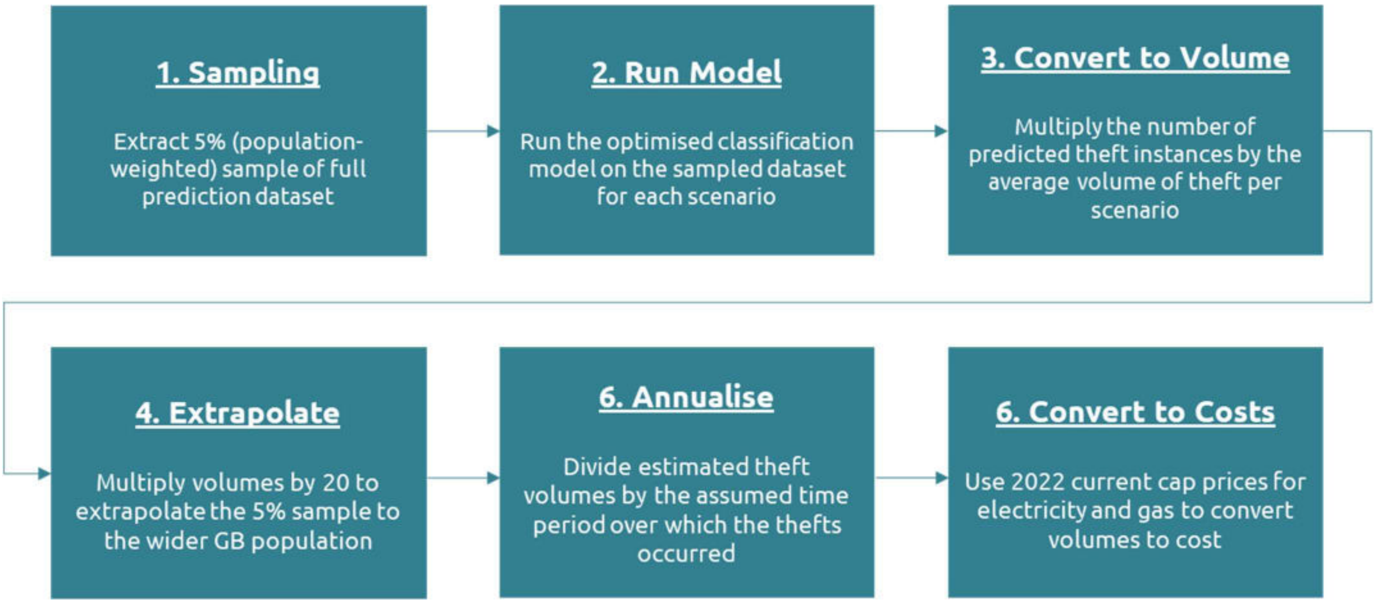
The performance metrics from the regression model on the other hand are very poor. The low R^2 values and high MSE/RMSE values, suggest that model performs no better than taking average volumes of theft across each scenario. This performance was in part expected since this analysis was limited to largely spatial datasets containing predictors which have limited/no correlation with volumes of energy theft. For this type of model to be effective, customer specific data will likely be needed.

Given the respective model performances and the current limitation of suitable predictor data for the regression model, the most appropriate methodology to pursue is to use the classification model to identify instances of theft for each scenario and multiply this by the respective average theft volume. The average theft was calculated at GSP level for each scenario, thereby generating the most appropriate average per each building type, fuel type and spatial location.

In addition to the model performance, we can extract the relative importance of each variable in predicting the final outcome which can be found in Appendices D and E for the classification and regression model respectively. In general, the supplier, normal payment method, meter type, meter location and the IMD ranking tend to be the most important features in classifying each investigation, which aligns with the findings from the hypothesis testing and the insights gained from discussions with SME's. The importance of these features tended to be greater amongst the larger datasets which are more likely to give reliable results since they are less skewed by outliers and less prone to overfitting.

5.2. PREDICTIONS

This section discusses how the models were used to run predictions using the wider GB energy consumption data, extracted from TRAS. The diagram below gives a high-level overview of the prediction process.



1. Sampling

Extract 5% (population-weighted) sample of full prediction dataset

2. Run Model

Run the optimised classification model on the sampled dataset for each scenario

3. Convert to Volume

Multiply the number of predicted theft instances by the average volume of theft per scenario

4. Extrapolate

Multiply volumes by 20 to extrapolate the 5% sample to the wider GB population

6. Annualise

Divide estimated theft volumes by the assumed time period over which the thefts occurred

6. Convert to Costs

Use 2022 current cap prices for electricity and gas to convert volumes to cost

The 'Predictor' data table was imported directly from SQL using Java Database Connectivity (JDBC) within a Python notebook environment. As previously outlined, the Predictor dataset contains a 5% sample of the total consumption data. It was ensured that an even distribution of data points within each LSOA were retained as part of the sampling process to keep it representative of the wider GB population. This reduces the size of the dataset from approximately 60 million data points to 3 million, thereby balancing accuracy with computational efficiency. The results obtained from this predictor dataset were then linearly extrapolated to estimate a total volume of energy theft (equivalent to 100% of the population).

Note that this estimate has no time series element, and therefore each instance of theft cannot be allocated to a particular year. For this reason, the following tables present the *total* number of theft instances and *total* volume of predicted losses. Table 11 presents the total number of No Theft, Confirmed Theft and Unproven Suspicion of Theft instances as predicted by the model, and for reference, includes the equivalent metrics for the TRAS data.

Table 11: Number of 'No Thefts', 'Confirmed Thefts' and 'Unproven Suspicions of Theft' in TRAS, and estimated by the model

			No Theft	Confirmed	Unproven Suspicion	Total
Electricity	Commercial	TRAS	10,088	3,745	659	14,492
		Model Predictions	4,416,300	43,720	640	4,460,660
	Residential	TRAS	69,405	47,497	17,980	134,882
		Model Predictions	33,470,580	641,520	80	34,112,180
Gas	Commercial	TRAS	4,896	856	34	5,786
		Model Predictions	14,984	1,980	-	16,964
	Residential	TRAS	25,959	9,778	5,398	41,135
		Model Predictions	19,358,880	114,340	40	19,473,260

To summarise the confirmed thefts cases, the model predicts the following:

- 43,720 electricity thefts instances in commercial buildings
- 641,520 electricity theft instances in residential buildings
- 1,980 gas theft instances in commercial buildings
- 114,340 theft instances in residential buildings

Table 12 presents this data as a proportion of each respective population.

Table 12: Proportion of 'No Thefts', 'Confirmed Thefts' and 'Unproven Suspicions of Theft' in TRAS, and estimated by the model

			No Theft	Confirmed	Unproven Suspicion	Total
Electricity	Commercial	TRAS	70%	26%	5%	100%
		Model Predictions	99%	1%	0%	100%
	Residential	TRAS	51%	35%	13%	100%
		Model Predictions	98%	2%	0%	100%
Gas	Commercial	TRAS	85%	15%	1%	100%
		Model Predictions	88%	12%	0%	100%
	Residential	TRAS	63%	24%	13%	100%
		Model Predictions	99%	1%	0%	100%

This shows that for each fuel type and building type, the proportion of predicted confirmed theft cases is lower than the proportion observed within TRAS. This is reasonable to expect since all TRAS entries are suspected thefts and will therefore inherently contain a higher proportion of confirmed cases than what occurs in reality.

Looking at the model predictions, aside from the commercial gas thefts, where 12% of cases are predicted as confirmed thefts, the remaining categories have an estimated 1-2% proportion of confirmed thefts. This suggests some overfitting in the model towards 'No Theft' cases, and that the subsequent volume and cost estimate are likely to be conservative. It should also be noted that the total volume of commercial gas instances reported by the model is lower than expected which may explain why the proportions are so difference for this grouping.

With this in mind, Table 13 presents the volume of energy theft, per fuel type and building type, in TRAS and as estimated by the model, in GWh

Table 13: Total estimated volume of energy theft in TRAS and as predicted by the model, in Gwh

			Confirmed	Unproven Suspicion	Total
Electricity	Commercial	TRAS	164.4	1.4	165.9
		Model Predictions	2,473.7	24.4	2,498.1
	Residential	TRAS	309.3	25.6	334.9
		Model Predictions	6,008.9	4.4	6,013.4
Gas	Commercial	TRAS	88.0	0.4	88.4
		Model Predictions	217.8	0.0	217.8
	Residential	TRAS	174.5	2.8	177.2
		Model Predictions	2,957.7	1.3	2,959.0

To summarise the total energy losses predicted by the model:

- 2500 GWh of electricity theft in commercial buildings
- 6000 GWh of electricity theft in residential buildings
- 200 GWh of gas theft in commercial buildings
- 3000 GWh of gas theft in residential buildings

To annualise these results, the total volume of energy theft has been divided by a range of time periods (annualisation factors) over which these thefts could occur. Since TRAS operates over a 7-year period.

Figure 26 (electricity) and Figure 27 (gas) assume a range in annualisation factors of between 1 and 7 years. This data is also presented in

Table 14.

Figure 26: Estimated Range of Electricity Theft (10^6 kWh), based on different annualisation factors

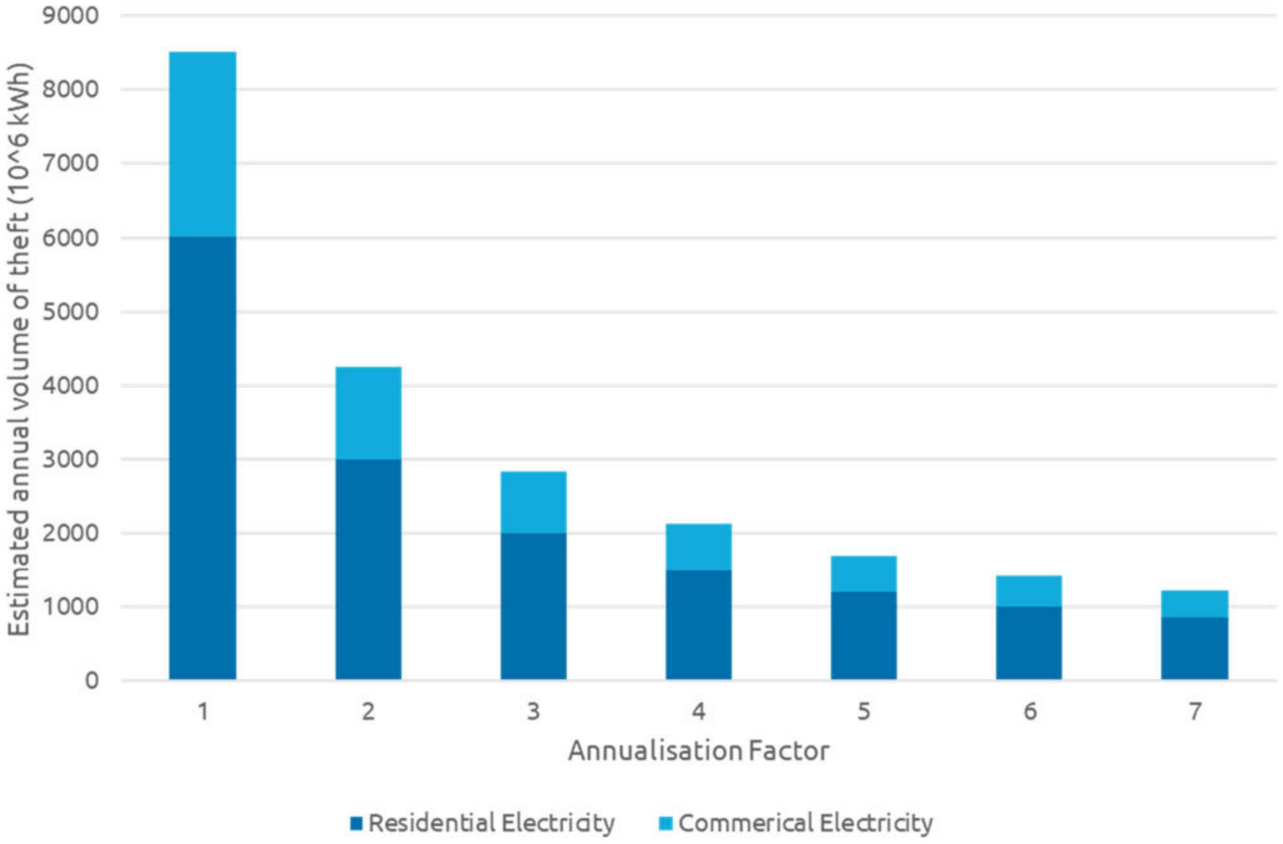


Figure 27: Estimated Range of Gas Theft (10^6 kWh), based on different annualisation factors

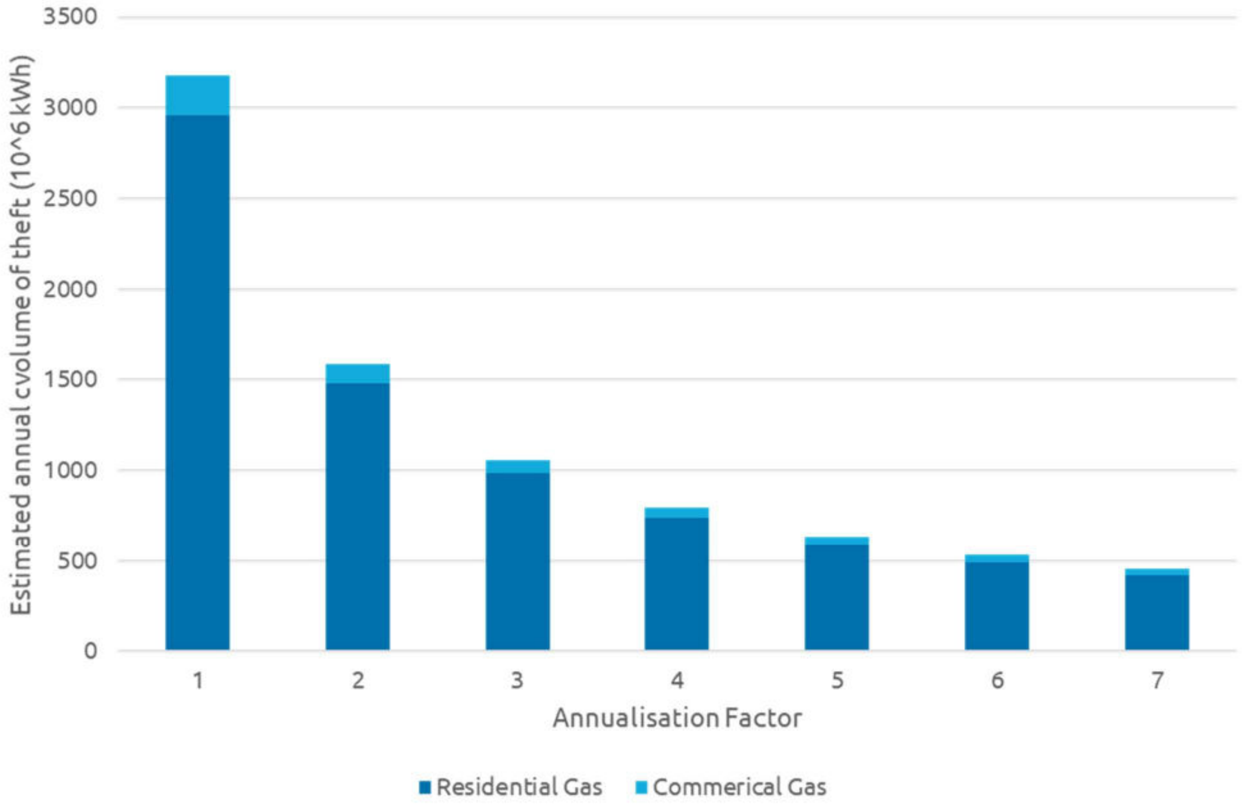


Table 14: Estimated range in the volume of energy theft predicted by the model, assuming varying annualisation factors (years over which the predicted thefts occurred), in GWh

Fuel Type	Building Type	Annualisation factor						
		1	2	3	4	5	6	7
Electricity	Residential	6013	3007	2004	1503	1203	1002	859
	Commercial	2498	1249	833	625	500	416	357
Gas	Residential	2959	1480	986	740	592	493	423
	Commercial	218	109	73	54	44	36	31

These predictions show that if all thefts predicted by the model occurred within one year, the estimated total annual volume of electricity and gas theft would be 8.5 TWh and 3.2 TWh respectively. Equally, if these thefts were to occur over a 7-year period, the estimated total annual volume of electricity and gas theft would be 1.2 TWh and 0.45 TWh respectively.

