DSA8023 Analytathon 1: Identification of EV users using Mixed Classification models

Varun Suresh Kumar (40364111)

01-06-2023

Table of Contents:

1. Introduction

Electric or hybrid vehicles are accounting for 20% of all new passenger vehicle sales by 2022. With over 70,000 EVs on the roads in Ireland (ROI), as of Q1 2023, the proportion of new EV ownership is growing rapidly (Prendergast, 2023). Energia assumes its market share as 10% of all electric vehicle users. Finding the remaining suspected EV users who are not enrolled in Energia's EV Tariff is the challenge. In the existing database, around 2,400 consumers have been identified as the EV Tariff consumers, indicating a considerable untapped market potential. The research on customer's preference to choose alternative energy suppliers and spreading awareness on the customer benefits of Energia's EV Tariff, is crucial for Energia's market share growth.

Energia can construct Classification models using ML algorithms that assess the likelihood of consumers acquiring an EV based on a variety of data points such as energy use trends, demographic information, and geographic location. These models can help Energia prioritise and target potential EV purchasers within its customer base who have not yet opted into the EV Tariff. By combining these ML predictions with additional sources of data, like vehicle registration databases and public charging station consumption data, Energia can improve the accuracy of identifying suspected EV customers.

This report will employ a number of research methodologies and analytical techniques to uncover an untapped market of prospective EV purchasers in Ireland. We will get valuable insights into the qualities, behaviour, and preferences of the target audience by examining customer data, conducting surveys, and researching market trends. These insights will assist Energia in developing data-driven strategies to meet the specific demands of the undiscovered niche, matching its offers to consumer expectations, and positioning itself as a leading provider in the EV energy market. By bridging the gap between recognised EV consumers and the untapped sector, Energia can contribute to the long-term growth of the EV industry, reduce carbon emissions, improve retention and onboarding of customers, provide better charging infrastructure, improve customer support.

2. Statement of the problem:

The challenge to Energia is to attract and convert an untapped segment of Irish Electric Vehicle (EV) owners who have either chosen other suppliers or are ignorant of the benefits offered by Energia. Despite having attracted 10% of the targeted 70,000 EV clients, Energia's database only shows 2,400 EV Tariff users. The main goal is to improve the EV user identification by employing specialised approaches that successfully explain Energia's benefits and resulting in a greater adoption rate among the untapped EV owner market. Energia can strengthen its market position and become the preferred provider.

3. Research Objective:

The objective of this study is to find non-Electric Vehicle (EV) consumers who have comparable socio-demographic profiles and monthly payment habits to existing EV users. The goal is to enhance the chance of these non-EV consumers adopting EVs by targeting them with targeted marketing campaigns and personalised solutions. Ultimately, this project will broaden Energia's client base, encourage sustainable mobility alternatives, and help to shape a better future.

4. Data overview:

The provided dataset contains **1,86,558 observations** and **25 variables** related to customer information and their bi-monthly billing data over a two years.

The variables are accountID, StartDate, ContractStartDateEV, contractStartDate, contractEndDate, saStatus, agedBand, signedUpGroup, title, mosaicType, EV, EV_New_or_Old, bill_1_2021, bill_2_2021, bill_3_2021, bill_4_2021, bill_5_2021, bill_6_2021, bill_1_2022, bill_2_2022, bill_3_2022, bill_4_2022, bill_5_2022, bill_6_2022, bill_1_2023. The **EV** variable is the target label.

4.1 Description of variables:

There are 13 billing variables *bill_1_2021* to *bill_1_2023* that represents the billing information of customers from January 2021 to January 2023. The billing variables are continuous numeric variables whereas the remaining variables are categorical.

