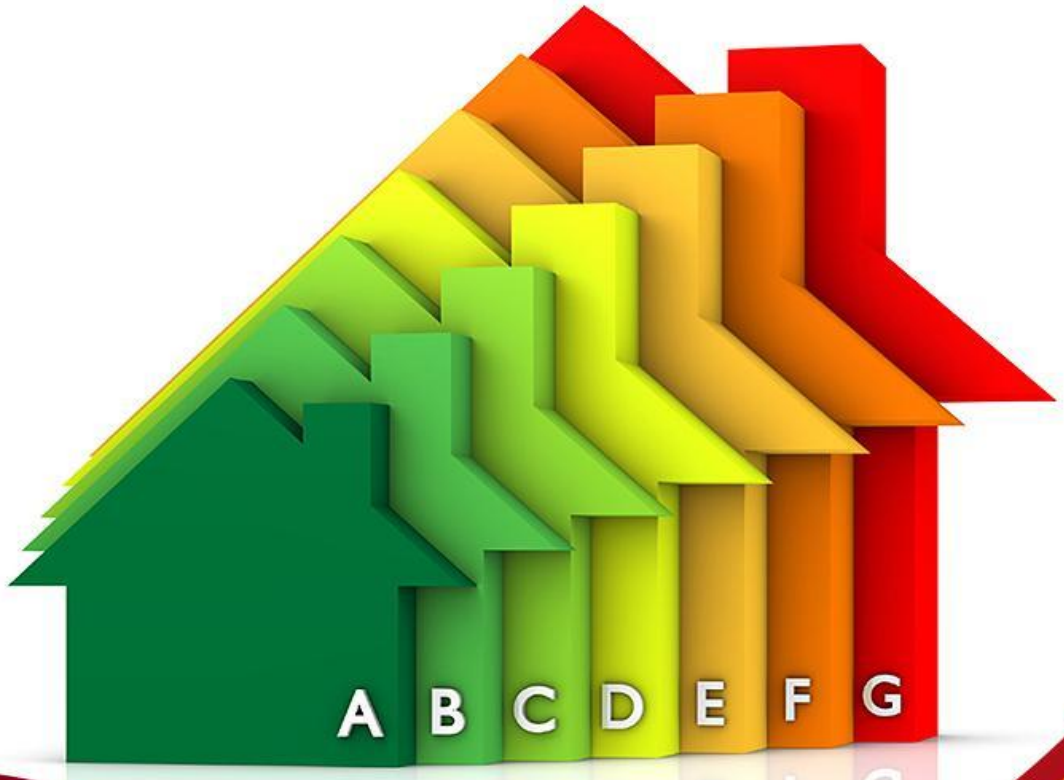
Can “fuel poverty” be estimated simply?



Introduction

As part of the Code First Girls Nanodegree Data Science specialisation, we were asked to complete a data analytics project. This involved defining a question, working as a team to utilise the skills and tools introduced to us by CFG and presenting our findings in a digestible format. This document is the artefact produced by the activities undertaken by our team.

The “cost of living crisis” is well documented in the UK media and it is anticipated that this will increase the incidence of fuel poverty in England (currently at 8% of all households). Estimating this is unfortunately not feasible due to the complex modelling that is required in estimating the relationship between fuel prices experienced at the household level and aggregates (such at the Ofgem cap or the energy component of the CPI). However, we can identify with some modelling techniques factors which contribute to the incidence of fuel poverty by using estimates of household fuel expenditure and factors that determine it in combination with income. The results of this data could be used to identify the likelihood of a household experiencing fuel poverty and therefore their entitlement to government benefits seeking to target this problem.

Aims and objectives of the project:

- Utilise python data visualisation for data exploration
- Utilise machine learning libraries and techniques to model fuel poverty

The report follows the structure below:

- Introduction
- Background
- Steps specification
- Data sourcing and pre-processing
- Exploratory analysis
- Implementation and Execution
- Model results
- Conclusion
- Appendices

Background

Our group is composed of three students who are sponsored by Experian, a large credit rating agency with expertise in consumer credit rating. In order to prepare for potential future roles within the company, the students decided to focus on a topic relevant for household finances and explore a topic with household level data.

The “cost of living crisis” is well publicised in the UK media. Of those households who reported a rising cost of living to the ONS Opinions and Lifestyle Survey (OPN) in January 2022, 79% cited higher gas and electricity bills as a cause following the October ‘21 hike of 12%. Since then there has been an additional 54% rise in the Ofgem default energy price cap in April ‘22 and there is another anticipated rise in October (by some estimates at around 34%.) Therefore, within one year, the energy price cap would have risen 128%. It is anticipated that increased energy prices will disproportionately impact those on lower incomes, especially those in low energy efficient housing.

It is near impossible to model the pass through effect of increased energy prices to household fuel expenditure given variation in prices paid by region/tariff/payment method. There is also the added complication of price elasticity of demand which is difficult to capture without longitudinal data or actual observed fuel expenditure. One thing is certain however, fuel poverty (defined as fuel expenditure of 10% or more of full household income) will increase as prices rise. Hence the focus of this project is to look at key factors determining fuel poverty.

The Department for Business, Energy and Industrial Strategy publishes an annual report called the “Annual Fuel Poverty Statistics in England”. They use the “English Housing Survey” to estimate fuel expenditure for a surveyed household using an estimate of consumption (determined by the energy efficiency rating, household composition and region amongst others) and price (determined by region, payment method and fuel type e.g. gas). Establishing whether a household is deemed as experiencing fuel poverty or not depends on the estimated fuel expenditure and ‘after housing cost’ equivalised income. The former is a complex calculation and they admit their model overestimates actual usage especially for households in low rated dwellings. Unfortunately, actual consumption data is owned by energy companies and is not accessible.

As these estimates are used for government policies to help those suffering from fuel poverty, we decided to use the modelling techniques taught to see if we could estimate the incidence of fuel poverty in a simple way. If successful, the model could be used at the grass roots level for instance in a JobCentre Plus where data for government benefits are collected and entitlement decided.

Steps specification

Framing the question & data gathering:

To understand the process of deciding the question as well as data gathering, please see Appendix 1. Given the focus on binary classification algorithms within the course, we decided to reorientate the project to modelling the fuel poverty binary variable. By this point the data sets were already prepared for training and testing a linear regression, so excluded many variables. However, the exclusion of some variables (e.g. surface area of the dwelling, presence of gas boiler) is realistic considering the data that would likely be collected in a JobCentre Plus offices (e.g. region, household composition etc).

Data sourcing and pre-processing:

Though we had arrived on a question that we would be able to answer with publicly available data, we needed to be able to secure the data sets that would form the backbone of our statistical analysis. Data was sourced from the UK Data Service - a resource that collected data on a variety of topics that would be of interest to researchers. In our case, we sourced datasets that focused on the prevalence of fuel poverty in England. It is important to note here that we are using the publicly available data sets (as opposed to the full dataset that is protected by a licence), so some data has been expunged. This mostly takes the form of rows being removed (most likely for privacy reasons). It is important to note here that there are no null values present in the columns of this dataset. The main reason for this has briefly been touched upon above - the UK Data Service provides validated data for research purposes. It is quite likely that this dataset has been used by a governmental think tank prior to us choosing to use it and as such the dataset is reliable and will allow us to draw valid conclusions.