5.3. EVALUATION AND JUSTIFICATION

To validate the predicted volumes of gas and electricity theft, the model outcomes were aggregated at a national level and compared against GB-wide energy network distribution data provided by Xoserve (for gas) and Elexon (for electricity). The goal of this exercise was to compare the annual volume of predicted gas/electricity theft against the total annual volume of unallocated gas/electricity within the wider GB network and ensure that the predicted theft volumes were:

i) Lower than the total unallocated volumes, and
ii) Of an appropriate order of magnitude compared to total unallocated volumes.

This section describes the approach taken for both gas and electricity. It was initially intended to compare energy volumes at an LDZ/GSP level, however due to limitations in the data this was not possible. A national level approach was therefore taken, which still provides overall surety that the estimated theft ranges are sensible.

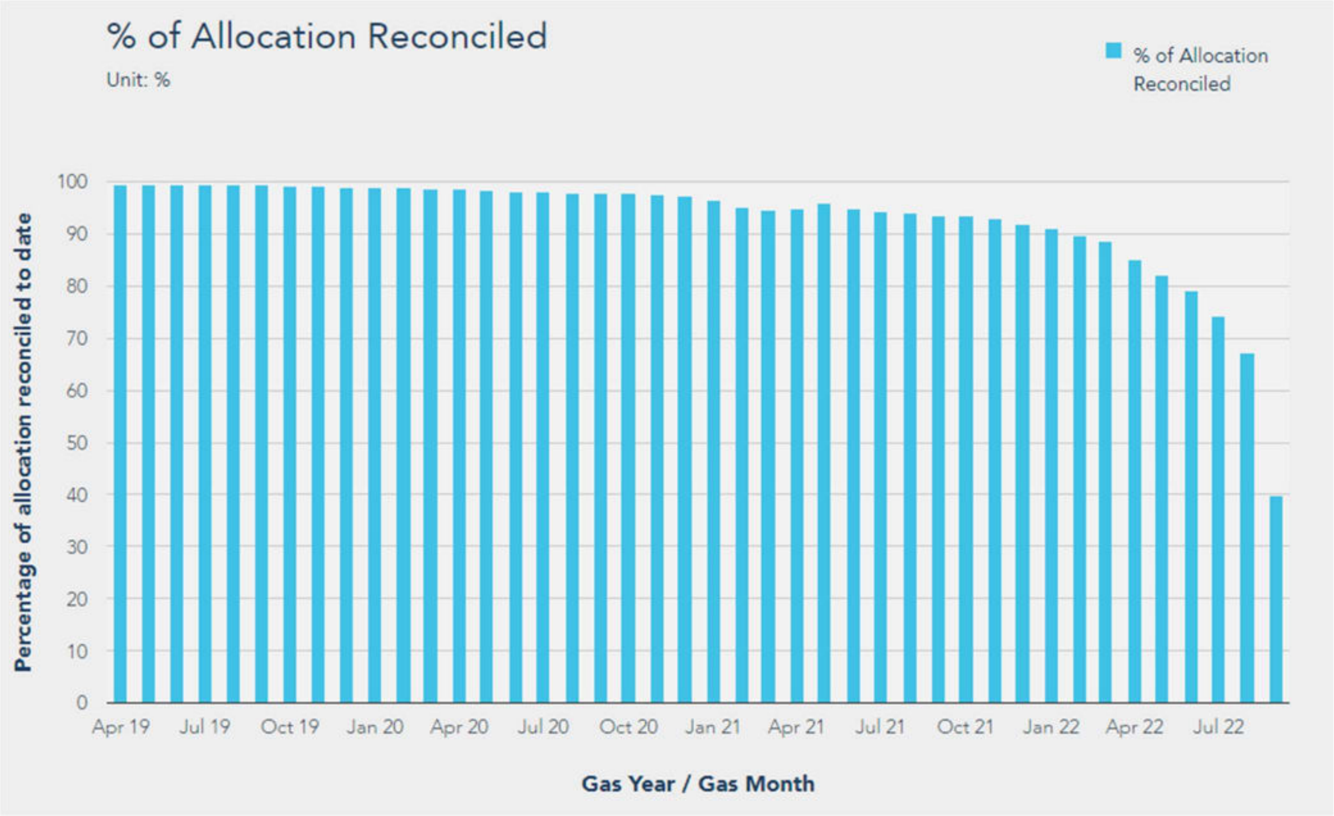
5.3.1. GAS RECONCILIATION

The primary data source used for this exercise was from National Grid's Data Item Explorer (see reference link in Chapter 13). The following datasets were extracted from this portal:

- *Demand Actual, D+6, per LDZ* - this gives the daily total measured volume of gas (in millions of m3) supplied to each LDZ, recorded 6 days after the settlement date and includes both daily metered, and non-daily metered properties.
- Shrinkage, NTS Shrinkage Factor – this gives the proportion of daily gas lost due to shrinkage.

These two datasets were combined with the data shown in Figure 28 below, a chart published by Xoserve showing the proportion of unallocated gas (out of the total volume of gas supplied) at a national level. For gas, reconciliation can take up to 3 years after the initial supply date before the volume of unallocated gas is revealed, and for this reason, the earliest year of available data was taken (April 2019 to March 2020) to ensure reconciliation was complete, or as near to completion as possible.

Figure 28: Percentage of gas allocation reconciled to date (source: https://www.xoserve.com/uig-charts/uig-as-of-total-throughput/)



The process taken to calculate an annual volume of unallocated gas was as follows, and the outcome of this is detailed in Table 15:

1. Factor the daily gas demand by the shrinkage factor to remove the proportion of gas lost within the transportation network
2. Aggregate the factored daily gas demand at a monthly level, and extract data for the 12-month period between April 2019 and March 2020
3. Multiply the monthly total volumes of gas supplied (excluding shrinkage) by the proportion of unallocated gas reconciled to date
4. Convert gas volumes in millions of cubic metres, to kWh using the formula below:

$$kWh = m3 \times Calorific\,Value \times 1.02264 \div 3.6$$

where:
Caloric value = 40.0 (this can deviate by +/- 5% depending on the quality of natural gas)
Correction factor = 1.02264 is a correction factor,
Conversion factor = 3.6.

5. Sum the total volume of unallocated gas to determine an annual volume

Table 15: Total Volume of Unallocated gas, estimated using National Grid and Xoserve data

	Gas Demand exc. Shrinkage (millions m3)	% Allocated Gas	Total Volume Unallocated (millions m3)	Total Volume Unallocated (kWh)
Apr-19	8258.5	0.995	41.3	469,193,149
May-19	6365.2	0.995	31.8	361,630,053

Jun-19	4746.1	0.9949	24.2	275,032,149
Jul-19	3436.4	0.9945	18.9	214,757,522
Aug-19	3632.7	0.9942	21.1	239,408,694
Sep-19	4374.3	0.9933	29.3	333,013,545
Oct-19	6620.5	0.9923	51.0	579,244,677
Nov-19	5845.9	0.9913	50.9	577,892,903
Dec-19	6413.8	0.9901	63.5	721,489,419
Jan-20	6432.9	0.9889	71.4	811,355,654
Feb-20	6445.5	0.9883	75.4	856,889,801
Mar-20	7587.2	0.9874	95.6	1,086,249,484
Totals	**70159.0**	**-**	**574.4**	**6,526,157,049**

This gives a total volume of unallocated gas, excluding shrinkage, of 6.53 GWh. With an estimated range of gas theft of between 3.2 GWh and 0.45 GWh depending on the annualisation factors used, this suggests that gas theft accounts for between 7-50% of the total volume of unallocated gas.

It should be noted that Shrinkage includes 0.2% of throughput for theft of gas upstream of the meter which equates to approx. 4% of shrinkage. This allocation for theft of gas is not included within the unallocated gas figure shown above. This theft of gas allocation within shrinkage would include theft in conveyance which is not being estimated by the methodology presented.

5.3.2. ELECTRICITY RECONCILIATION

In response to a FOI request to Ofgem submitted by Jon Dixon on 12th May 2022, the data presented in Table 16 was provided which details out total annual losses of electricity between 2016 and 2021 in Great Britain.

Table 16: Total annual losses of electricity between 2016 and 2021 in Great Britain

Year	2016	2017	2018	2019	2020	2021	Average
Total GB Electricity Losses (GWh)	132,806	135,374	134,045	124,709	125,925	129,543	130,400

The total losses are a measure of the difference between units entering and exiting the DNO network through different connection points. It is noted that these totals include both technical and non-technical losses. Since theft forms part of the non-technical losses, it would be useful to compare this proportion only with the predicted electricity theft estimates. However, due to the time constraints of this project, it has not been possible to find methods to accurately split out these categories of losses.

Nonetheless, a comparison between annual electricity losses predicted in the model, which ranges between 1,200 and 8,500 GWh (depending on the annualisation factor used), is lower, but of similar magnitude to the total losses provided by Ofgem. Based on the average annual electricity loss of 130,400 GWh, this analysis suggests that theft accounts for between 1-6.5% of the total volume of unallocated electricity.

5.4. SUMMARY

To conclude, the model development and assessment led to the inclusion of the classification model only based on the outcome of the statistical tests. The classification models achieved an accuracy score of between 68% and 93% for each scenario with an average of 79%, which was deemed reasonable. The classification model was used to identify instances of theft on the sample prediction dataset, which represented a 5% sample of the population. In lieu of the regression model, the average assessed loss for each scenario was calculated and used to convert theft predictions to an energy volume. This was then extrapolated for the total population. To convert this to an annual volume, a range of annualisation factors were used to provide a range of theft volume estimates.

Table 17: Estimated range in the volume of energy theft predicted by the model, assuming varying annualisation factors (years over which the predicted thefts occurred), in 10^6 kWh

Fuel Type	Building Type	Annualisation factor						
		1	2	3	4	5	6	7
Electricity	Residential	6013	3007	2004	1503	1203	1002	859
	Commercial	2498	1249	833	625	500	416	357
Gas	Residential	2959	1480	986	740	592	493	423
	Commercial	218	109	73	54	44	36	31

Finally, the predictions were then validated against the volume of unallocated energy. The data for this was discussed thoroughly with the theft SMEs. Unallocated electricity losses over the last 6 years were considered to have averaged 130,400 GWh whilst unallocated gas, excluding shrinkage, was taken to be 6.53bn kWh for a 1 year period (April 2019 – Mar 2021). This confirmed the estimates provided by the model met the validation criteria of being within the maximum threshold of losses.

6. ESTIMATION

This section of the report will summarise the estimation from the modelling prediction and recap on the limitations to the estimation. It will put into context what that means in terms of revenue lost. It also explains how the losses are annualised.

6.1. MISSED REVENUES

To put into predicted volumes from Section 5.2 into context, it was agreed to attribute a financial value to them. Due to the fluctuating wholesale costs and current challenges of price rises, the use of OFGEM's price caps were considered the most appropriate reference point. Average residential pricing was taken using the Default tariff cap for October 2022 – December 2022. Commercial pricing was taken from the 'Prices of fuels purchased by non-domestic consumers in the United Kingdom (including the Climate Change Levy)' for the 2nd quarter of 2022.

Applying these price caps to the estimated volumes,

Figure *29* and Figure 30 below show the estimated range in annual costs associate with energy theft, assuming different annualisation factors. This is summarised in Table 18.