- ***accountID*** : unique identifier for each account
- ***startDate*** : Start date of customer
- ***ContractStartDateEV*** : EV contract start date
- ***ContractStartDate*** : Energia contract start date

- ***ContractEndDate*** : Energia contract end date
- ***saStatus*** : Service status of customers
- ***agedBand*** : age band of customers
- ***signedUpGroup*** : customer's channel of signup
- ***title*** : title of customer
- ***mosaicType*** : socio Economic class of customers
- ***EV*** (target label) : Whether customer has EV or not
- ***EV_New_or_Old*** : Is EV vehicle new or old
- ***Bill_1_2021*** to ***Bill_1_2023*** : Bi-billing information of customer

4.2 Exploratory Data Analysis:

Missing Values

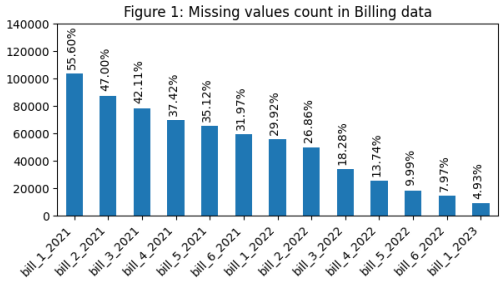
Based on the Table 1, *ContractStartDate* and *mosaicType* have a significant percentage of missing values in the dataset. There are no missing values in the following columns *StartDate, saStatus, signedUpGroup, EV, EV_New_or_Old*.

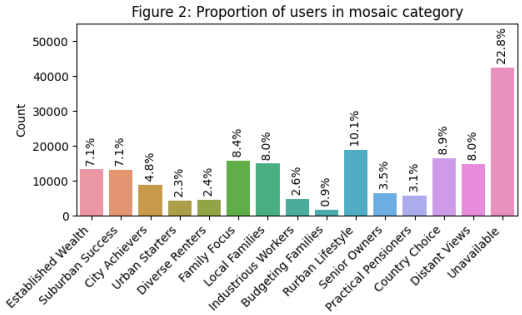
Table 1 Significant Missing values in the dataset

Variable	Description	No of Observations	Missing data (count)	Missing data (%)
ContractStartDateEV	EV contract start date	2516	184042	98.650000
mosaicType	Socio Economic class of customers	144036	42522	22.790000
contractEndDate	Energia contract end date	185591	967	0.520000
contractStartDate	Energia contract start date	185981	577	0.310000

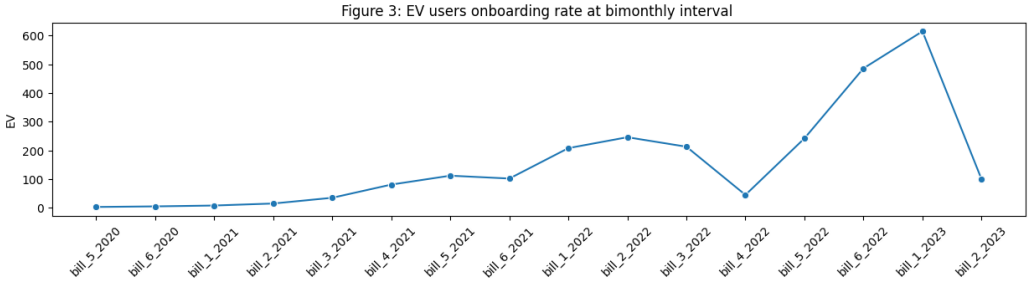
On the basis of the study of the billing data, there are a lot of missing numbers, as shown in Figure 1. Notably, the category with the highest percentage of missing values is bill_1_2021. Nevertheless, it is reassuring to see that the number of missing values has been declining over time. This pattern points to an increase in the reliability and completeness of the data. On the other hand, when taking into account EV consumers who signed up in 2022, a special observation becomes apparent. Their recent enrollment makes missing values for the previous year more common. However, as shown in Figure 3, the study based on enrollment shows a downward trend in missing values within the invoices. These results demonstrate the overall improvement in collecting accurate and thorough billing data over time, emphasising the benefit of enrolment.