The initial dataset comprised 16 years worth of data (dated 2003 to 2019). During the data exploration period we realised that additional variables were added, and that these variables in particular would add significant value to our analysis. Our priority became restructuring the dataset and ensuring that these variables were accessible to us. One of the new variables added in the post 2014 datasets was gorEHS. GorEHS is the variable that represents different regions of England and Wales (which had a value of 3, but was not included in this dataset as it was outside the scope of our project). Prior to data construction and cleaning, the combined values of the initial 2003 to 2019 dataset was nearly 300,000 rows. This proved too much data for our personal computers to handle, so slicing the dataset into a more manageable format became integral to our data exploration activities.

The introduction of the 2014 region data, and its importance are explained as follows. It would be fair to assume that colder regions are more likely to require more fuel during cold snaps (eg. winter), in addition to this the regional variable would enable us to make a more educated guess of price estimates per unit of fuel and power used.

Bearing these factors in mind, the decision was made to focus on the data available from 2014 onwards. Again, we intended to use the entirety of the dataset (71, 000 rows across nearly 30 columns) but it proved quite cumbersome. Considering the refining that the question had gone through, we decided that it would be reasonable if we dropped extraneous columns to produce a dataset that held just 13 columns. In order to do this a fairly unwieldy piece of code was written, which would ensure that the same 13 columns would be pulled from the necessary year datasets (2014-2019).

From here, these trimmed year based datasets were merged into one 'master' or main dataset. This dataset would be used for operations that would be used to make generalised statements. As we were interested in using data learning models to see if they were capable of predicting fuel poverty, it was necessary for us to split this dataset into two datasets: testing and training. Treating a pooled master dataset as one and split arbitrarily would not be an effective way of creating a test sample. Therefore we needed to ensure that we took a random sample of 20% from each year to create a test and left the remainder for training. Fortunately the share of those in fuel poverty was the same in both the test, train and master samples.

Initially, the datasets were available in SPSS and TAB delimited formats. As neither of these were formats we were comfortable using, we opted to convert the TAB delimited files into comma separated files. This was done with a simple read/write file facilitated by the pandas python library.

In order to better understand the variables and their marker values (eg. gorEHS, hhcompx) please see Appendix 1 below.

Exploratory analysis:

Using the historical dataset from 2014 to 2019, various graphs were plotted to understand the relationship between the variables that factor into fuel poverty. Below are a few graphs that have been analysed in great depth, there are many more graphs that have been plotted within the project jupyter notebook.

Fig1 in Appendix 2 shows the percentage of households that are either in fuel poverty (represented by 1) or above fuel poverty (represented by 0) from 2014 to 2019. It shows that the percentage of households within the fuel poverty flag have decreased a little over the years as more people are able to get out of fuel poverty. This could be due to government initiatives, the price of fuel (or rising incomes but unlikely). We know from analysis in our notebook that this figure standard is around 8% in 2019 which is also the same as found in our pooled master file. The decrease is seen more clearly in Fig2. What is interesting here is that households in fuel poverty suddenly increased from 2017 to 2019. This can be explained as during this time the UK experienced fuel shortages and so fuel costs increased resulting in households to be in fuel poverty. There was also the Brexit negotiations during this time, many households would have suffered from income loss due to jobs being removed and businesses that trade with the EU were failing. The increase in 2019 can be down to many people losing jobs due to COVID, as restrictions blocked people going to jobs that required to be within the office building. This resulted in a loss of income for many households and many struggled to pay for basic necessities such as fuel for the home.

Fig3 in Appendix 2 shows the number of houses in each FPEER band from 2014 to 2019. Over the years the number of houses that have a rating band of 1 have increased and the number of houses with a rating band of 5 have decreased. The majority of houses have a rating band of 2 during 2014 to 2017, this meant most houses were energy efficient and required less fuel to heat up homes resulting in less fuel expenditure.

Fig4 in Appendix 2 shows the number of houses in each FPEER band for all the years combined. This shows that the majority of homes are in rating band 2 which is good as this means their homes are very energy efficient. It is also reassuring to see that rating band 1 is the second most common for homes. The homes with rating band 5 are not many which is to be expected as many homes are being renovated to ensure they meet the requirements set by the government.

Fig5 in Appendix 2 shows the number of homes in rating band 5 in relation to the households within the fuel poverty from 2014 to 2019. This graph shows more detail on the homes owned/rented by people in fuel poverty and people above fuel poverty. From this, it is clear to see that many more homes with the band rating of 5 are occupied by households in fuel poverty. This is very understandable as these homes were probably much cheaper than homes with better band ratings.

Fig6 in Appendix 2 shows the average housing costs and average fuel expenditure of different households from 2014 to 2019. The household compositions are explained in Appendix 1. The graph shows that housing composition 3, who are couples with dependent children, have the highest housing costs compared to other households. This is understandable as a larger home needs to be bought or rented to accommodate the children and ensure everyone has space according to the law, as the government has restrictions to how many people can live in one dwelling. The fuel expenditure is also high for this household composition as more fuel is needed to heat a larger house compared to a smaller house. The household composition 1, who are couples with no children and are under 60, experience high housing costs behind household composition 3. This can be due to housing prices, which increase every year and therefore making mortgage prices increase as well. Housing costs now are more costly compared to 10 or even 20 years ago, so for couples under 60 finding homes it will be difficult to find anything affordable which is why the cost is so high. The household composition 2, who are couples over 60 with no children, have the lowest housing costs among all the different household compositions. This can be explained as a couple over the age of 60 could have already purchased a house 10, 20 or even 30 years before when housing prices were far lower than they are now. There are a few reasons this could be much lower compared to other household compositions.

Fig7 in Appendix 2 shows the number of households in different income deciles from 2014 to 2019. The graph shows that most households are placed within the second decile which means the majority of households have very low income and if faced with a sudden increase of fuel expenditure this could push many of the households into fuel poverty. The decile with the least households is the tenth, this is expected as this decile is for households that have the highest 10% annual income. The first decile contains households that are in fuel poverty. The number of households in income decile 1 have fluctuated over the years. A decrease over 2014 to 2016 and then an increase from 2016 to 2019. This means that the same number of households are in fuel poverty now as in 2014. Which shows that not much has been done to decrease the number of households in the first full income decile.

Models:

To give each group member the experience of modelling, each of us sought to try one binary classification algorithm. See the Model Results section.

Logistic Regression: Logistic Regression is a binary classification algorithm used to predict binary response variables. It does not predict the exact category that a given observation should be in, but a *probability* of it falling into category '1'.

Decision tree: Decision Trees are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. It is simply a series of sequential decisions using each variable to eventually reach a specific result (i.e. which category).

Random Forest: A random forest is a supervised machine learning algorithm that is constructed from a number decision tree algorithms. Each tree only looks at a randomly selected subset of the complete feature set. By randomly selecting subsets of features, some trees can isolate more important features and improve the accuracy.