Figure 29: Estimated Range of annual Electricity Theft (£m), based on different annualisation factors

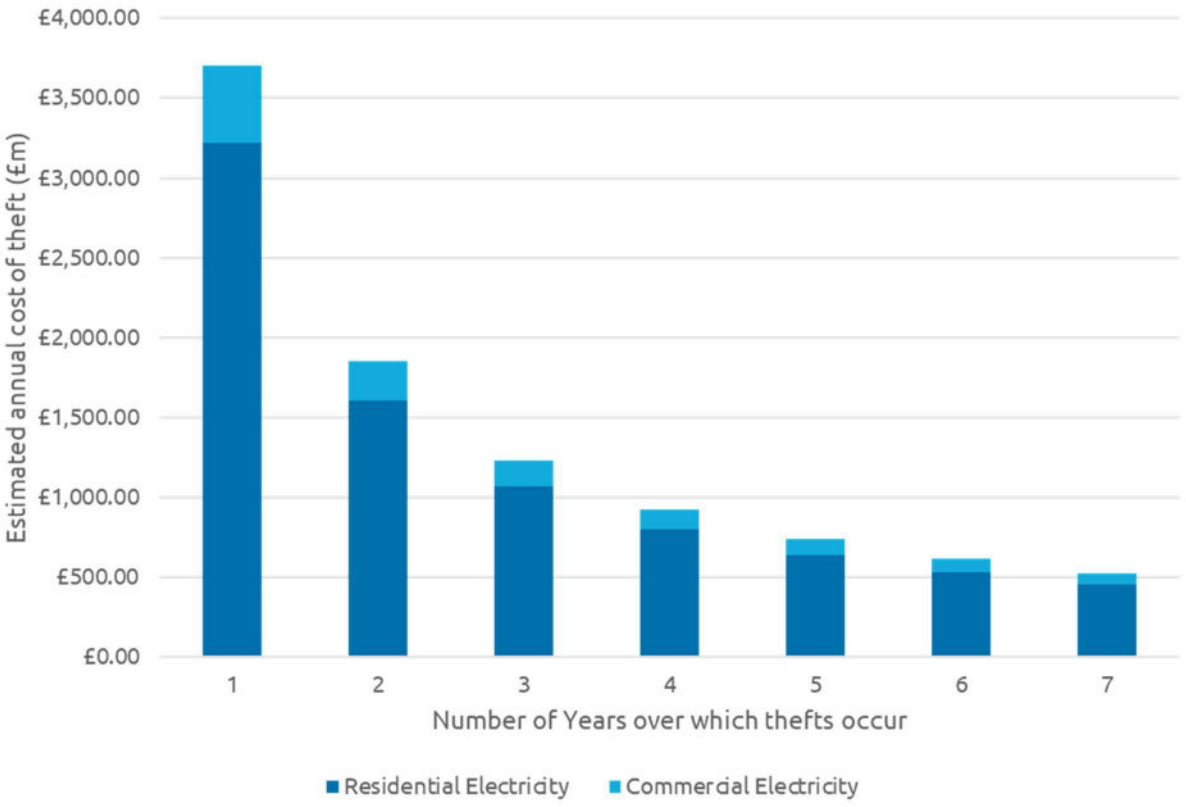


Figure 30: Estimated Range of Gas Theft (£m), based on different annualisation factors

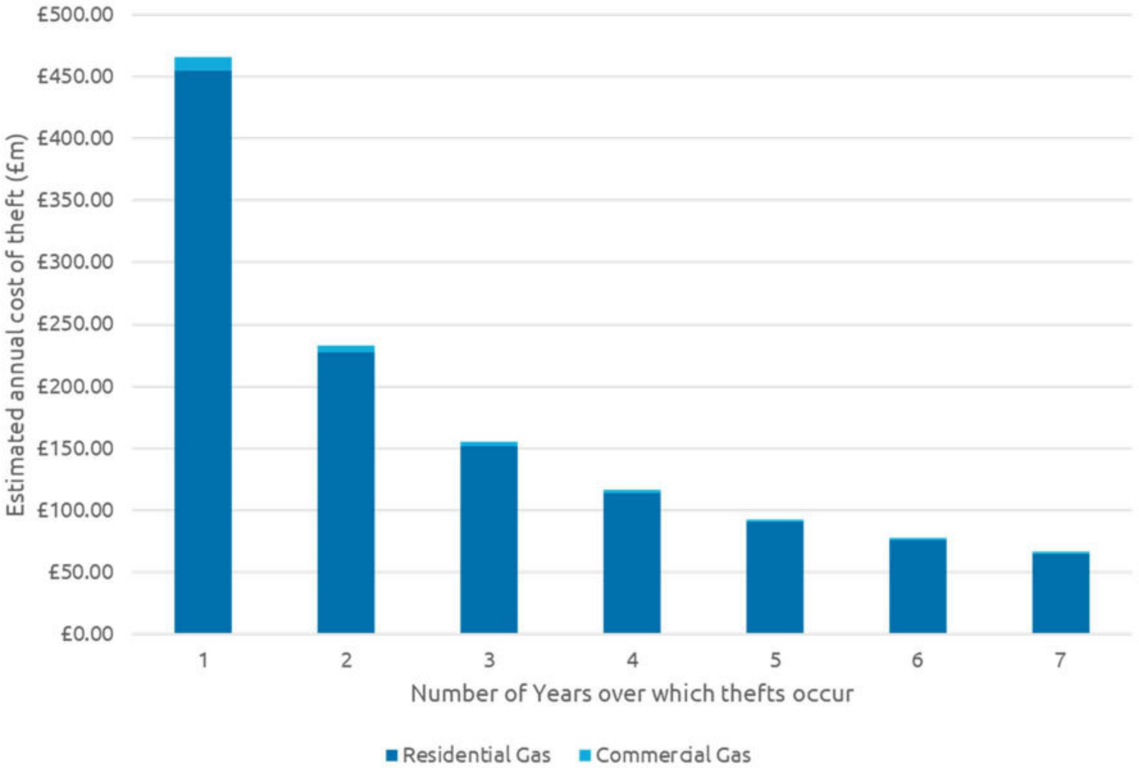


Table 18: Estimated Range of annual Energy Theft (£m), based on different annualisation factors

Fuel Type	Building Type	Annualisation factor						
		1	2	3	4	5	6	7
Electricity	Residential	£3,218.56	£1,609.28	£1,072.85	£804.64	£643.71	£536.43	£459.79
	Commercial	£482.96	£241.48	£160.99	£120.74	£96.59	£80.49	£68.99
Gas	Residential	£454.53	£227.27	£151.51	£113.63	£90.91	£75.76	£64.93
	Commercial	£10.88	£5.44	£3.63	£2.72	£2.18	£1.81	£1.55

These predictions show that if all thefts predicted by the model occurred within one year, the estimated total annual cost of electricity and gas theft would be £3.70bn and £0.47bn respectively. Equally, if these thefts were to occur over a 7-year period, the estimated total annual volume of electricity and gas theft would be £0.53bn and £0.07bn respectively.

Based on the bulk of the theft within the prediction occurring just over a 3/4-year period, it would be expected that the theft identified within the wider population would be spread over a similar timeframe. It would be prudent to attribute the annualisation of the theft to 3 to 5 years providing a range of:

Table 19: Estimated range of annual theft in volume (GWh) and value (£m)

Fuel Type	Building Type	Volume (GWh)	Value
Electricity	Residential	1203 – 2004	£643m - £1072m
	Commercial	500 – 833	£96m - £161m
	TOTAL	**1703 - 2837**	**£737m – £1233m**
Gas	Residential	592 – 986	£91m - £151m
	Commercial	44 - 73	£2.1m – 3.6m
	TOTAL	**636 - 1059**	**£93m - £155m**

Given the accuracy of the model ranging from 68% - 93% with an average of 79%, the range could expect to be ±21%.

6.1.1. LIMITATIONS

There are number of factors which need to be considered when using this estimate.

- The classification accuracy scores ranged from 68% to 93% in testing so there will be some mis classifications within the wider population
- A 5% sample size was used to balance the additional computing needs within the timeboxed exercise
- Without accurate consumption data at a granular level, it is difficult to provide better than an average of volumes lost due to theft
- The dataset contains early unvalidated TRAS data from the early collection period prior to validation checks being implemented which may still be within the dataset
- RPA data contained corrupted data due to MPXN formatting to scientific format, in addition to missing data such as 10% supplier investigation ID which prevented its inclusion within the model
- A short history of time series data with significant influencing factors (covid) meant that time series analysis was not an option at this time
- It is recognised, there are additional attributes which are likely to support a better model such as business type and more specific properties such as hygiene rating
- Generic meter data can only predict so far, and therefore more tailored attributes for meters such as tamper indicators, actual readings could lead to a greater confidence in the model. Indicators such as tamper alerts could hold patterns which indicate a theft is likely to occur
- Without the ability to reconcile the inputs and outputs to the low voltage network it is not possible to identify the theft at a granular level so the reconciliation can only be completed through a top down approach
- This analysis has used the TRAS consumption data as a base reference for the MPRN/MPAN population, however this will exclude those not registered on a network such as conversions to multiple properties, plot to postcode delays for new builds
- Basing the analysis off the properties connected to the network will miss the theft occurring upstream of the meters

6.2. GENERAL TRENDS

It is important to recognise this is a snapshot in time. However, these are fast-moving times and there are challenges which will need to be considered alongside the estimations provided at this time. Some factors which are likely to have an influence on energy theft and changing motivations and the ability to detect it:

- Cost of living pressures
- Bit coin mining
- Cannabis cultivation
- Expansion of EV charges
- Expansion of micro-generation and battery storage

6.3. CONCLUSION

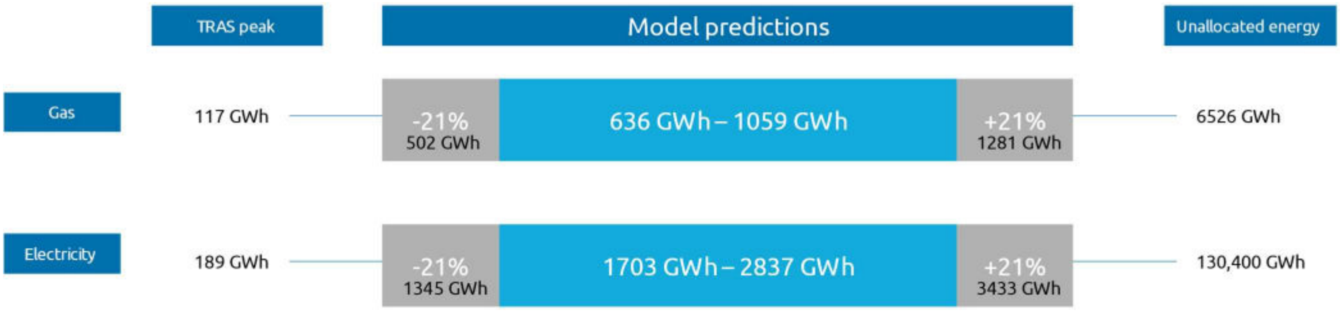
With all the above, the range of losses due to energy theft has been derived.

This equates to:

- Gas losses: 636million kWh to 1059million kWh per year
- Electricity losses: 1703million kWh to 2837million kWh per year

An estimate has been provided with a 21% tolerance. This is considered the best prediction available within the constraints documented within this report. In addition, it does sit within the extreme validation points.

Figure 31: Overview of gas and electricity theft loss estimates within the constraints of peak TRAS losses and unallocated energy volumes



To put this into context, in 2019, 323.8 TWh of electricity was generated within the UK[4].The losses predicted here are just 0.5 – 0.9% of this despite nearly 40% being unallocated. The demand for Gas in 2019 was 859.8 TWh. The losses predicted here make up just 0.1% of this whilst overall losses are around 0.8%.

Despite these losses making up such small proportions of total energy in the UK, taken at OFGEM's capped prices, the losses are worth around £737m - £1233m for electricity and £93m - £155m for gas. The current costs for these losses end up with the paying consumer and equate to around £29 to £48 per household per year.

[4] UK Energy in Brief 2021 (publishing.service.gov.uk).

In terms of carbon equivalent emissions[5], the electricity losses estimated equate to 397,000 $kgCO_2e$ to 661,400 $kgCO_2e$ and carbon equivalent emissions for the gas losses estimated equate to 117,000 $kgCO_2e$ to 194,700 $kgCO_2e$.

[5] Greenhouse gas reporting: conversion factors 2020 - GOV.UK (www.gov.uk)

7. WHAT NEXT

RE-RUNNING THE METHODOLOGY

The data has been stored in a purpose-built Azure environment to enable the re-use of the data with the potential to rerun the model. Datapipes have been established to transform the data as required from raw format to feature in a workable table. These are referenced in Chapter 2. The code has been created and provided to RECCo for the ingestion and transformation of data, as well as the code within Azure data bricks to run the models again. The steps defined in Chapter 5.1 align with the setup of the code to be able to run the analysis. In addition, the data is connected to a Power BI interface within the Virtual Machine which RECCo can access to gain further insight into the theft data.

EXPANDING THE DATA SOURCES TO ENHANCE THE METHODOLOGY

There were several data sources identified through the project which could not be justified within the timeframe. It is hypothesised that these could provide a greater level of robustness or accuracy to the modelling:

- Smart Alerts - Smart Meters send not only meter readings but also alerts about their environment and status, there is potential that these alerts could aid in identifying instances of energy theft.
- Half-Hourly (HH) smart meter data – the roll out of smart meters is leading to increasing amounts of HH data being collected. There is the potential that analysing consumption at this lower level of granularity will lead to greater visibility of consumption anomalies and, as a result, enhance theft detection.
- Feeder Meter Data - Distributors are increasingly rolling out Half Hourly metering at feeder level to aid them in managing their networks. This feeder meter data could be used to compare the volumes going into specific sections of the network against that exiting it, to identify where energy is going unaccounted for which might indicate theft.
- Hygiene rankings for food relating business – there is a view that those with a lower score have an increased likelihood to tamper with their meters. This data is publicly available but was not able to be gathered within the timeframe of the project.
- Business classification – this is available from Company’s House on a case-by-case basis as well as available for purchase. This could be a data source to better classify the model if a suitable route to acquire the data was available.
- If a future ‘TRAS 2’ was to be developed, careful consideration of the data validation and consistency of submissions should be taken to ensure the data can be utilised to strengthen the model.

It is recommended that RECCo explore the opportunity to add additional data sources ahead of any future TEM calculations, to enable the methodology to be enhanced.

8. APPENDIX A – DATA EXTRACTION

Overview of the extractions and manipulations required to create the datasets required for modelling purposes.

EXTRACTING METADATA FOR METERS AND ACCOUNTS FROM TRAS CONSUMPTION DATA

From the TRAS consumption dataset, which is a flat dataset, it was required to extract metadata for meters and accounts in a way that allows a join between each theft investigation to the metadata pertaining to it. This is done as follows.

- The consumption dataset is first broken down into meter points.
- The subset of data for each meter point is broken down into meters.
- The subset of data for each meter is broken down into periods that will be called "meter installations" which start at the meter installation date and ends at the next meter installation date. This is necessary because in the consumption dataset a meter might have multiple installation dates.
- To determine a theft investigation's meter metadata there is a search for the meter point, meter and meter installation period that match the investigation.
- The subset of data for each meter point is also broken down into accounts.
- The subset of data for account is broken down into periods that will be called "account modifications" which start at the account start date and ends at the next account start date. This is necessary because in the consumption dataset an account might have multiple account start dates.
- To determine a theft investigation's account metadata there is a search for the meter point, meter, account, meter installation period and account modification period that match the investigation.

For the approach above, for each meter, the construction of the start date and end date for each meter installation period as follows.