In dataset, a variety of Mosai factors have varying percentages of users as members (Figure 2) . In terms of user population, factors like Established Wealth, Suburban Success, Family Focus, Local Families, Rurban Lifestyle, Country Choice, and Distant Views account for 7% to 10%. Conversely, less than 4% of users fall into categories such as City Achievers, Urban Starters, Industrious Workers, Budgeting Families, Senior Owner, and Practical Pensioners.





From Figure 4, 53% of the participants are classified as Mr. In the Mr group, non-EV users make up 52% of the population, while EV users make up only 0.9% of the total. Ms, Mrs., and Miss, on the other hand, make up 42% of the dataset as a whole. Only 0.4% of people in this category are EV users, leaving the majority of people to be non-EV users. These results provide insight into the distribution of EV users by title, showing that the Mr group has a higher percentage of non-EV users and that the Ms, Mrs, and Miss categories have a considerably smaller percentage of EV users.



The distribution of users among various categories is shown in Figure 5. 16.9% of users fall into the PCW category, with 0.2% of them being EV users and 16.7% of them being non-EV users. 21.3% of users fall into the Telesales group, with 21% of them being EV users. 14.9% of users are in the Web category, of which 14.5% are non-EV users and 0.4% are EV users. A minimum of 0.1% of users are solely EV users, with field sales and OBS being the other 23.6% of users. The remaining 29.4% of customers fall into the "Others" category, with 0.3% of them being designated as EV users.

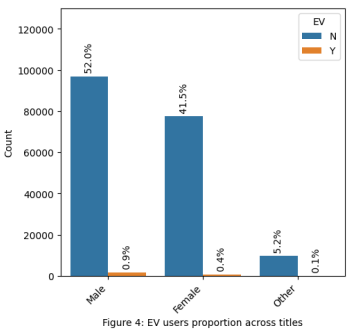


Figure 4: EV users proportion across titles

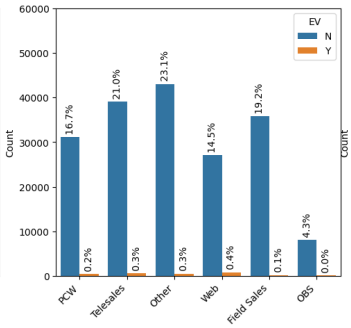


Figure 5: EV users proportion in Signed up Group

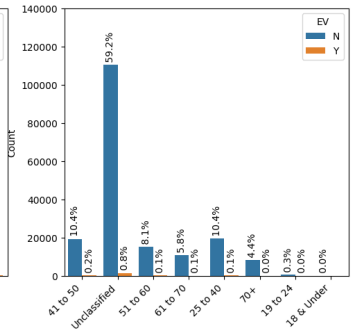


Figure 6: EV users proportion across age groups

From Figure 6, 10.5% of the population is between the ages of 25 and 40, with 10.4% of that group using electric vehicles and 0.1% of the population are solely EV users. When it came to customers who were between the ages of 41 and 50, 10.6% of them were found, and only 0.2% of them were EV users. A significant 60% of users were found to be between the ages of 51 and 60, and 0.8% of these users were solely electric vehicle (EV) users. Only 0.1% of the 5.14% of customers who were 60 and older were EV users.

It is significant that 60% of users belonged to an unspecified age bracket, and that only 0.8% of this group were found to be EV users. These results demonstrate the diverse percentages of EV users across various age groups, highlighting possible chances for targeted advertising and outreach initiatives. No users under the age of 19 were available.

4.3 Measure of Association:

In the feature selection procedure, a univariate analysis is performed to analyze relationship between the EV status and the variables of age group, signed-up factor, mosaic factor, and title. Assuming the null hypothesis that there is no relation between variables and target variable, we perform Chi-squared test to establish association between them. Based on the p-value calculated, the statistic is highly significant at 95% confidence level which shows significant correlation with target variable as shown in Table 2. So we are considering these variables in the Machine learning models.

Table 2 Testing Significance between variables using Chi Square test.