Model results

All models perform extremely well on accuracy, but a closer look at the Precision and Recall for a 'positive' fuel poverty reveals a different story. The models all accurately identify 62-64% of those actually suffering from fuel poverty, but of those it deems to be in fuel poverty, it is only correct 72-76% of the time. It actually does much better when it comes to those not in fuel poverty (not reported here, see Jupyter notebook for exact results however both precision and recall are >90% in every model.) The result is due to a very unbalanced dataset: 92% of the data is NOT in fuel poverty and we are trying to capture just 8%. Hence, if any of these models was implemented for government policies, it would likely face criticism of not helping those in need and incorrectly giving to benefit fraudsters.

Table 1: Models on unrestricted dataset

Metric/Model	Logistic Regression	Decision Tree	Random Forest
Accuracy	0.94	0.95	0.95
Precision (FP=1)	0.72	0.76	0.76
Recall (FP=1)	0.62	0.64	0.62

The feature importance (See Appendix 3) in the Decision Tree and Random Forest are very close (and that's to be expected, after all there are only 4 features!) Overwhelmingly the income decile is the most important whereas the energy efficiency band of the dwelling is only half as important. Likewise, the logistic regression reveals a negative relationship with income decile and positive with FPEER band (i.e. the smaller your income or the higher the energy efficiency band, 5 being the lowest, the more likely you are to be in fuel poverty). How much of a surprise is this? Given that we know from the exploratory analysis that a large percentage of those in fuel poverty are in the lowest decile, it's obvious that this is a strong predictor. Furthermore, we also know that the 'higher' (actually lower) the energy efficiency rating, the likelihood of a household being in fuel poverty increases exponentially (50% in FPEERband=5).

Given this result, it is worthwhile looking at the lowest decile alone and seeing if we can improve the accuracy to the benefit of those who are really suffering.

Table 2: Models on Decile 1

Metric/Model	Logistic Regression	Decision Tree	Random Forest
Accuracy	0.76	0.78	0.79
Precision (FP=1)	0.78	0.80	0.85
Recall (FP=1)	0.77	0.79	0.75

At first glance the models perform worse - the accuracy falls to 76-79%. However, it is a considerable improvement when it comes to those in fuel poverty (which really is the purpose of this model). By focusing solely on the lowest decile, the models accurately identify 75-77% of those actually suffering from fuel poverty (up from 62-64%), and of those the model deems as being fuel poverty, it is correct 78-85% of the time (vs 72-76% previously).

When it comes to identifying those not suffering from fuel poverty it doesn't perform as well, falling to the mid-late 70% for both precision and accuracy. However, identifying those who don't need help is not the purpose of this exercise.

Restricting the model to the lowest decile improves the chances of correctly identifying those in need. Furthermore, there is much less likelihood of overlooking someone in need and considerably less likelihood of benefit fraud taking place.

In terms of features, the decision tree and random forest models again assign the most importance to income albeit less, and the energy efficiency rating increases in importance considerably. Interestingly, the logistic regression determines that income carries no weight at all, and the energy efficiency rating coefficient remains the same. This is encouraging as it implies that the government policy to raise the energy efficiency of all houses to C rating or more would be effective in reducing fuel poverty.

"By how much?" you may ask. Here lies the limitations of these models. A pooled linear regression model on the data would help as the coefficients would be able to identify the determinants of fuel expenditure and forecasts like this could be made.

An additional limitation of our models has been the restrictions in the variables used. It is possible that there are more important variables in the dataset that we have not included (e.g. gas boiler presence, floor space). Reasons for doing this were discussed earlier in this document. It is perhaps unrealistic to collect such variables from every applicant at the Job Centre Plus in need of help tackling fuel poverty.

Conclusion

This analysis has confirmed that low income is the major decider as to whether a household experiences fuel poverty. However, the energy efficiency rating is much more important to whether the lowest decile experiences fuel poverty, much more so than those in higher deciles. Interestingly regions and household composition have little bearing on determining fuel poverty which is perhaps surprising given the variation in consumption that both are likely to bring about. We have observed the likely impact of Brexit and Covid which indicates that we could see fuel poverty numbers increase with soaring energy prices. Time series analysis could confirm what we witness in some of our figures regarding real world events. Whilst the initial model provides obvious results due to the unbalanced nature of the dataset, the restricted dataset provides a very good accuracy when trying to capture those suffering and hence it has potential to be used to determine entitlement to benefits targeting fuel poverty.

Appendix 1: Question framing, Data sourcing and Data description

Given rising energy prices, our initial idea was to try to model the impact of the increased Ofgem cap of the share of household energy expenditure using a linear regression model. Household level data that is representative of the whole country (ideally with a decent history) is not easy to find. Initially we sought to connect a few sources: EPC rating at the postcode level, take an average on a small regional basis (so called 'Lower Layer Super Output Areas' (LSOA) and match this to income deciles as found in the Index of Multiple Deprivation. Using the annual ONS Living Costs and Family Expenditure Survey we would know the energy expenditure share of consumption per decile and could assign it to each LSOA. From here we could create a panel of data and use it for regression analysis using the energy component of the consumer price index.

The problem with this approach is that income deciles are very broad categories which ignore the household characteristics important in determining consumption. Hence, we decided to use source data from the Department of Business, Energy, Industrial Strategy's Annual Fuel Poverty Statistics. This contains a large back history of survey data from the English Housing Survey. A representative sample of 6000 households is taken each year which is representative of the population of England. The survey includes an interview with a household and inspects the condition of the property. From this survey we are able to obtain after-housing-cost income, the energy efficiency rating, household composition, region and an estimate from the BEIS on the household's fuel consumption given the rating and household characteristics.

Whilst a pooled linear regression aimed at estimating fuel expenditure would have been ideal, a forecast would have been extremely difficult. This would require a model to forecast how a change in the ofgem cap translates into the region prices used in the Fuel Poverty Statistics. Furthermore, to accurately estimate the share of fuel expenditure to income, we would need forecasts of income growth at least by decile. A further issue was that the latest available dataset covers the financial year 2018-19 so we would need to apply a number of changes to assume the data as 'current' and apply the forecasted changes. Finally, diagnostic tests in order to ensure accurate statistical inference of linear regression models was not covered in the course (e.g. heteroskedasticity, autocorrelation etc), and without this knowledge there was a high chance of creating and interpreting a misleading model of spurious correlation.

Data description:

fpflgf - 10% definition Fuel Poverty flag - full income definition, 0= not in fuel poverty, 1=fuel poverty. The full annual income of the household, which is based on the net income, including housing benefit, SMI, MPPI and net council tax payments. This includes income for the whole household from all sources, including benefits and savings and investments.

AHCIncomeEQ - After Housing Cost Income Equalised. Incomes are calculated after housing costs to reflect the fact that money spent on housing costs cannot be spent on fuel. Therefore, mortgage and rent payments are deducted from the full income of each household to give an after housing costs measure of income. Once housing costs have been deducted, incomes are also equivalised, to reflect the fact that different types of households have different spending requirements.

GorEHS - Government Office Regions as per the English Housing Survey. This relates to the government office regions now abolished. They are North East, North West, Yorkshire and The Humber, East Midlands, West Midlands, East of England, London, South East, South West. It is not documented which number relates to which region in this dataset documentation. Unfortunately the BEIS did not reply to repeated emails.