- Each combination of `(mpxn, supply_postcode, meter_serial_number)` is taken to represent a meter.
- If this meter has multiple records with the same installation date, which might be null, these records are compressed into one by sorting the meter types and locations alphabetically and taking the first values.
- If this meter has a record with no installation date, this is taken to represent an installation that started in the year 1900.
- If this meter has a record with an installation date, this is taken to represent an installation that started on this installation date.
- If this meter has another record with a subsequent installation date, this is taken to mean that the installation in the previous step ended on this subsequent installation date.
- If this meter has an installation without an end date, its end date is taken to be the year 2100.

The construction of the start and end dates for each account modification is in a similar way.

CREATING THE PREDICTION DATASET FROM TRAS CONSUMPTION DATA

From the TRAS consumption dataset, which is a flat dataset, a list of meter points was extracted. Each meter point could have multiple meters and multiple accounts, from which a selection is made of a meter and an account for use by the ML. To make predictions about energy theft for each meter point, the ML will use the following predictors:

- The selected meter's metadata, specifically, the meter type and meter location.
- The selected account's metadata, specifically the normal payment method.
- Socio-economic data, such as crime rates and fuel poverty in the area, which is identified using the supply postcode.

TRAS THEFT DATA

TRAS theft data was uploaded to the following tables in the raw data schema

- Commercial theft: `raw_data.tras_comm_tfto_data` (150,000 rows)
- Residential theft: `raw_data.tras_resi_tfto_data` (900,000 rows)

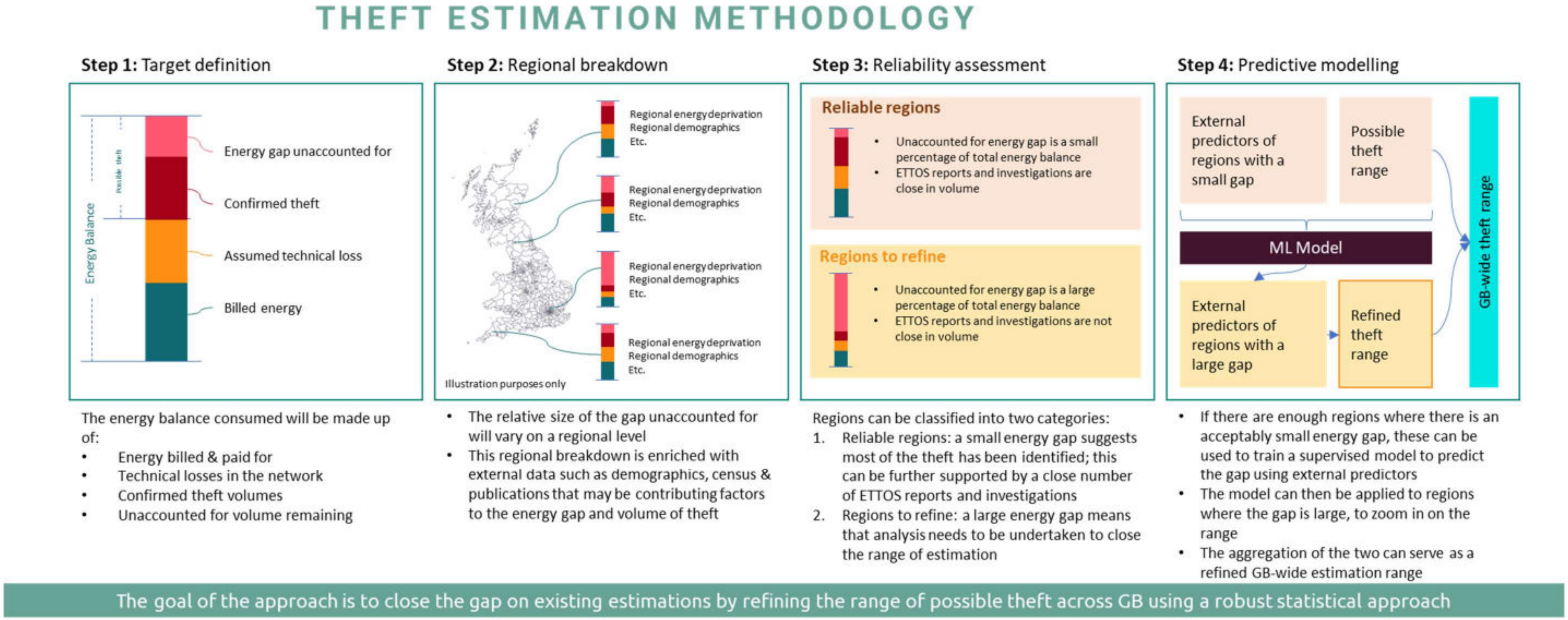
These tables are used to identify the instances of theft investigation. For each instance:

- The supply postcode is used to identify spatially based predictors such as crime rates in the local area of this theft investigation.
- The combination of (`mpxn`, `supply_postcode`, `meter_serial_number`, `applicable_date`) is used to search for the meter associated with this theft investigation instance, in order determine the meter type and meter location involved in this theft investigation.
- The combination of (`mpxn`, `supply_postcode`, `account_number`, `applicable_date`) is used to search for the account associated with this theft investigation instance, in order to determine the normal payment method involved in this theft investigation.
- The list of meters and accounts that is searched in is constructed from the TRAS consumption dataset, as described in the next section. From this dataset, a start date and an end date for each meter or account is constructed, which is also described in the next section.
- In general, each combination of (`mpxn`, `supply_postcode`) can have multiple meters, so a selection is made of the meter whose start date and end date enclose `applicable_date`.
- In general, each combination of (`mpxn`, `supply_postcode`) can have multiple accounts, so a selection is made of the account whose start date and end date enclose `applicable_date`.
- The `applicable_date` is defined as follows,
    - If the theft investigation confirms that there is a theft and there is an assessed theft start date, then it is taken as `applicable_date`.
    - If the theft investigation is closed without confirming that there is a theft, then the date of closing the investigation is taken as the `applicable_date`.
    - If the theft investigation is not closed, the latest tamper report date is taken as `applicable_date`.

Otherwise, the theft investigation file date is taken as `applicable_date`.

9. APPENDIX B – ORIGINAL METHODOLOGY

The figure below defines the initial proposed methodology, which would have been undertaken had the required data been available.



THEFT ESTIMATION METHODOLOGY

Step 1: Target definition

The energy balance consumed will be made up of:

- Energy billed & paid for
- Technical losses in the network
- Confirmed theft volumes
- Unaccounted for volume remaining

Step 2: Regional breakdown

Regional energy deprivation
Regional demographics
Etc.
Regional energy deprivation
Regional demographics
Etc.
Regional energy deprivation
Regional demographics
Etc.
Regional energy deprivation
Regional demographics
Etc.
Illustration purposes only

- The relative size of the gap unaccounted for will vary on a regional level
- This regional breakdown is enriched with external data such as demographics, census & publications that may be contributing factors to the energy gap and volume of theft

Step 3: Reliability assessment

Reliable regions

- Unaccounted for energy gap is a small percentage of total energy balance
- ETTOS reports and investigations are close in volume

Regions to refine

- Unaccounted for energy gap is a large percentage of total energy balance
- ETTOS reports and investigations are not close in volume

Regions can be classified into two categories:

1. Reliable regions: a small energy gap suggests most of the theft has been identified; this can be further supported by a close number of ETTOS reports and investigations
2. Regions to refine: a large energy gap means that analysis needs to be undertaken to close the range of estimation

Step 4: Predictive modelling

External predictors of regions with a small gap
Possible theft range
ML Model
External predictors of regions with a large gap
Refined theft range
GB-wide theft range

- If there are enough regions where there is an acceptably small energy gap, these can be used to train a supervised model to predict the gap using external predictors
- The model can then be applied to regions where the gap is large, to zoom in on the range
- The aggregation of the two can serve as a refined GB-wide estimation range

The goal of the approach is to close the gap on existing estimations by refining the range of possible theft across GB using a robust statistical approach

10. APPENDIX C – HYPOTHESIS

REF	STATEMENT	RATIONALE	Test	ENERGY	PRIORITY
H1	Some suppliers have been historically better at investigations from different lead types	Suppliers with a greater field force in a region have been able to investigate more risk of theft and therefore proactively investigated	Conversion rates by source code across regions / LDZ/DNO	Gas Electricity	Must do
H2	Those committing energy theft do not change supplier	Changing supplier may trigger a meter reading which could identify theft occurring	Duration with supplier of confirmed thefts vs non confirmed thefts vs not at risk	Gas Electricity	Should do
H3	Theft can occur equally in the properties with or without smart meters	Smart meters are no different, just needs one power down and back up	Rates of theft per population TRAS & ETTOS	Electricity Gas	Would do
H4	Thefts may incorrectly be reported as faulty meters	Faulty meters doesn't account consumption and the meter reads is not recorded. AT times, these meters are deliberately broken, consumed power and gas is theft	Rates of reason code for no theft where investigated (do we have rates of faulty meters where no theft is suspected)	Gas Electricity	Would do
H5a	Parties who have stolen in one instance are likely to do so in others	A landlord may steal at all properties they own. People who have moved homes and have tampered in one instance are likely to tamper at the new place Same customer may tamper at the same address (Future – ability to track customers across address / suppliers)	TRAS & ETTOS Evidence of same landlord recurring in suspected / confirmed theft	Gas Electricity	Could do
H5b	Theft may recur at the same property	The same customer may continue to tamper with the meter or there may be new occupants who then tamper with the meter		Gas Electricity	Could do
H6	Thefts are more likely to be detected by suppliers in prepaid meters than credit meters	Targeting pre paid meters was an easy win to identify theft	Conversation rates of thefts / investigations to credit/prepay meters, source (proactive/responsive) Rate per population of confirmed theft	Gas Electricity	Must do
H7	Suppliers don't monitor credit meter enough where	PP may have relation to theft based on the current TRAS data however suppliers don't monitor credit meter enough to find theft	PP/CM ratios across different suppliers and source for theft / investigations	Gas Electricity	Should do

	significant amount of theft occurs				
H8	Potential thefts in remote locations are less likely to have been investigated	It may have been too time consuming to send resources to remote locations when there was a possibility a legal reason was behind identification e.g. holiday let (more likely reactive than proactive)	Conversation rates in rural / urban locations, spatially, by source code / supplier code Rate per population of confirmed thefts	Gas Electricity	Must do
H9	Incidence of theft is likely to have increased due to the financial impact of Covid and lockdowns	COVID period has tested the financial aspects of many people and may see increased theft	Time series analysis of confirmed thefts Energy gap time series analysis	Gas Electricity	Should do
H10	Cost of living impacts will increase theft	With stretched budgets, there is an increased propensity to energy theft (more theft and potentially less investigations)	Conversion rates over time / gas vs electricity Energy gap time series analysis Investigation rates over time, by lead type	Gas Electricity	Should do
H11	Potential thefts in difficult to resource locations are less likely to have been investigated	It may have been too costly to send resources into areas such as central London due to higher charges (congestion zone etc)	Conversation rates in alignment with removal of services, by source code / supplier code Rate per population of confirmed thefts	Gas Electricity	Must do
H12	Crimestoppers campaign areas will have increased theft rates	Expected to be an increase in 'tip off' source of data in locations where marketing is occurring	Conversion rates by source data over duration of campaign	Gas Electricity	Should do
H13	There has been a bias to investigations within deprived areas	It was easier to identify energy theft through pre pay meters / crime reports rather than more complex scenarios	Conversion rates in deprived vs affluent areas / crime rates	Gas Electricity	Must do
H14	High deprivation are likely to have higher levels of theft	Financial obligations may drive theft incidents for areas with high level of deprivation	Confirmed theft per population rate? Conversion rates of investigations	Gas Electricity	Must do
H15	Fuel poverty levels may show high association with the levels of theft	Financial obligations may drive theft incidents where levels of fuel poverty is high	Rates /volumes of confirmed theft by poverty levels Rates of investigations by poverty levels	Gas Electricity	Must do
H16	Newly subdivided property area have more unregistered meters (error /theft)	Not all installations of additional meters have been registered, knowingly or unknowingly	Comparison of areas where population have increased / properties increased, tampering codes	Gas Electricity	Would do

			Rates /volumes of confirmed theft		
H17	Social housing areas are likely to have higher levels of theft	Fully maintained properties by local authority, housing association may see theft incidents Is this due to additional checks that it is being discovered, different types of tenants	Rates /volumes of confirmed theft by poverty levels Rates of investigations by poverty levels	Gas Electricity	Could do
H18	Non energy crime is an indicator of energy theft	Those with a propensity to commit energy theft are already involved in other illegal activities, e.g. motoring offenses, petty theft, building regulation disputes	Rates /volumes of confirmed theft between crime stats and energy theft – calling out specific crime types	Gas Electricity	Would do
H19	Energy theft levels strongly relate to the levels of other types of crime	Energy theft levels are likely to correlate geographically with levels of other types of crime	Rates /volumes of confirmed theft against crime levels in area	Gas Electricity	Must do
H20	There may be correlation with water theft for some theft types (e.g. cannabis farms)	Energy as well as water requirement is high for cannabis farms. Also water theft may be closely related to energy theft	Correlation of water theft against energy theft locations	Electricity	Would do
H21	Theft is more likely to be detected in cannabis farms	Cannabis cultivation needs 18/12/6 hours of energy for 6/6/4 weeks and theft is therefore more likely	TRAS, published information for cannabis	Electricity	Would do
H22	Theft is more likely in hospitality relating businesses, nursing homes, laundrettes	There have been experiences that these types of companies have historically committed energy theft.	Explore conversion rates of risk theft and rates of theft within groups to test for bias (investigations vs theft found)	Gas Electricity	Should do
H23	Theft is more likely in poor quality businesses such as those with low hygiene levels	Revenue protection have historically used hygiene ratings to identify business to investigate for fraud	Explore conversion rates of risk theft and rates of theft within those with low scores vs other scores to test for bias (investigations vs theft found)	Gas Electricity	Should do
H24	Small/ medium business are more likely be involved in energy theft than large corporations	Large national / multinational corporations and government bodies are less likely to participate in energy theft whereas smaller companies may be more inclined	Test theft rates of investigations / confirmed thefts with company classifications (potentially link to CIS with companies house data)	Electricity Gas	Should do

11. APPENDIX D – CLASSIFICATION MODEL METRICS

11.1. CONFUSION MATRICES

Figure 32: Confusion Matrix for Scenarios 1 and 2 Classification Model (Gas, Residential, England/Wales)