Variable	Null Hypothesis	p-value	Significance
agedBand	There is no association between agedBand and EV variables	3.13213811595283e-09	Significant
signedUpGroup	There is no association between signedUpGroup and EV variables	1.415185184357035e-149	Significant
title	There is no association between title and EV variables	9.310374982118508e-62	Significant
mosaicType	There is no association between mosaicType and EV variables	1.324534407799008e-66	Significant

4.4 Data Wrangling:

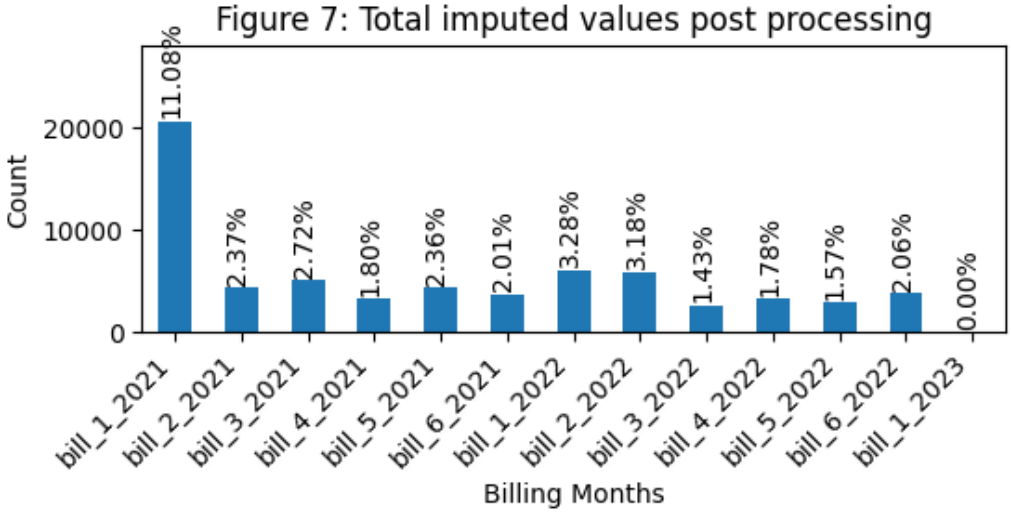
The mosaic factors variables, which originally had 44 sub-classes, were reduced to 15 classes during the data wrangling process to facilitate analysis. Mosaic classes were unknown for about 22.8% of the users, and to fill in the gaps in the data, these cases were labelled as Unavailable. Classes by age group, title, and signedup group were encoded for model training. Each customer's 13 invoices from the dataset,from January 2021 to January 2023 are included. The lack of values for earlier invoices for members who joined the tariff in 2022, however, had a negative influence on the accuracy of our models. At least 44% of the users had one missing bill out of a total of 186,558 participant records. We set a threshold of eliminating observations with greater than eight zero's to ensure data quality. After this cleaning process the dataset has 143,849 observations and they are considered for modelling.

Based on the fact that, if reading for a billing cycle is missed, the reading will be included in the next billing cycle. So average reading value is applied for previous and current month. Only one intermittent zero value is filled this way Consecutive zero values are ignored. Notably, Bill_1_2023 did not undergo imputation since forward values were not available.

The findings showed in Figure 7, that for Bill_1_2021, almost 11% of the missing data were imputed, compared to least imputation variables, which ranged from 1.5% to 3%, this shows zero values are significantly reduced in Bill_1_2021.

SMOTE (Synthetic Minority Over-sampling Technique) sampling was used on the dataset to rectify the imbalance between minority and non-minority classes. By interpolating between instances of the minority class that already exist, SMOTE creates synthetic samples for the minority class. SMOTE was only applied to the EV (Electric Vehicle) user class, which represented the minority class in our dataset. With 1,678 EV users and 142,171 NON-EV users in the dataset, we were able to achieve balance by creating new instances for the EV users using SMOTE. This equal representation of the two classes made it possible to create a model that could manage the class disparity and produce reliable predictions for both EV and Non-EV customers.