FullincDeciles - The first decile relates to the households with the lowest 10% of annual fuel poverty full incomes and the tenth decile relates to the households with the highest 10% of annual fuel poverty full incomes. This varies by year.

FPEERband - This is the Fuel Poverty Energy Efficiency Rating Band. is a measure of the energy efficiency of a property based on the Standard Assessment Procedure (SAP) but accounts for policies that directly affect the cost of energy. Similar to SAP, the FPEER methodology generates a rating between 1 and 100, which is then translated into an energy efficiency Band from G (lowest) to A (highest).

The bands are as follows:
FPEER band: SAP rating
1: A-C
2: D
3: E
4: F
5: G

hhcompx - Household composition. This variable shows the type of people who live in the household.
The different categories are:
1: Couple, no children, under 60
2: Couple, no children, over 60
3: Couple with dependent children
4: Lone parent with dependent child
5: Other multi-person households
6: One person under 60
7: One person 60 or over

fuelexpn - Total fuel costs (£). The value in £/year for the cost to the household of the fuel they use for space heating, water heating, lights & appliances energy use and cooking energy use.The fuel cost is based on BREDEM modelled consumption (characteristics of the dwelling, household composition, energy efficiency rating among others) and fuel prices, which vary by region and method of payment.

Appendix 2: Exploratory analysis charts

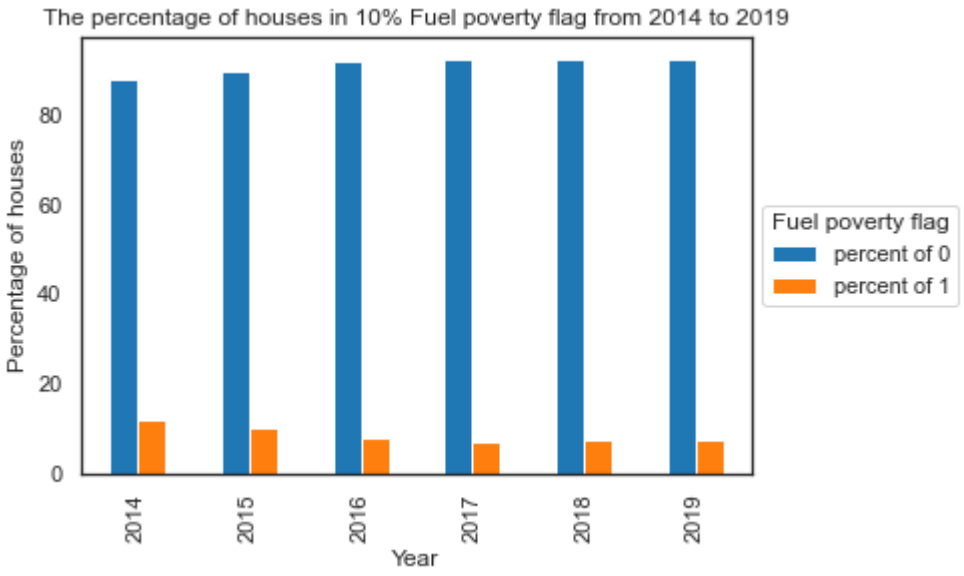


Fig1. The percentage of houses in 10% Fuel poverty flag from 2014 to 2019

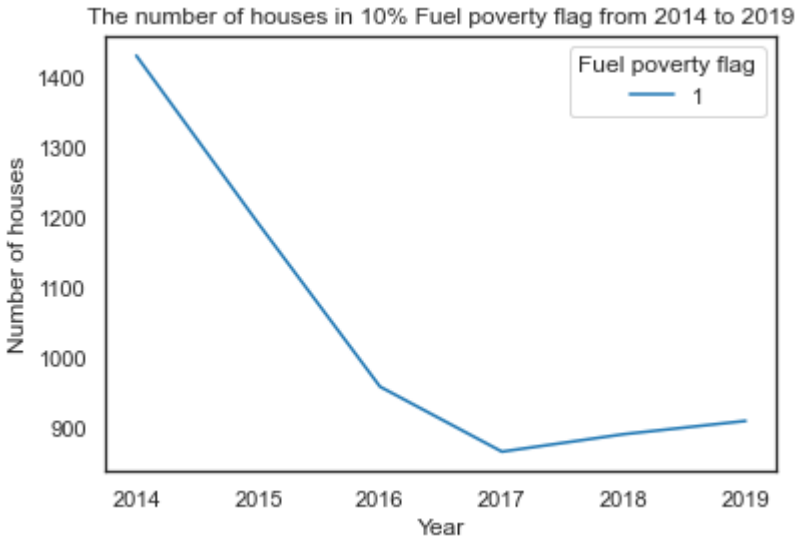


Fig2. The number of households in 10% Fuel poverty flag from 2014-2019

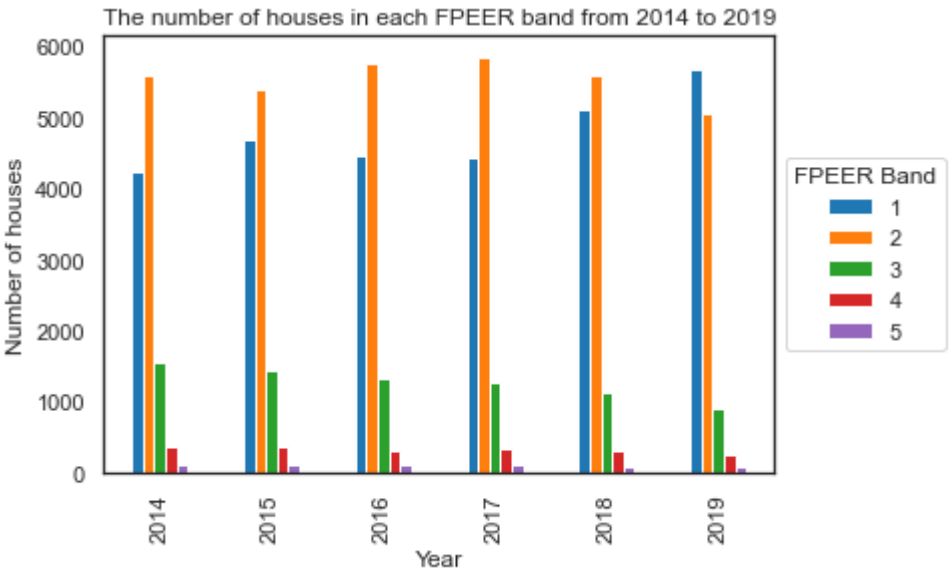


Fig3. The number of houses in each FPEER band from 2014 to 2019

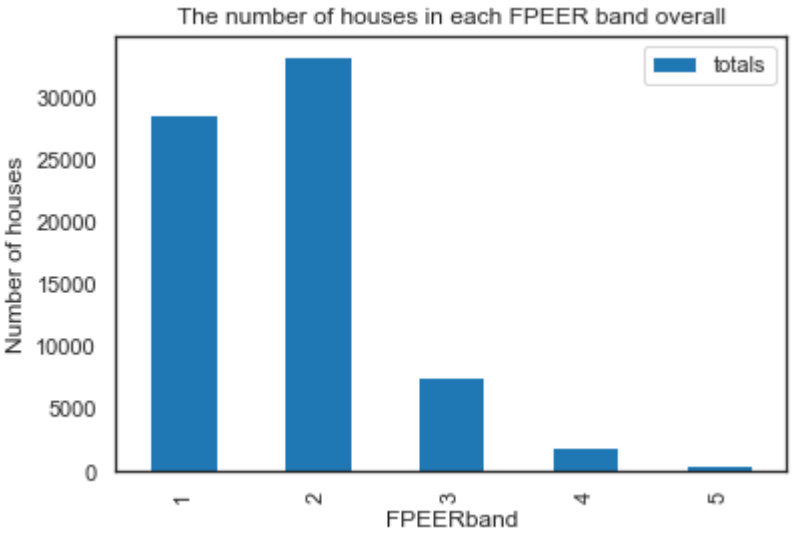


Fig4. The number of houses in each FPEER band overall

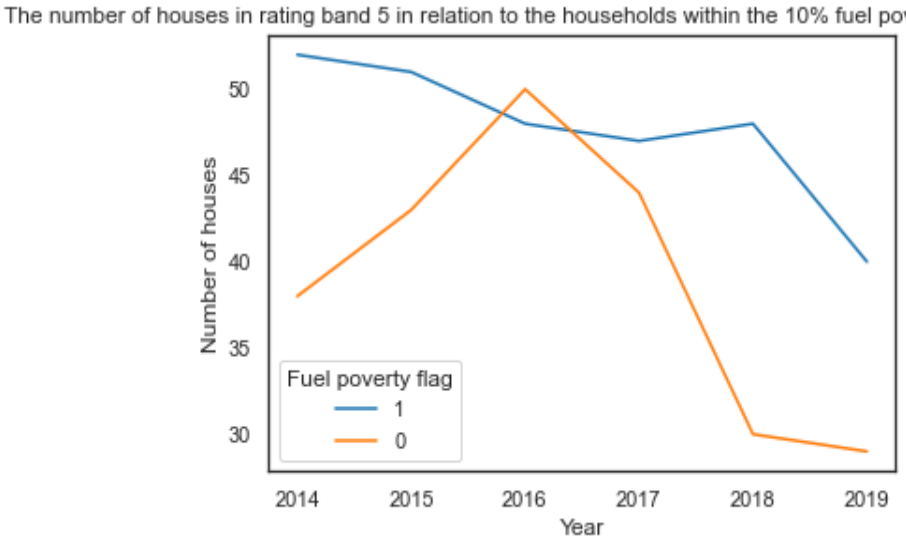


Fig5. The number of houses in rating band 5 in relation to the households within the 10% fuel poverty flag from 2014-2019

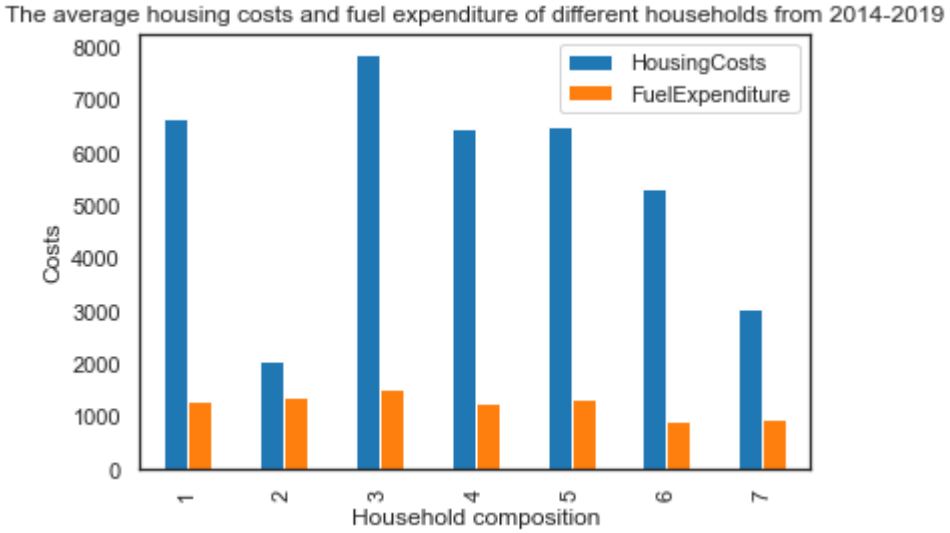


Fig6. The average housing costs and average fuel expenditure of different households from 2014-2019

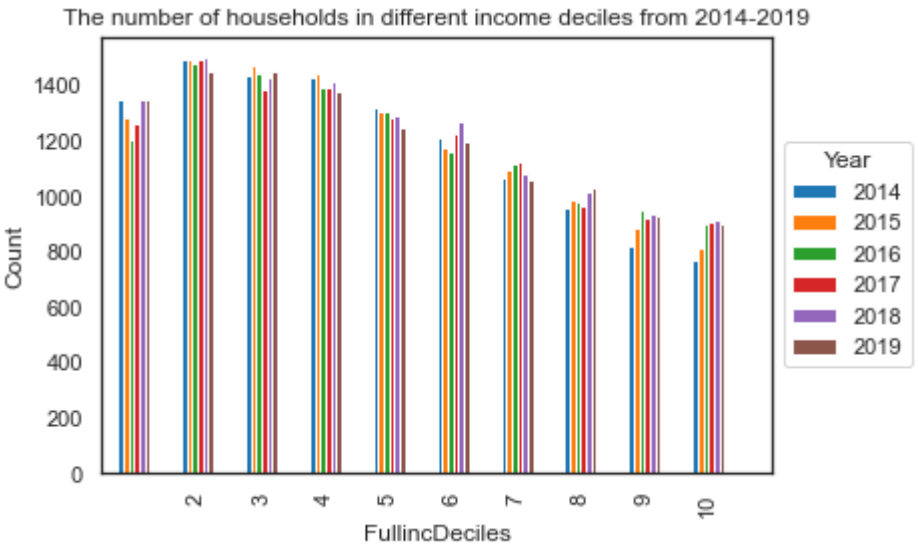
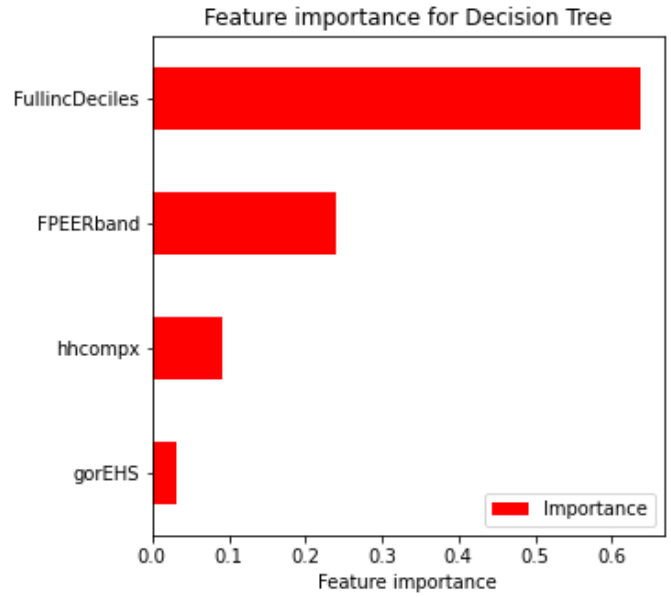