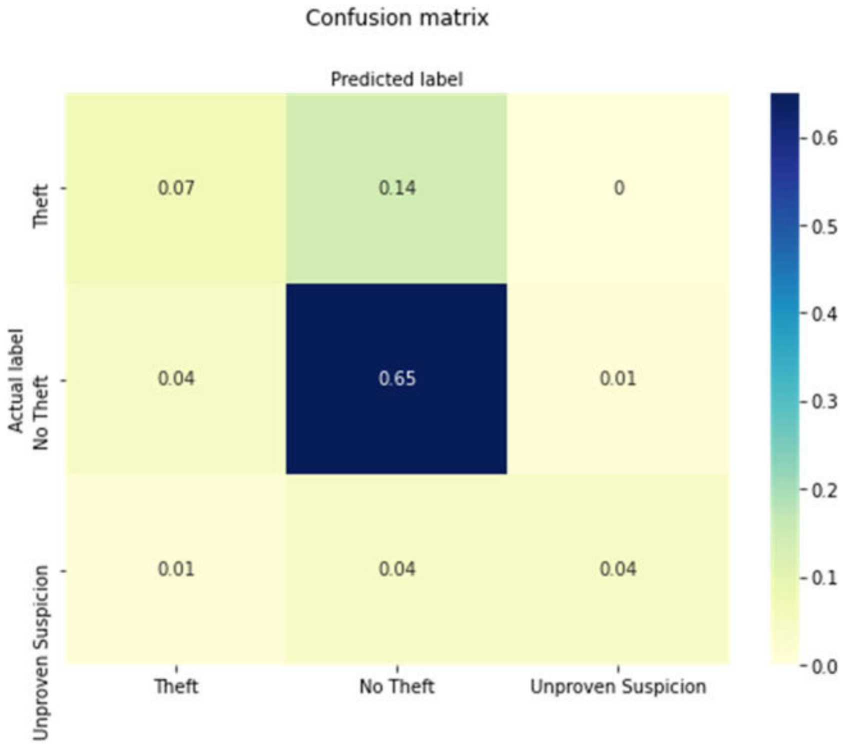


Figure 33: Confusion Matrix for Scenario 3 Classification Model (Gas, Residential, Scotland)

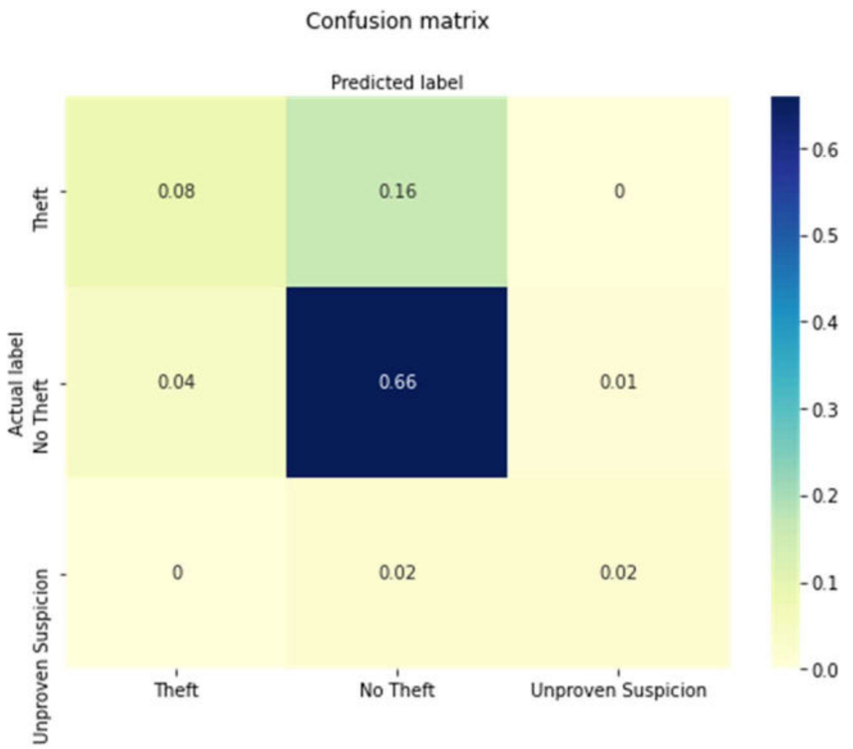


Figure 34: Confusion Matrix for Scenarios 4 and 5 Classification Model (Gas, Commercial, England/Wales)

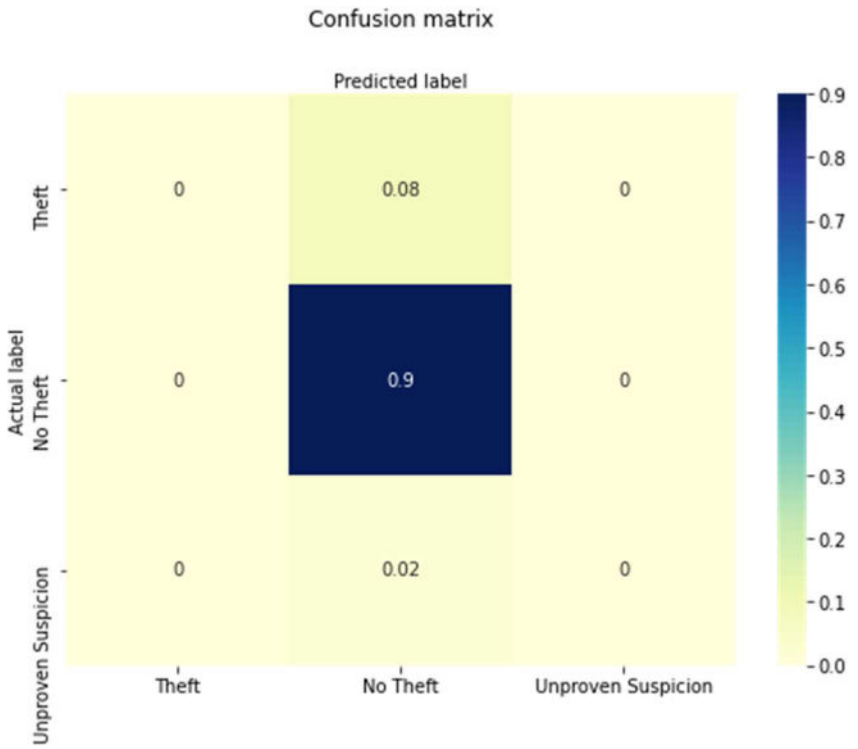


Figure 35: Confusion Matrix for Scenario 6 Classification Model (Gas, Commercial, Scotland)

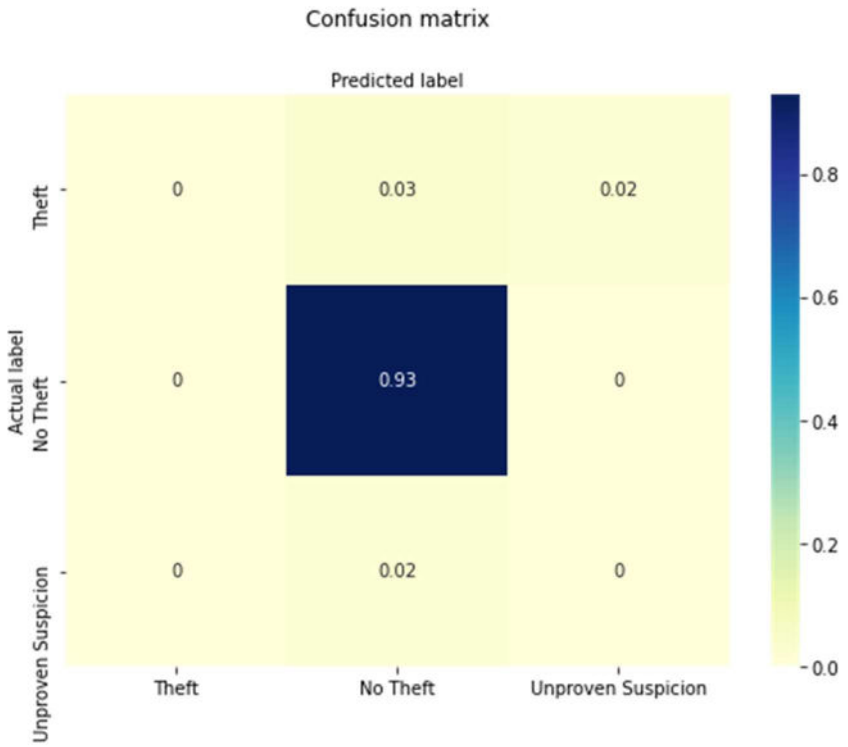


Figure 36: Confusion Matrix for Scenarios 7 and 8 Classification Model (Electricity, Residential, England/Wales)

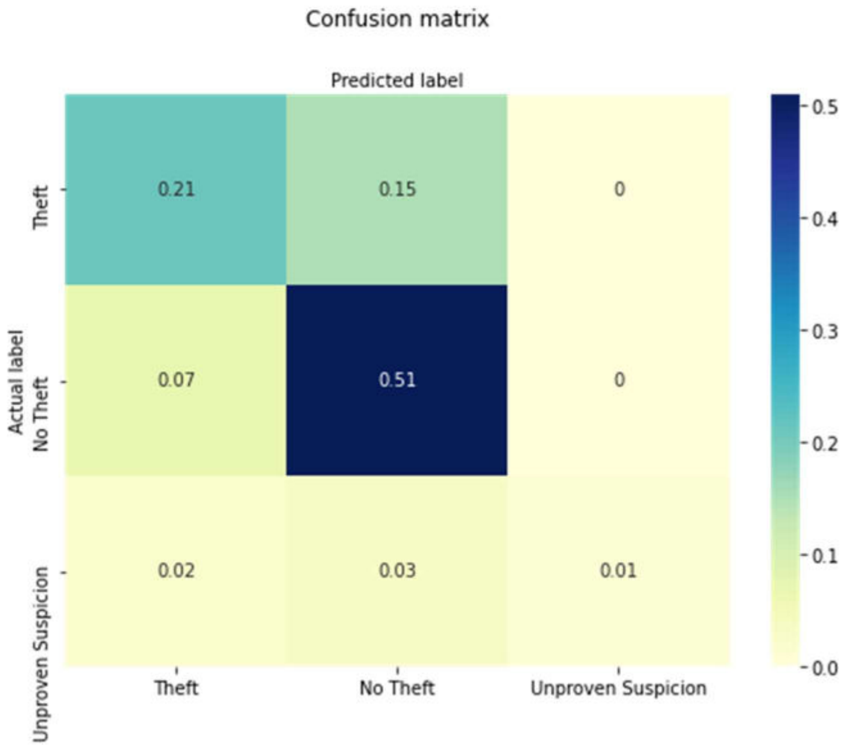


Figure 37: Confusion Matrix for Scenario 9 Classification Model (Electricity, Residential, Scotland)

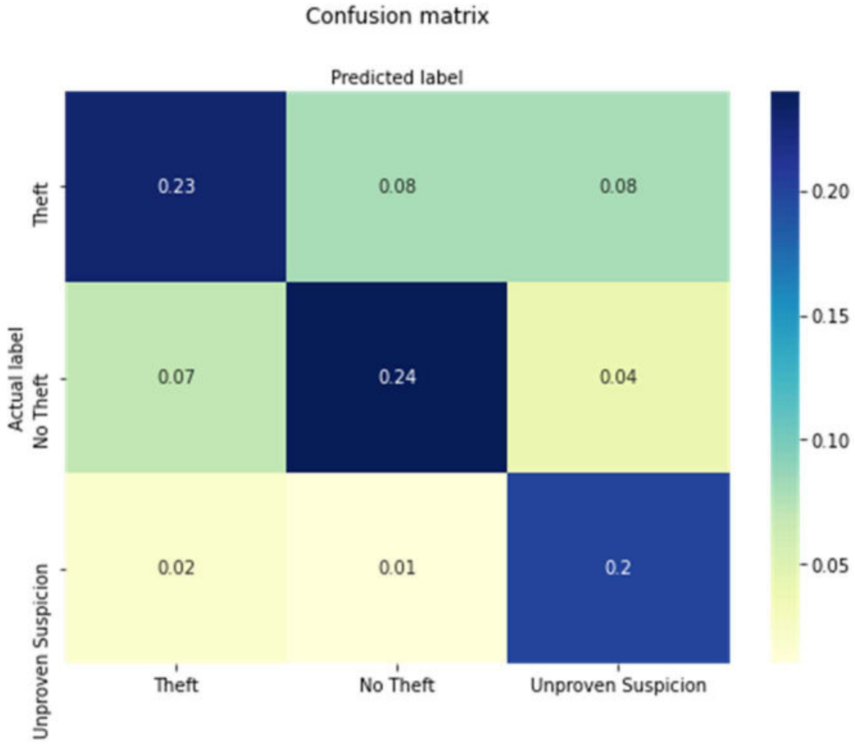


Figure 38: Confusion Matrix for Scenarios 10 and 11 Classification Model (Electricity, Commercial, England/Wales)