Figure 7: Total imputed values post processing



5. Methodology:

From the given dataset of 186,558 observations, which reduced to 143,849 observations by cleaning and pre-processing techniques. From our initial analysis, we observed our dataset is highly imbalanced. To solve this problem, Synthetic Minority Oversampling Technique (SMOTE) was used to synthetically generate additional instances on the training data for analysis. Logistic regression, Naive Bayes, and K-nearest neighbours (KNN) models were used to classify from EV users from non-EV users those exhibit similar features as EV users. Naive Bayes does classifcation based on computed probabilities on feature occurrences and probabilities. Logistic regression is a binary classification model that computes log arithmatic for features that over comes non linearity. KNN utilised feature similarity score without assuming a specific data distribution to assign new data points to the target class with its closest K neighbours. Based on feature values and related probabilities, it determines if a user is an EV or non-EV, and then chooses the class with the highest likelihood as the anticipated class. A new data point is assigned to the majority class among its k nearest neighbours in the feature space by the non-parametric classification method K-nearest neighbours (KNN). Without making any explicit assumptions about the distribution of the data, KNN predictions are based on feature similarity.

After model building, we tested their performance on a separate test dataset of 503 EV users and 2013 non-EV users. The primary goal was to identify these user classes using the three models and calculate the number of true positive, true negative, false positive, and false negative classifications produced by each model. This allowed us to examine the model's accuracy, sensitivity and specificity in discriminating between EV and non-EV users. The accuracy of the model predictions is measured by accuracy, while sensitivity (a.k.a., recall or true positive rate) is measured by the proportion of real EV users correctly recognised by the model. The proportion of real non-EV users accurately detected by the model is measured as specificity (a.k.a., true negative rate).

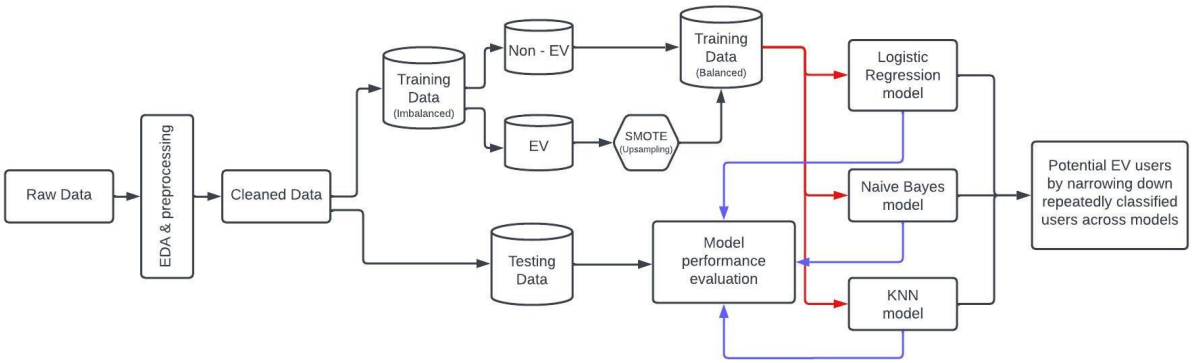


Figure 8: Flowchart of modelling methodology

By integrating the classifications from all three models, we hoped to gain a more robust and reliable categorization of users who are more likely to follow the same pattern as EV users. This strategy used the aggregate insights and strengths of numerous models, possibly enhancing the accuracy and confidence in detecting people displaying similar behaviour to EV consumers.This combined categorization technique might give useful insights for targeted tactics and interventions, allowing for improved understanding and segmentation of users based on their proclivity to adopt EV user habits.

6. Comparison of Modal performance

The below table (Table 3), shows the classification model's findings on how well the models performed in classifying EV (Electric Vehicle) users.