Fig7. The number of households in different income deciles from 2014-20

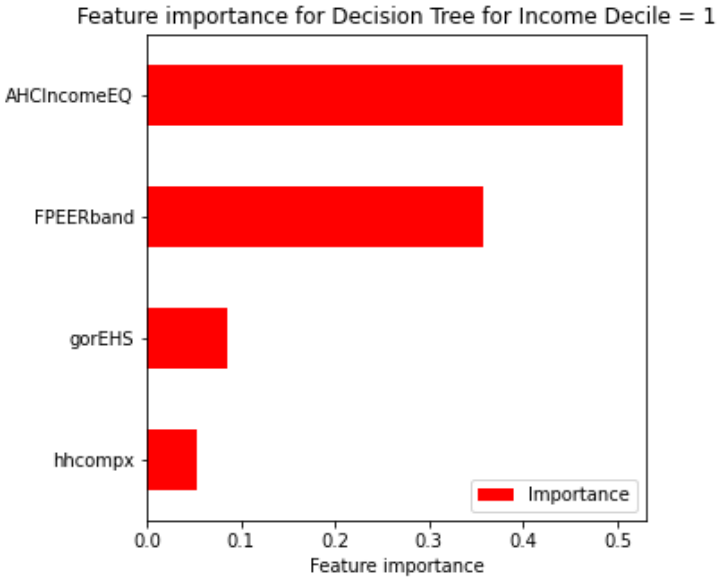
Appendix 3: Results charts

Decision Tree:

Unrestricted model

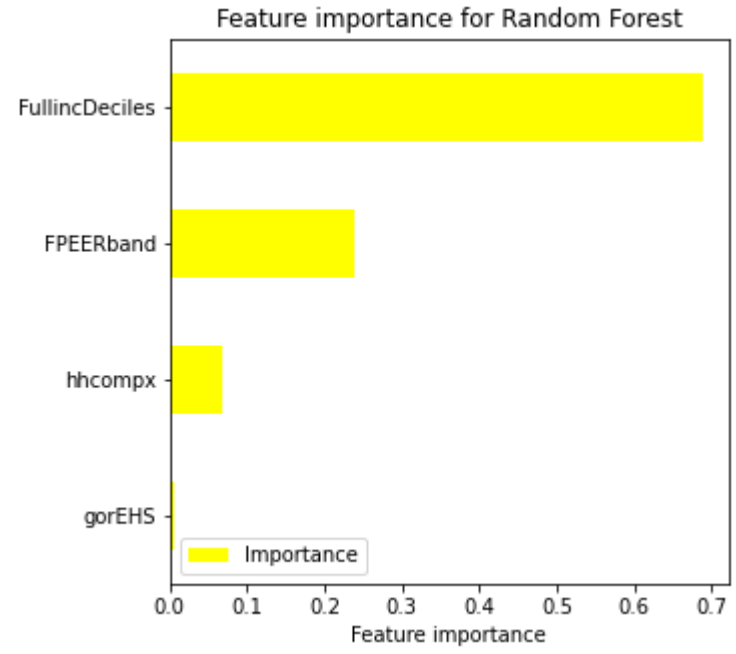


Restricted model to Decile 1

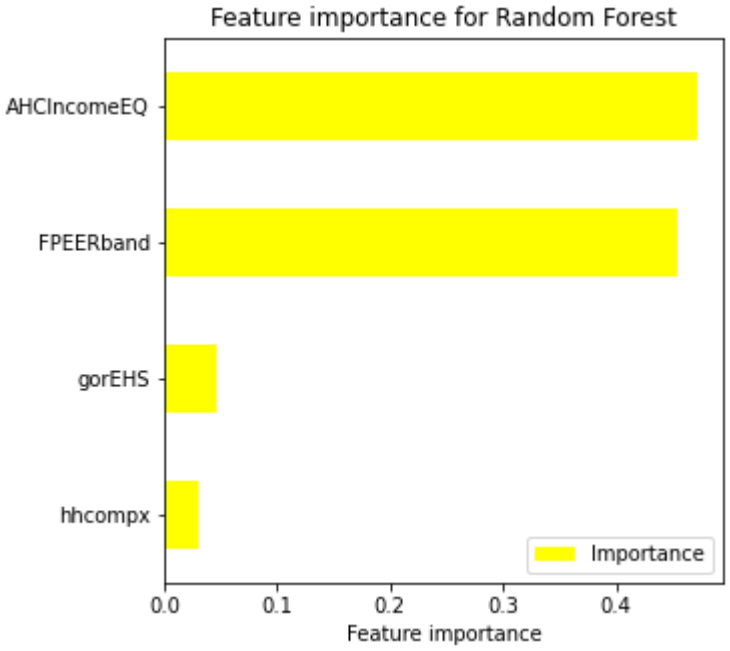


Random Forest:

Unrestricted model

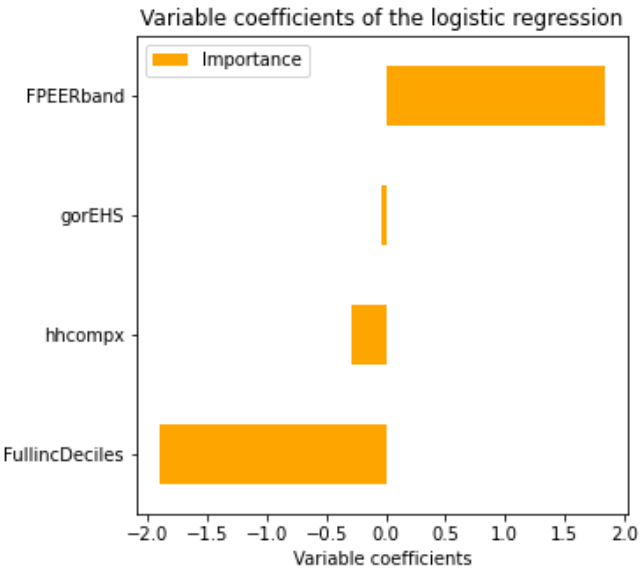


Restricted model to Decile 1

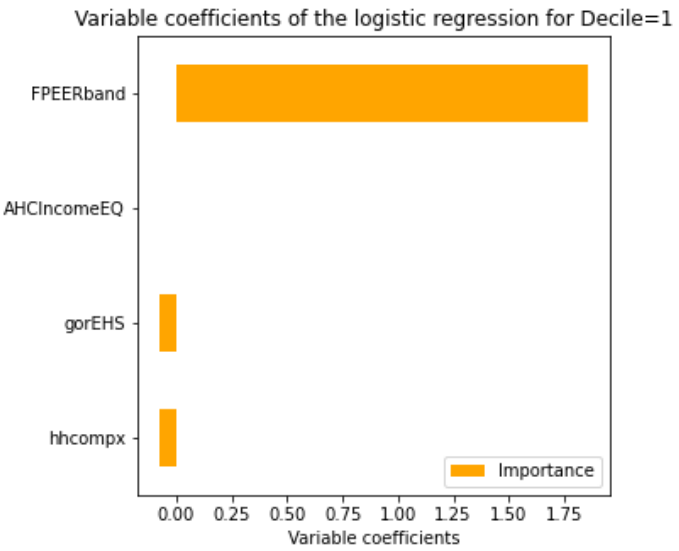


Logistic Regression:

Unrestricted model



Restricted model to Decile 1



Appendix 4: Implementation & Execution

We did not assign discrete roles to one another due to the fairly limited scope of the project and the small group size. Rather than this, each group member took on responsibilities that related to their skillset. To see examples of discrete coding responsibilities, please check the headings in the project notebook to see attributions.