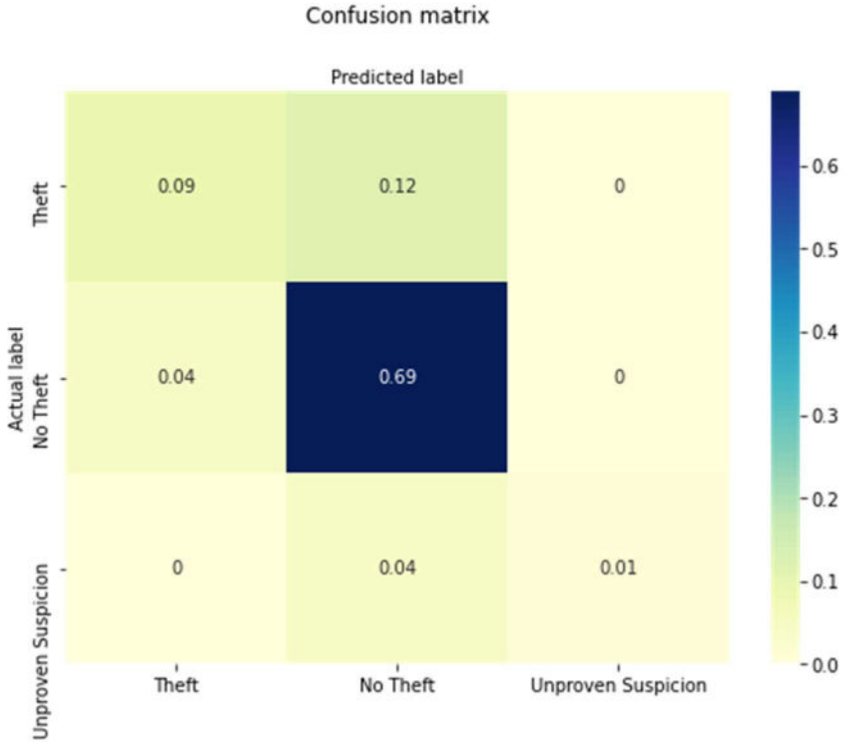
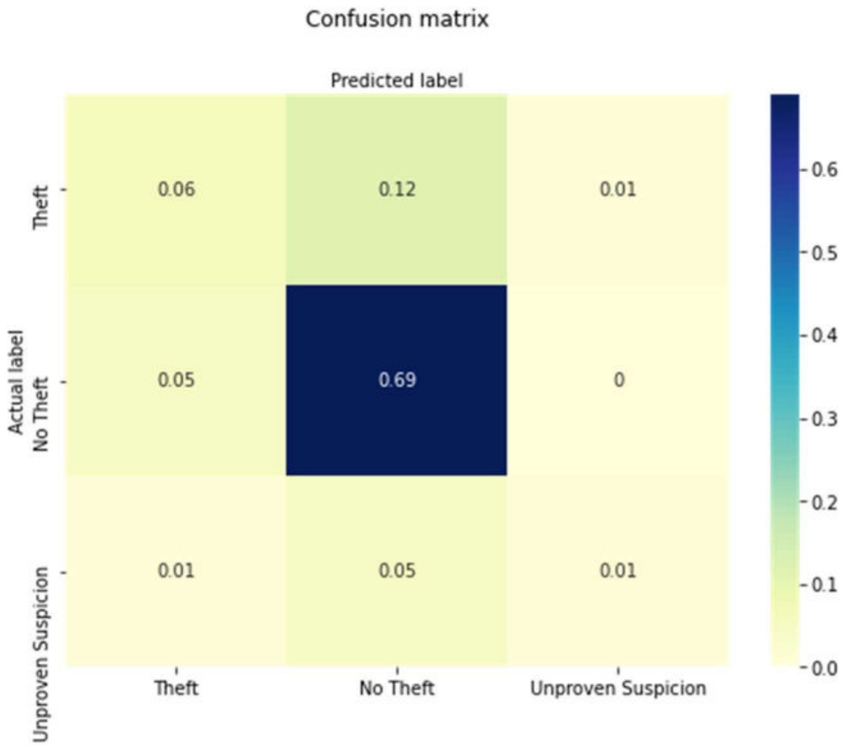


Figure 39: Confusion Matrix for Scenario 12 Classification Model (Electricity, Commercial, Scotland)



11.2. RELATIVE FEATURE IMPORTANCE

Figure 40: Relative Feature Importance Plot for Scenarios 1 and 2 Classification Model (Gas, Residential, England/Wales)

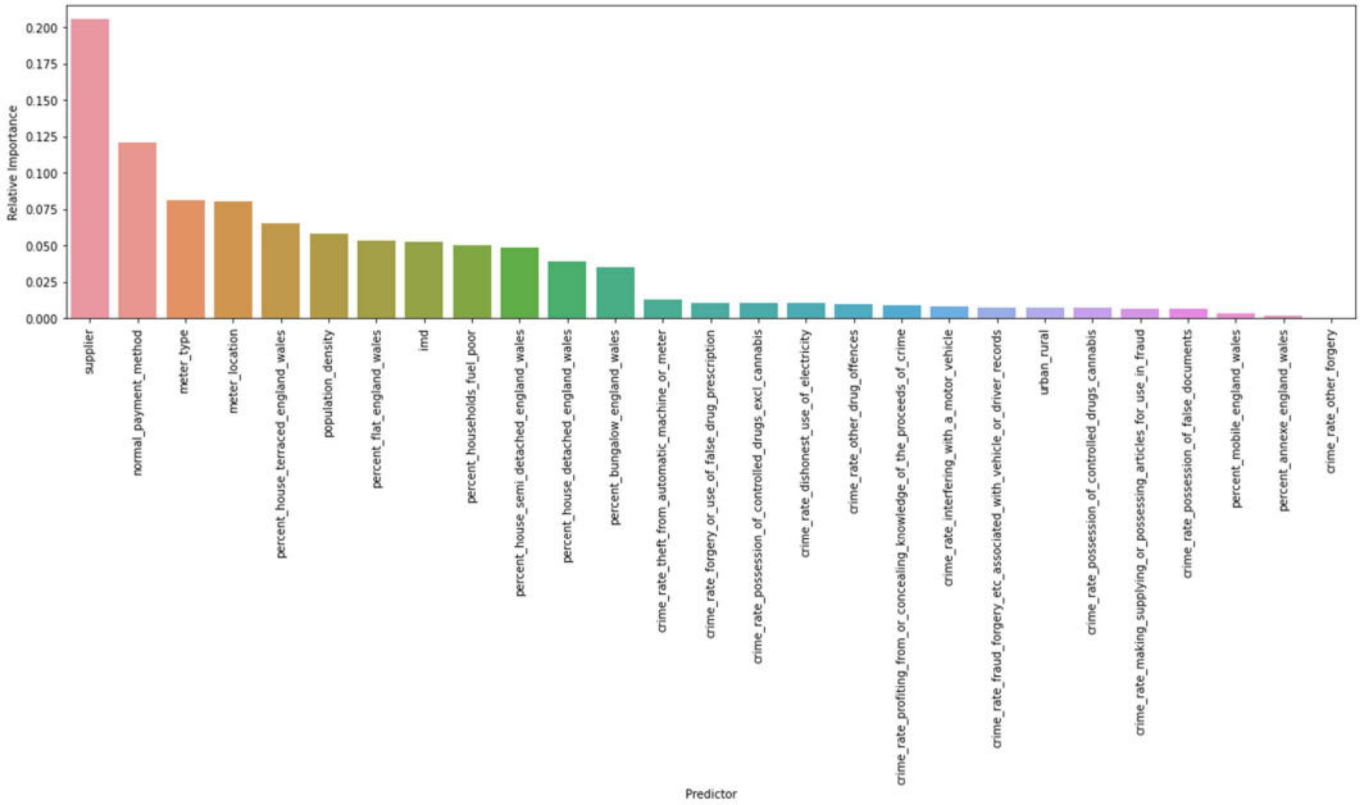


Figure 41: Relative Feature Importance Plot for Scenario 3 Classification Model (Gas, Residential, Scotland)

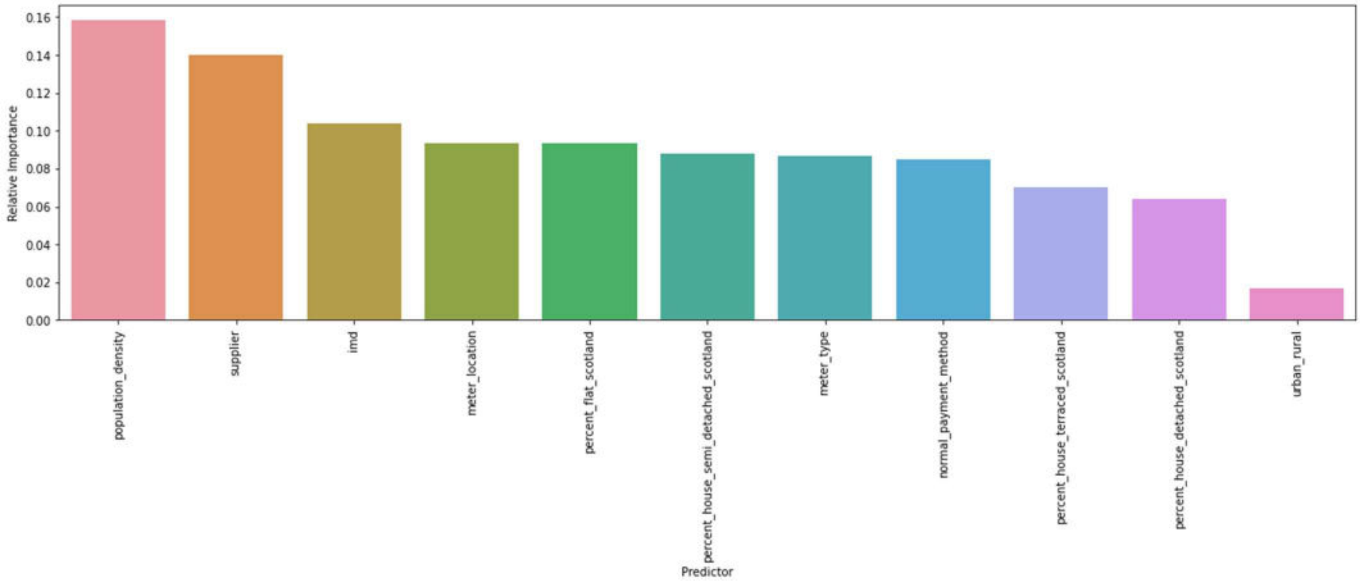


Figure 42: Relative Feature Importance Plot for Scenarios 4 and 5 Classification Model (Gas, Commercial, England/Wales)

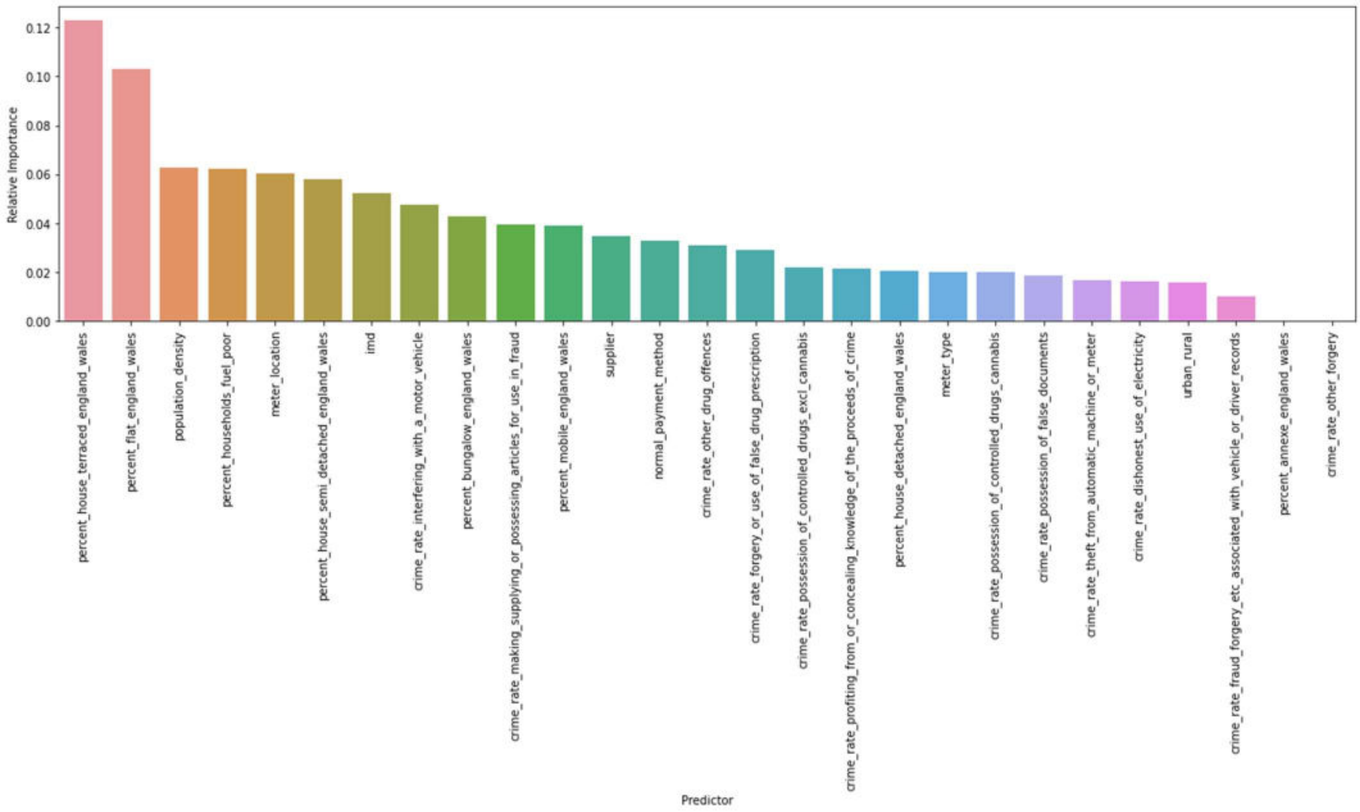


Figure 43: Relative Feature Importance Plot for Scenario 6 Classification Model (Gas, Commercial, Scotland)

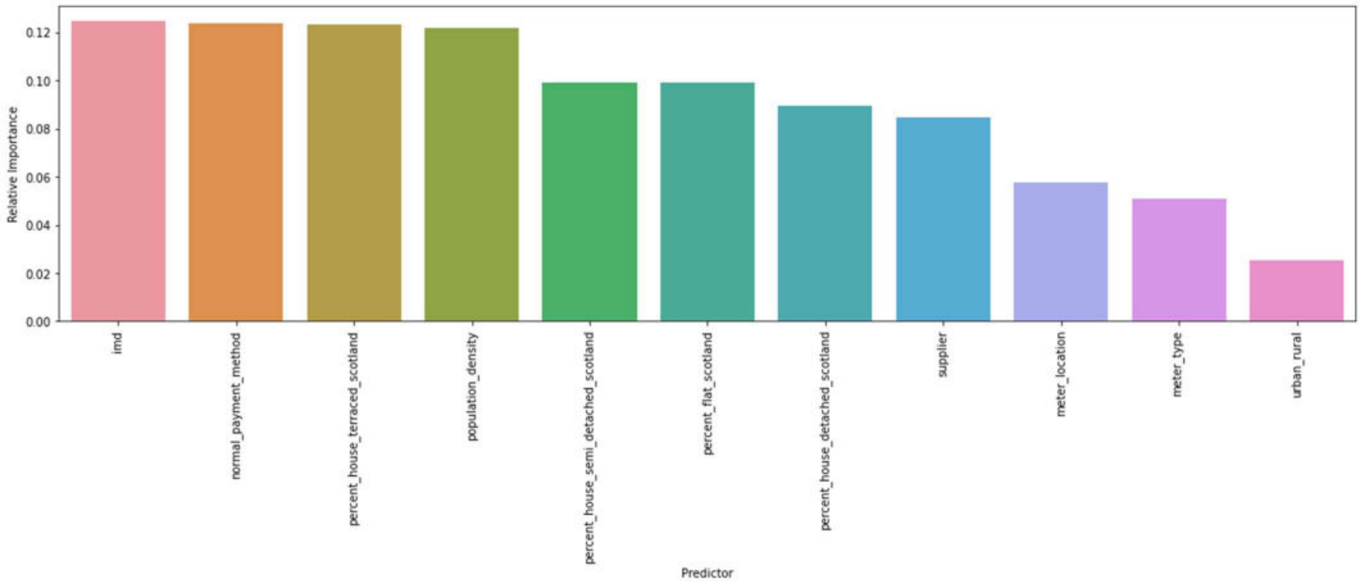


Figure 44: Relative Feature Importance Plot for Scenarios 7 and 8 Classification Model (Electricity, Residential, England/Wales)