Table 3: Performance metrics and counts of Identified users

Models	Accuracy	Specificity	Sensitity	No of EV users Identified
K Nearest Neighbour	**77.410000**	**59.450000**	**63.820000**	**76684**
Logistic Regression	**84.880000**	**84.870000**	**85.710000**	**22839**
Naive Bayes	**91.160000**	**92.050000**	**32.560000**	**15102**

KNN attained an accuracy of 77.40%, a sensitivity of 63.81%, and a specificity of 59.44%. Logistic Regression performed better, with an accuracy of 84.87%, a sensitivity of 85.71%, and comparatively highest specificity of 84.86%. Naive Bayes had the maximum accuracy of 91.2%, with a sensitivity of 32.55% and a specificity of 91.04%.These findings show that all three models correctly classified EV and non-EV consumers. Logistic Regression had the best overall accuracy,

suggesting a good capacity to accurately identify both sorts of users. It also has a high sensitivity, suggesting a low number of false negatives, and a good specificity, indicating a low rate of false positives.

While KNN's accuracy was significantly lower than that of Logistic Regression, it comparatively performs well in categorising EV users. Among the three models, Naive Bayes scored the best accuracy, reflecting its overall strength. It did, however, have a decreased sensitivity, implying a larger probability of false negatives while keeping a high specificity.Overall, these findings illustrate the models' usefulness in categorising EV and non-EV consumers, with each model displaying its own set of strengths and drawbacks.

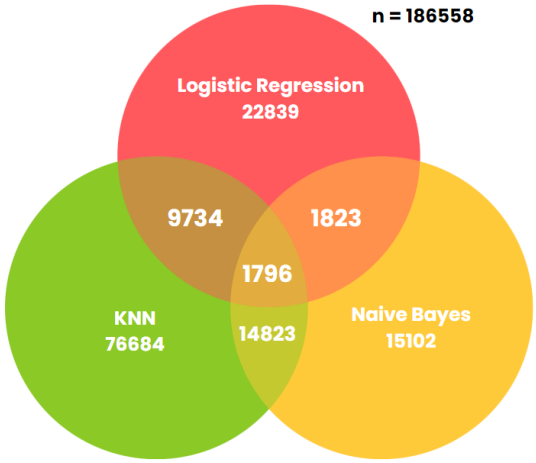


Figure 9: Venn diagram of count of Classified EV users

From Figure 9, it is evidant that Logistic Regression model identified 22,839 users, KNN model identified 76,684 users, and Naive Bayes model classified 15,102 users as more likely to follow the same patterns as EV users. as more likely to be EV users based on characteristics such as Mosai characteristics, Age group, Signed Up Factor, and user titles. By merging two models, it shows 9,734 users were categorised by both Logistic Regression and KNN, 1,823 users were classified by both Logistic Regression and Naive Bayes, and 14,823 users were classified by both KNN and Naive Bayes.

By combining the repeatedly categorised users from the two models, we can improve classification performance since these users are more likely to display patterns similar to EV users. However, when the common users from all three models were combined, around 1,796 persons were discovered. Although these individuals are likely to be EV users, our strategy may deviate on more accurate significant EV users. As we are confident that integrating two models is more efficient, it allows us to infer more EV users without sacrificing critical information.

7. Results and discussion

```
To prove likeliness of classified EV users being truely EV users, we compare the
proportion of distribution of features. Based on Figure-10, which examines
patterns of known EV users and classified users from Logistic Regression based
on Mosai category, Age Band, and Signed Up channel features, the percentage of
proportion of features are substantially identical between the two groups.
```

The Logistic Regression model efficiently captures the patterns and features of existing EV users, allowing for reliable categorization of new users. This similarity in class proportions makes Logistic Regression an ideal model for EV users classification.