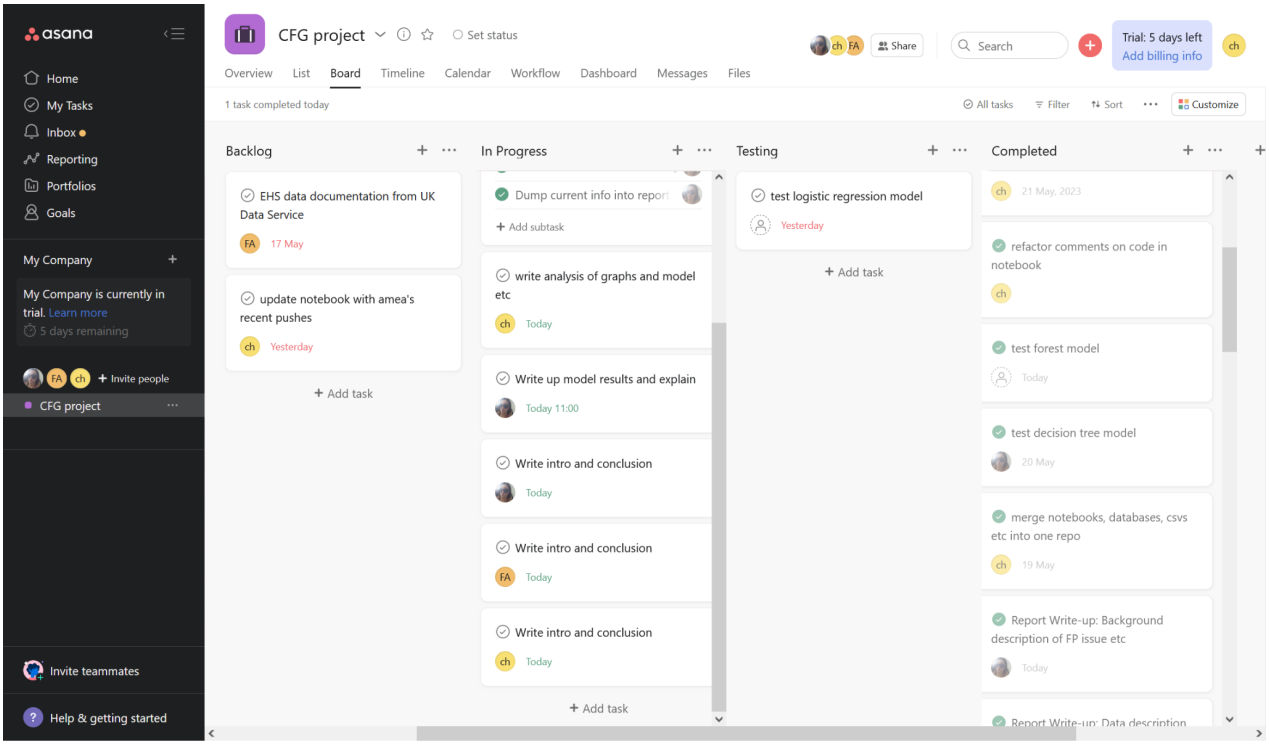
Amea demonstrated a strong leadership ability and took on the responsibilities of a project manager. Part of this required Amea to liaise directly with our contacts in Experian in order to best leverage their resources to aid us in the completion of this project. In addition to this, Amea was extremely proactive with regards to communication, task definition and allocation and proved invaluable by setting the project pace and providing momentum. In terms of coding and technical aspects, Amea was responsible for model data preparation and wrote the code that underpinned the Decision Tree Learning Models.

Farha demonstrated her coding proficiency by authoring many of the graphs present in the project notebook. She also took part in data processing by creating a mySQL database and python connector that were used in conjunction with the CSV files. Farha also demonstrated a strong understanding of graphical statistical analysis, providing many of the key insights present in the project notebook. The Random Forest Machine Learning Models were authored by her utilising the mySQL database and connector.

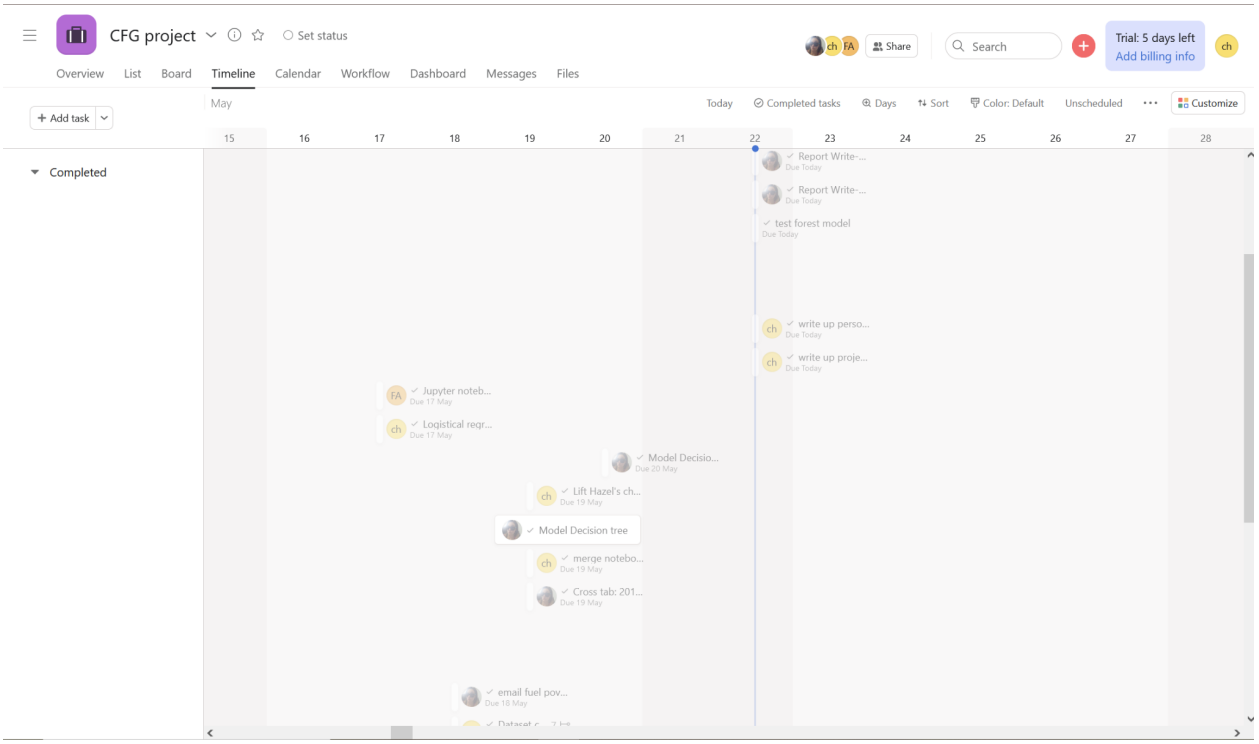
Hazel demonstrated an eye for detail and consistency by writing and performing the initial data cleaning code and managing version control for the project notebook and datasets. A key part of version control was cross referencing any uploaded versions of the project notebook to ensure the consistency of python comments and regrouping sections of code to prevent code duplication. The Logistic Regression Learning Model was written by her and she performed some perfunctory analysis on these results.