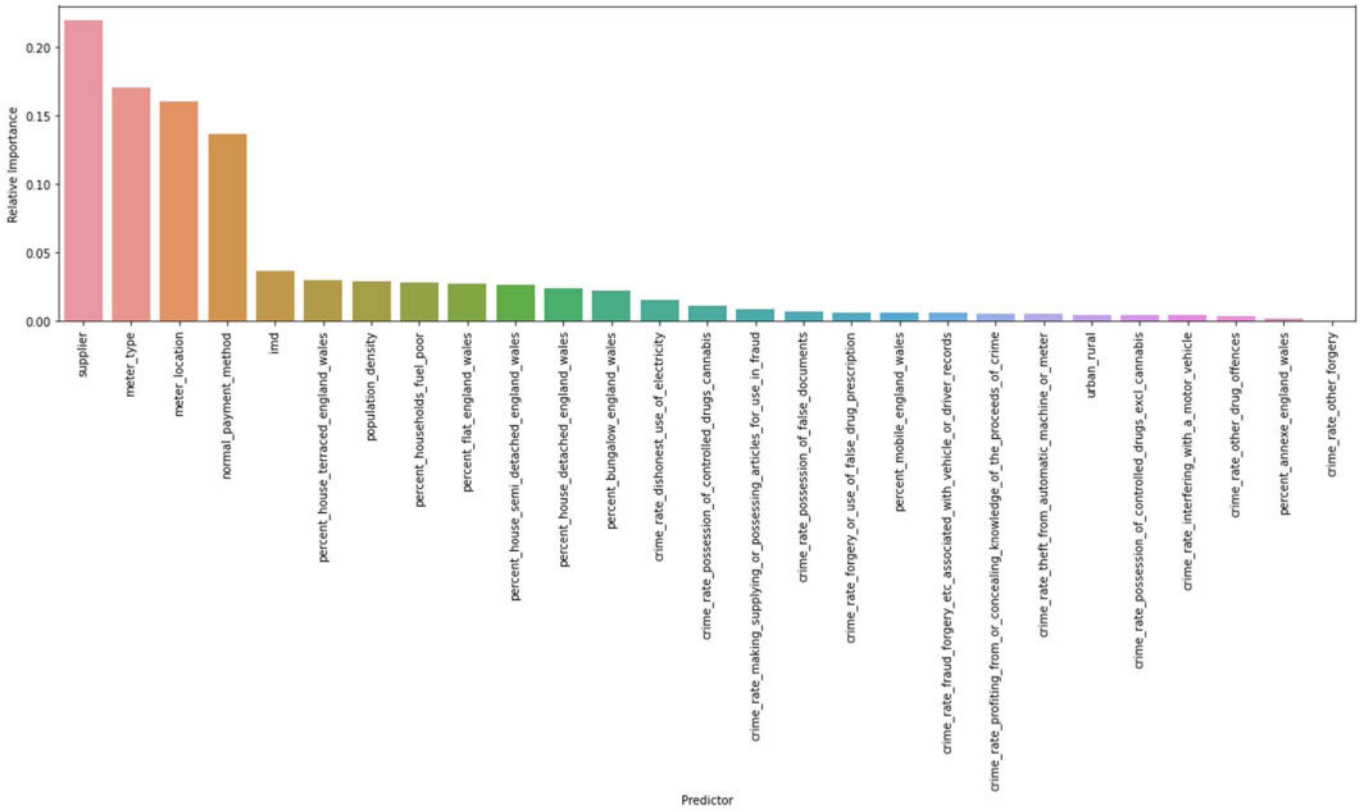


Figure 45: Relative Feature Importance Plot for Scenario 9 Classification Model (Electricity, Residential, Scotland)

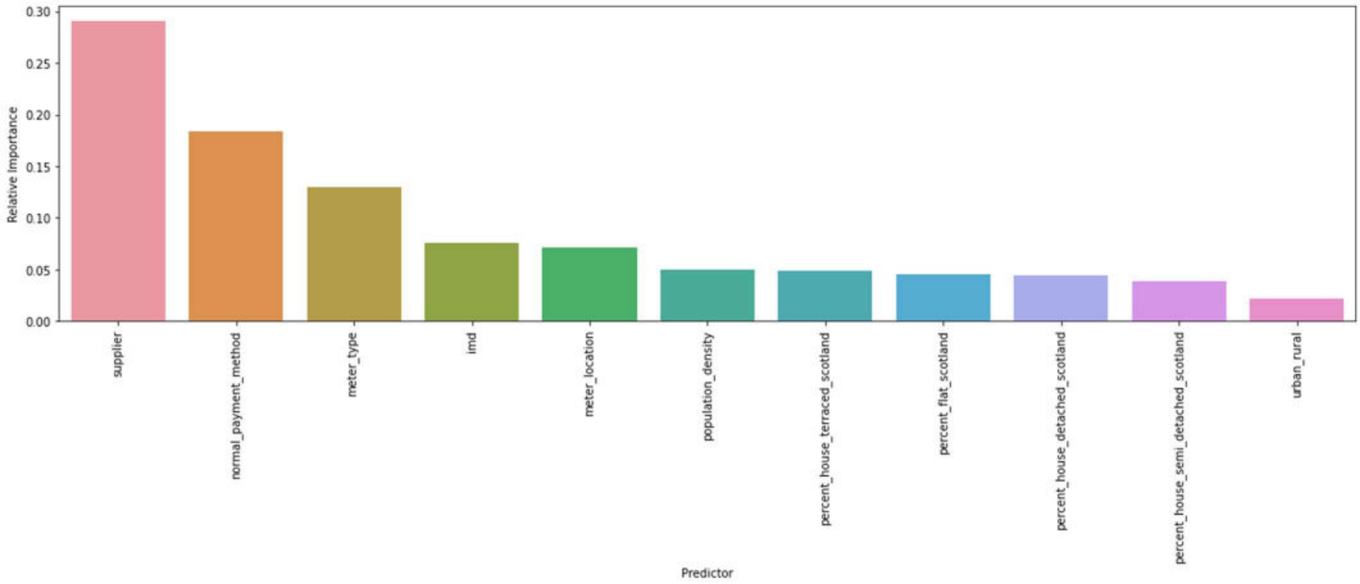


Figure 46: Relative Feature Importance Plot for Scenarios 10 and 11 Classification Model (Gas, Commercial, England/Wales)

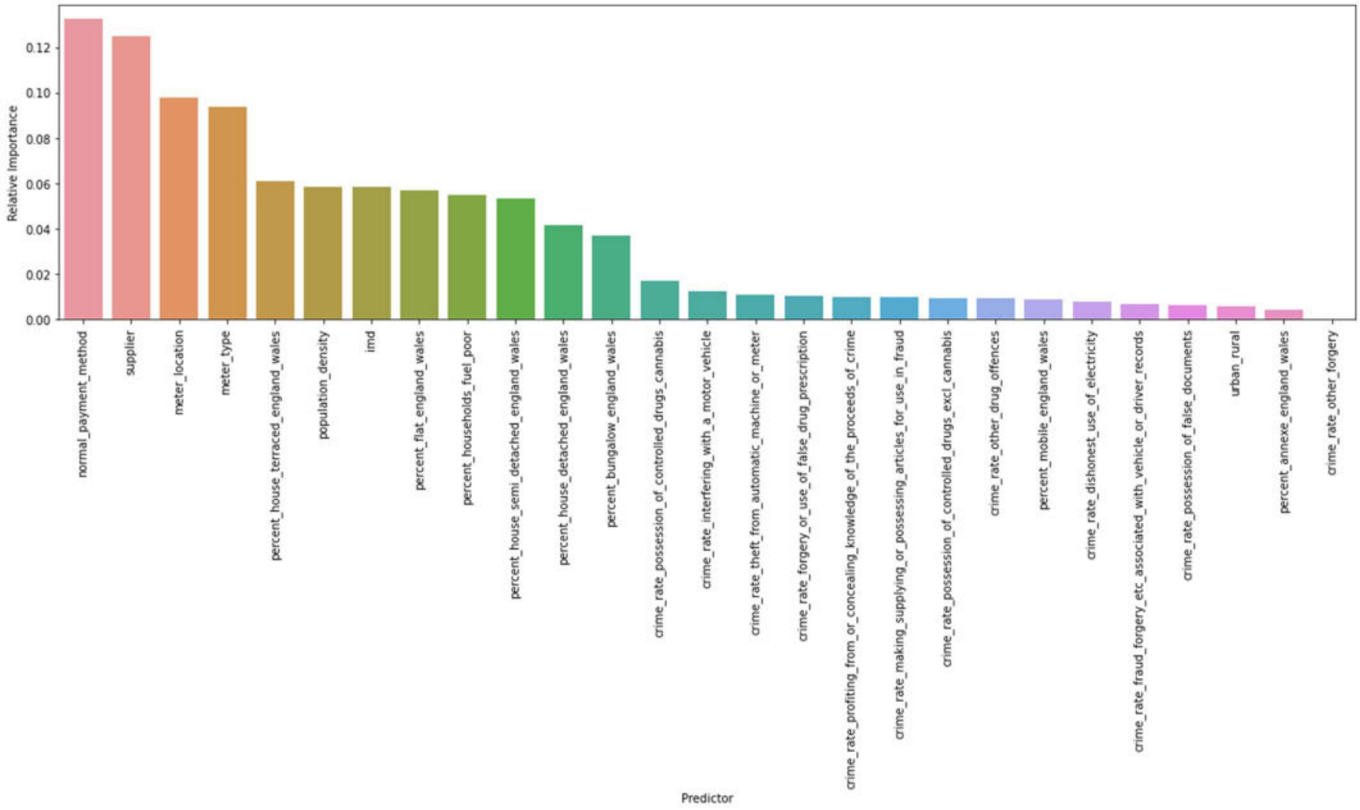
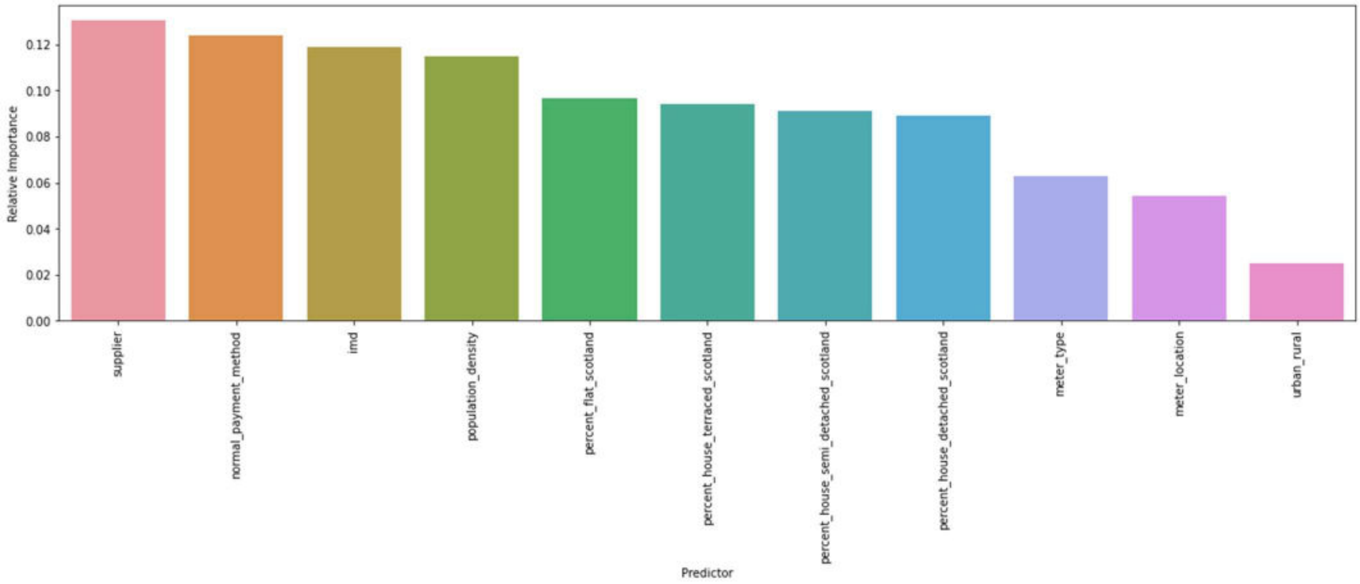


Figure 47: Relative Feature Importance Plot for Scenario 12 Classification Model (Gas, Commerical, Scotland)



12. APPENDIX E – REGRESSION MODEL METRICS

Figure 48: Relative Feature Importance Plot for Scenarios 1, 2 and 3 Regression Model (Gas, Residential)

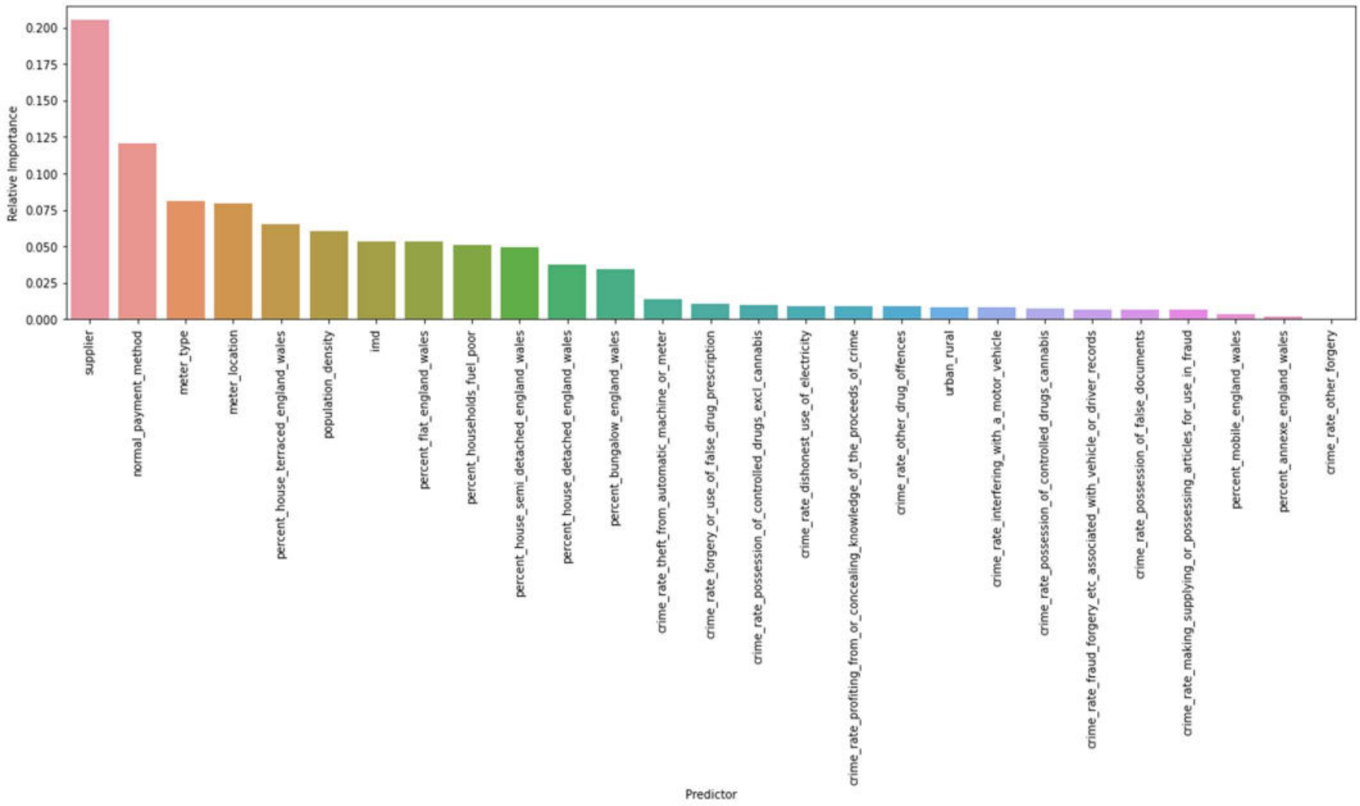


Figure 49: Relative Feature Importance Plot for Scenarios 4, 5 and 6 Regression Model (Gas, Commercial)