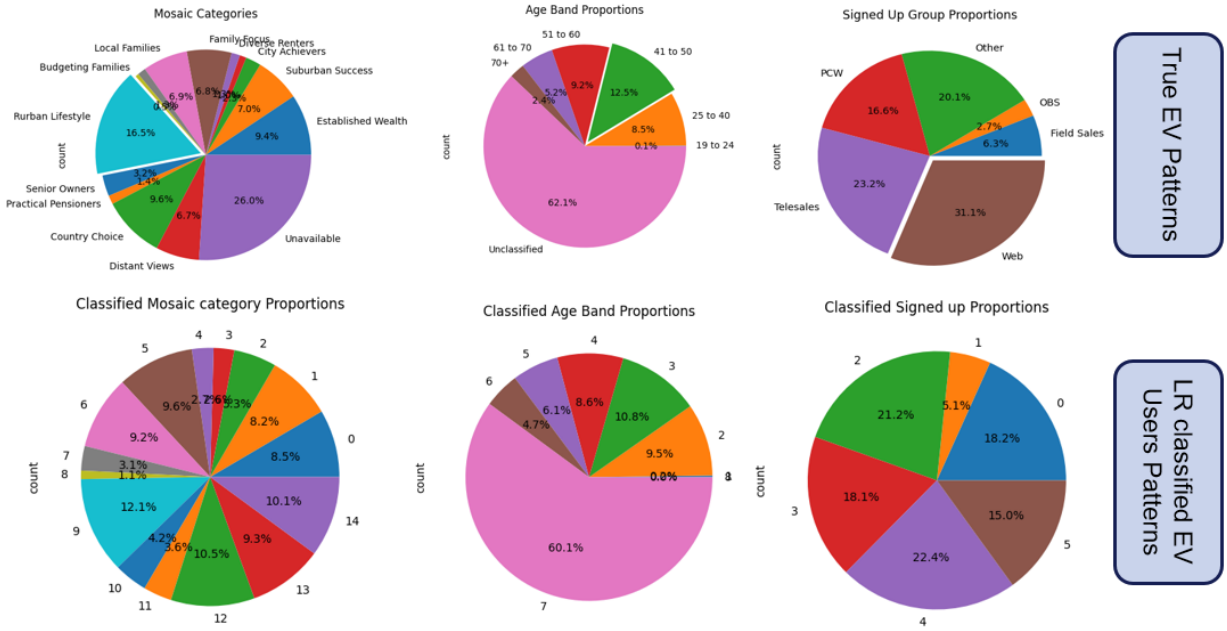


Figure 10: Feature Comparison of Logistic Regression model's predicted users.

From Figure-11, the proportions of classes for known cases and classified users are slightly following the same pattern. However there are some discrepancies in proportions when compared to Logistic Regression.