The project was undertaken using a mix of agile and waterfall methodologies. In terms of communication and code structuring, we used an approach that relied heavily on agile techniques. For example we held nearly daily meetings to check in with progress and to identify any potential issues that members are facing. In addition to this an iterative coding approach was used to ensure that each code block functioned before building subsequent blocks. This approach also leaned somewhat into the waterfall approaches. For example, the overall project structure was quite rigid in nature. It followed the following structure; brainstorming -> data sourcing -> data processing -> model construction -> code review -> results analysis.

A wide variety of tools were used to facilitate this project. In order to facilitate day to day task management we used an application called Asana. Asana comprises an interactive task board (Backlog, In Progress, Testing and Completed). Once tasks were decided and delegated, they were assigned to the team member, along with a due date. This was important in allowing team members to see how the project was progressing. Should team members be unable to attend a meeting, they could check their Asana tasks to see what work needed done, and helped us understand when we were struggling with certain tasks. We used a mix of stand alone tasks and larger tasks with subtasks in order to delegate work fairly. An example screengrab of the Asana dashboard can be seen here;



In addition to the kanban board, Asana also creates a timeline based on the tasks assigned and completed by a given date. This would represent a gantt chart like structure, which would allow us to review activities and the amount of time they took in order to reduce the time each activity took in future coding sprints. An example of the Asana timeline in action is below;



We opted to use the coding language python in order to complete this project. There are several reasons for this. Firstly, python is open source which means that we did not need to worry about licensing. In addition to this, python is flexible enough to facilitate both procedural and object oriented programming techniques. Finally, python is an extremely forgiving programming language, featuring dynamic typing and no problems with variable capitalisation.

During the course, we were introduced to jupyter labs and its notebook module. These are powerful tools that can be used to perform statistical analysis using python as the basis. Like python itself, jupyter notebooks are free to use and reduce reliance on paid software.

Finally, we used several python libraries to construct our models. In order to read and process data we used both Pandas and Numpy. Pandas allows csv files to be read and viewed in jupyter notebooks as a data frame, and this data frame can be used in order to produce data visualisations. We used Matplotlib and Seaborn to make easily readable data visualisations (Seaborn being built on Matplotlib's capabilities) and supplemented these with statsmodels. Finally we used SciKitLearn to construct the test/train datasets and power the machine learning models.

File version control was managed via google drive and GitHub - google drive allows fast synchronous collaboration across documents and GitHub provides a safe way to view, edit and manage file updates. Data files themselves consist of .sql files (which hold the Fuel_poverty database's table values), .csv files (which hold the data values being manipulated) and a .ipynb file which is the document that holds the code and outputs.

Our main medium of communication was via Slack, which made sending messages to ask and answer brief queries extremely convenient. We used a mix of Zoom and google meetings to facilitate in person meetings, which allowed us to screen share in order to highlight issues and things needing to be fixed.

Challenges faced on this project fall into three broad categories; time commitments, knowledge gaps and data availability.

As this project was undertaken in a non-work setting, the amount of time each group member had to dedicate to the project fluctuated in response to prevailing circumstances. This meant that the workload could be imbalanced at times.

We all came on this course to learn, so knowledge gaps were to be expected. Due to information being unavailable until certain points in the course, we often created issues by trying to use more complex techniques, only to realise that there was an easier way to do it. Some of the knowledge gaps present relate to modelling (machine learning modelling as opposed to traditional man-powered modelling) and the language and coding backbone necessary to perform this. Another issue was the mathematical underpinning of certain statistical measures. As the focus of this course was not on machine learning itself, some of these measures were glossed over in favour of time. Finally, the group is not specialised in this subject area - each of us come from a different academic background. As such, there are elements of the dataset and the subsequent analysis that were difficult to process and slowed down progress.

Finally, as briefly mentioned in the report, sourcing the 'perfect' dataset to answer our initial question proved quite difficult. Given more time and the ability to access larger (possibly private) datasets would most likely have led to a more robust, rounded modelling experience.

Appendix 5: Reflections from the team

Hazel:

Personal statement

Responsible for maintaining and marrying jupyter notebook versions. Coded initial data processing, and wrote notes on this in the notebook. Coded the Logistic Regression machine learning model, but did not write comments. Drew some simple graphs to demonstrate relationships between elements of the dataset.

Things I learned

The limitations and difficulties surrounding branch merging and management on github. How to use python (and possibly other oop languages) in order to perform statistical analysis and machine learning. Using a kanban board to assign and track tasks. Refreshed basic statistical knowledge and learning new concepts (regression models, decision trees).

Limitations

Not agreeing on certain conventions beforehand led to version control and consistency being a bit difficult. A similar agreement for programs and processes would have been invaluable as well. Time to allocate to this project was limited - attending a workload heavy university course whilst studying a new subject area meant there was not enough time to fully digest some concepts.

Farha:

Personal Statement

Created SQL database with the relevant tables needed to run queries. Connected SQL with jupyter notebook to run queries and plot graphs to view relationships between the variables. Queries converted to graphs and explanations added in jupyter notebook. Exploratory analysis of graphs. Random forest tree models, feature importance graphs explained.

Things I learned

Modelling data and viewing the accuracies and errors. To use github to collaborate with the team for many different files. Using SQL database to produce graphs. Learning to divide out tasks and using project management tools like Asana.

Limitations

Duplication in some graphs and using csv in some parts and SQL database in others resulted in the notebook running into some errors initially. Variables are not the same so jupyter notebook throws back errors. Time constraints: Currently working full-time and at university full-time in my final year.

Amea:

Personal statement

Project managed the process which started with designing the core question, sourcing of data, design of data sampling and test/train, selecting the modelling process and allocating tasks. Wrote template for Jupyter notebook and project document, undertook exploratory analysis and consistency between models. Own model was the decision tree.

Things I learned

Git version control and it's limitations with notebooks. Exploratory analysis techniques and how to frame data to answer questions. Interpreting different machine learning models.

Limitations

Project planning and designing a question to address without yet knowing any machine learning techniques. Communication on changing and updating the project as relevant material was taught made the scope of the project change constantly.

References

BEIS (Department for Business, Energy and Industrial Strategy) **"Annual Fuel Poverty Statistics in England, 2022"**

BEIS (Department for Business, Energy and Industrial Strategy) **"Fuel Poverty Dataset Documentation 2020"** (Accessible through UKDataService upon downloading data.