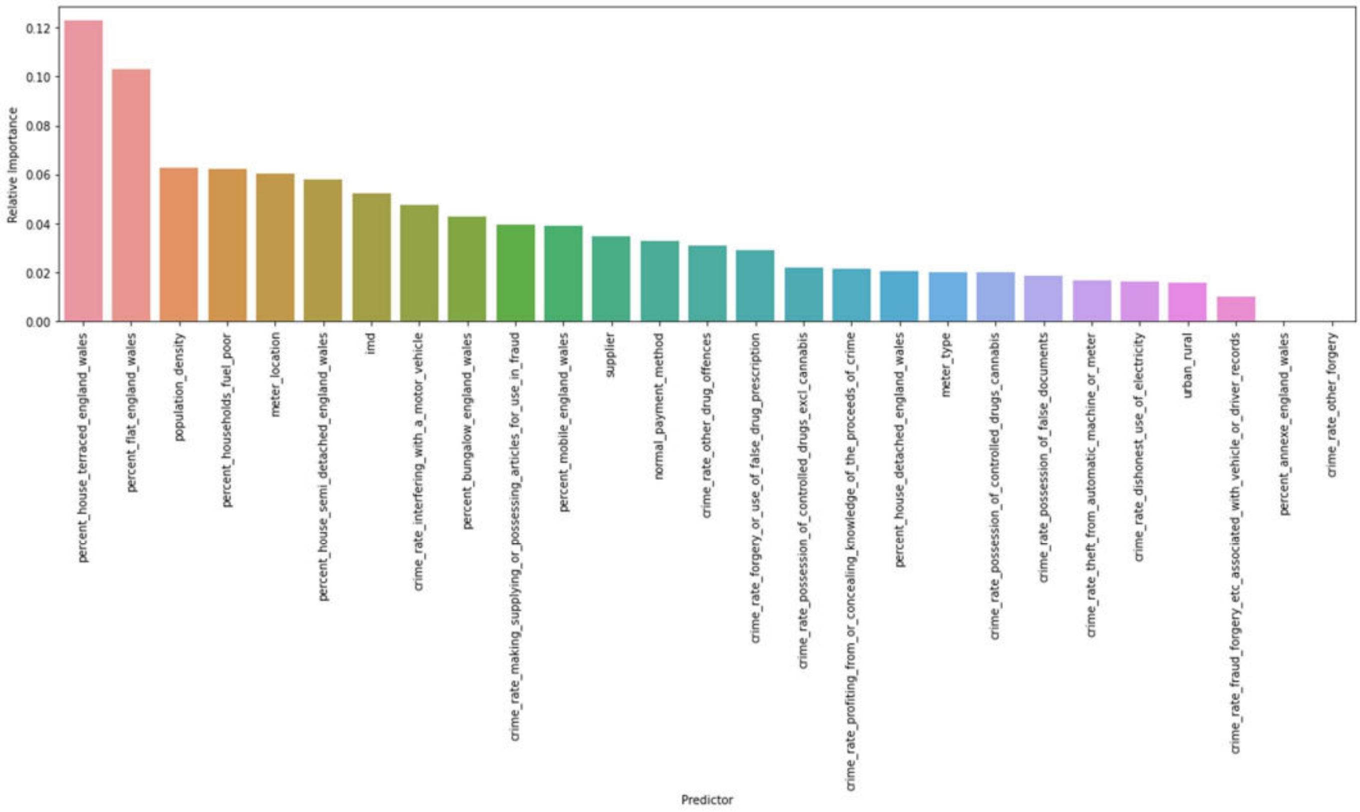


Figure 50: Relative Feature Importance Plot for Scenarios 7, 8 and 9 Regression Model (Electrcity, Residential)

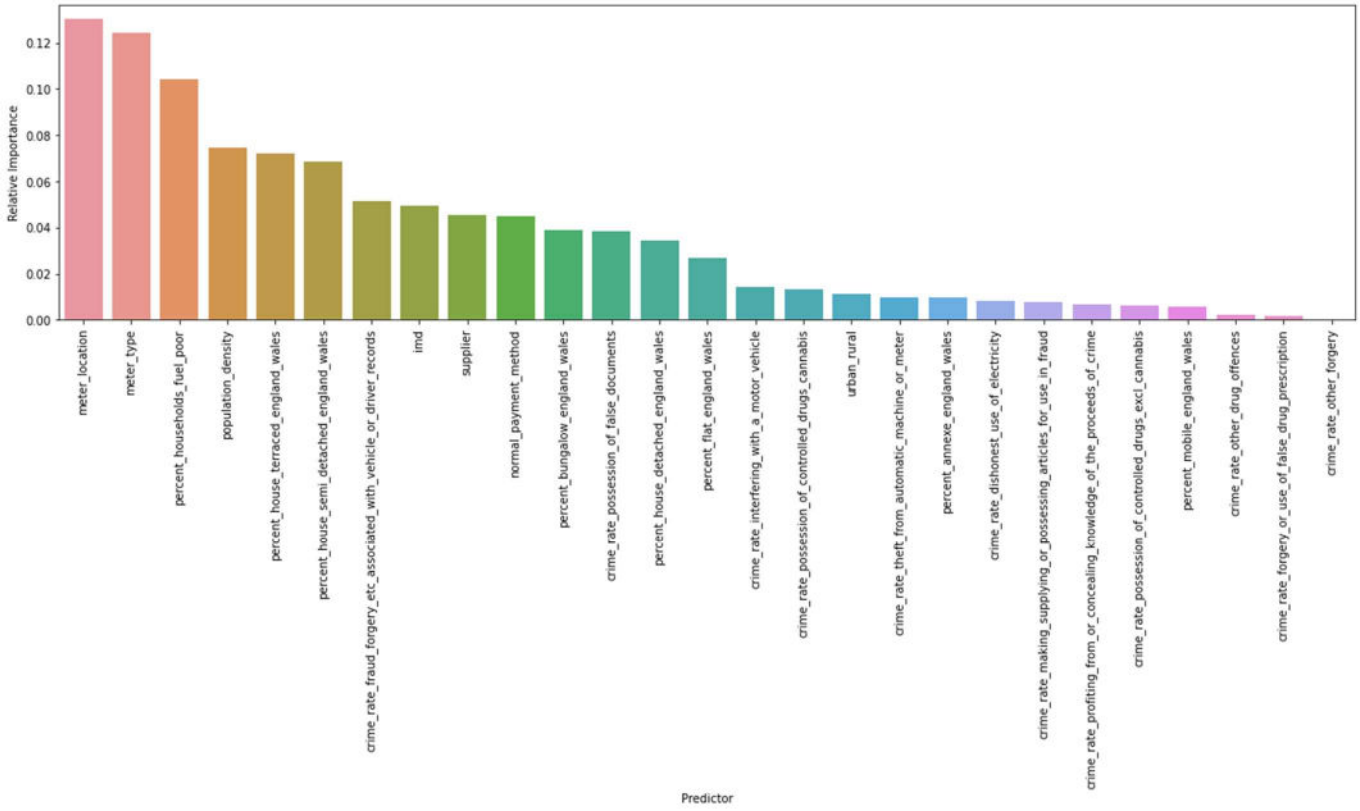
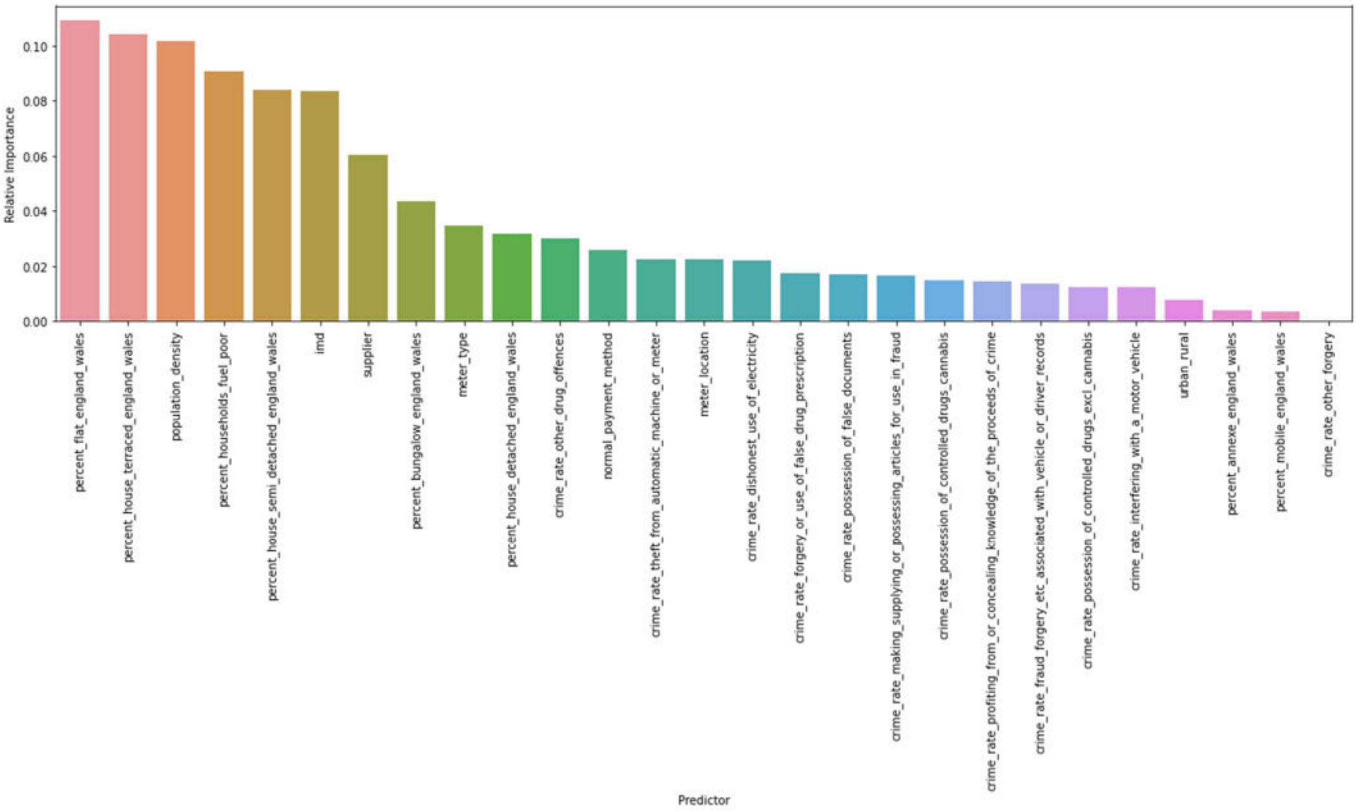


Figure 51: Relative Feature Importance Plot for Scenarios 10, 11 and 12 Regression Model (Electricity, Commercial)



13. APPENDIX F – GLOSSARY OF TERMS

ACRONYM	DEFINITION
AAD	Azure Active Directories
ADF	Azure Data Factory
DCC	Data Communications Company
DNO	Distribution Network Operator
ETTOS	Energy Theft Tip Off Service
EV	Electric Vehicle
GB	Great Britain
GDPR	General Data Protection Regulation
GSP	Grid Supply Point
HH	Half Hourly
IMD	Index of Multiple Deprivation
LDZ	Local Distribution Zone
LSOA	Lower Layer Super Output Area
ML	Machine Learning
MPAN/MPRN	Meter Point Administration Number/ Meter Point Reference Number
MSE	Mean Squared Error
ONS	Office of National Statistics
PAYG	Pay as you go
RECCO	Retail Energy Code Company
RMSE	Root Mean Squared Error
RPA	REC Performance Assurance
SME	Subject Matter Exert
TDIS	Theft Detection Incentive Scheme
TEM	Theft Estimation Methodology

TRAS	Theft Risk Assessment Service

14. APPENDIX G – SOURCES

The table below provides links to the spatial and predictor datasets used for this project:

Data Description	Format	Link to Source/ Geography
LSOA Data	Shapefile	GB
GSP Data	Shapefile	GB
Spatial Datasets	Csv	Postcodes to LDZs Postcode to OA to LSOA
Housing Data	Csv	England and Wales Scotland
IMD	Csv	England Wales Scotland
Population	Csv	England and Wales Scotland
Fuel Poverty	Csv	England
Urban / Rural	Csv	England and Wales Scotland

In addition, the following datasets were used as part of the methodology:

- National Grid Gas Data Item Explorer:

  https://mip-prd-web.azurewebsites.net/DataItemExplorer

- Xoserve Chart showing Percentage of Unallocated Gas:

  https://www.xoserve.com/uig-charts/uig-as-of-total-throughput/

- Gas and electricity prices in the non-domestic sector - GOV.UK (www.gov.uk) Table 314

  https://www.ofgem.gov.uk/publications/default-tariff-cap-level-1-october-2022-31-december-2022

Sources used for context

UK Energy in Brief 2021 (publishing.service.gov.uk)