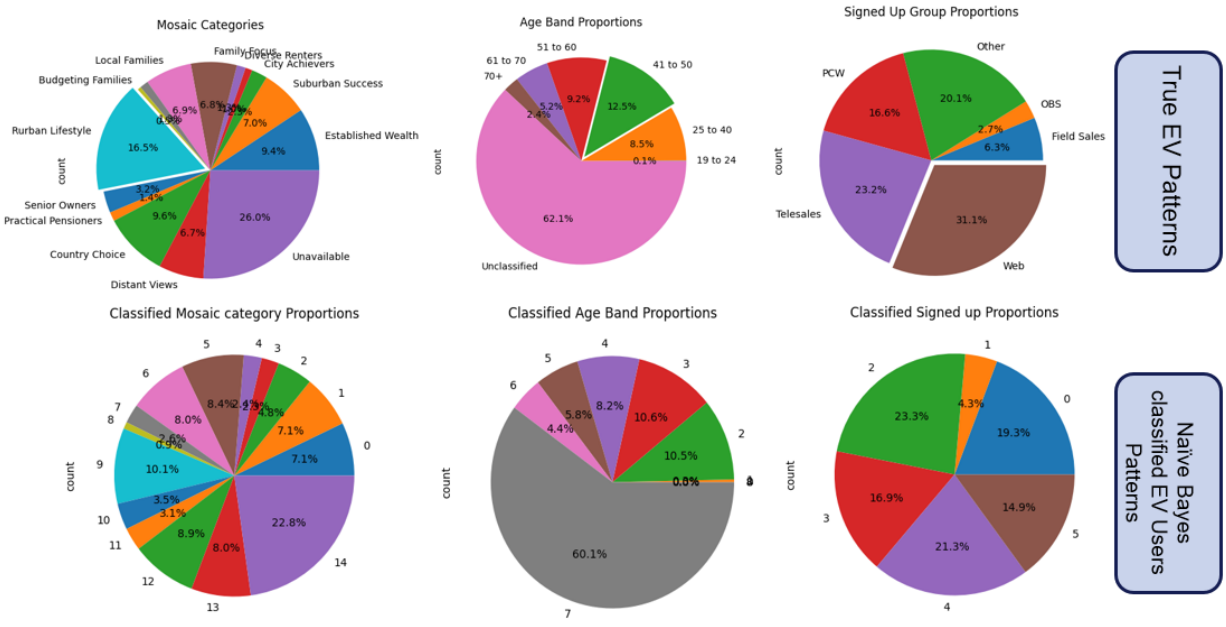


Figure 11: Feature Comparison of Naïve Bayes model's predicted users.

While Naive Bayes has a good overall accuracy, its sensitivity is significantly lower than that of Logistic Regression. Furthermore, the capacity to identify true EV users is fairly limited. These findings imply that, when compared to Logistic Regression, Naive Bayes may fail to reliably distinguish actual EV users. Users who are not categorised using Naive Bayes may have a modest possibility of being EV users.

However, because of increased sensitivity in distinguishing actual EV users, Logistic Regression looks to be a superior candidate for a more robust and accurate classification. In summary, though Naive Bayes has acceptable overall accuracy, it may have lesser sensitivity and mediocre performance in distinguishing true EV users.

8. Conclusion:

Based on our analysis, Logistic Regression had the greatest overall accuracy of 84.87%, surpassing KNN (77.40%) and Naive Bayes (91.20%). Logistic Regression shows a high sensitivity of 85.71%, and far better specificity of 84.86%, indicating a low percentage of false positives. We can conclude Logistic Regression is a good model for effectively categorising EV users and Non-EV users, with a balanced performance in terms of accuracy, sensitivity, and specificity. Furthermore, the findings reveal that the Logistic Regression model is excellent at categorising users based on relevant variables, with promising alignment in terms of class proportions between known EV instances and categorised users. Finally, Logistic Regression exceeds Naive Bayes in effectively categorising EV and Non-EV users based on the features presented, with greater sensitivity and overall performance. In the context of EV adoption, Logistic Regression is advised for more reliable categorization of users, but Naive Bayes may still provide some insights but with lesser sensitivity.

- From the models likelihood to classify EV users, we are 85.71% confident that Logistic Regression will comparatively perform well.
- We are 77.7% confident that the users predicted by combining KNN and logistic regression models exhibit patterns similar to EV users.
- Naive Bayes model performs well in identifying Non-EV users, by eliminating the classified Non-EV users using this model, we are highly suspicious that remaining users will exhibit EV characteristics.

Limitations and Suggestions:

- The provided dataset has bi-monthly information, if we could capture the hourly consumption information of the customers, we can classify better using time series and DTW techniques (Chawla, Nitesh & Bowyer, Kevin & Hall, Lawrence & Kegelmeyer, W.. [2002])
- The provided tabular data has customers starting Energia contract and EV tariff at different dates, this makes the dataset contain more zero /NA values, this makes missing values imputation difficult.

9. References:

- Shane Prendergast, Direction of Travel - the growing EV markets in Ireland (26 April 2023) https://www.seai.ie/blog/ev-direction-of-travel/ (https://www.seai.ie/blog/ev-direction-of-travel/) [accessed 26 May 2023].
- Chawla, Nitesh & Bowyer, Kevin & Hall, Lawrence & Kegelmeyer, W.. (2002). SMOTE: Synthetic Minority Over-sampling Technique. J. Artif. Intell. Res. (JAIR). 16. 321-357. 10.1613/jair.953.