UNSW Business School/
Information Systems and
Technology Management

TEAM ASSIGNMENT COVER SHEET

GroupID: __________T14A-1_____________________________________

Course: __________INFS2822______________________________________

Tutorial Session:__________ T14A ______________________________________ (e.g., H09A)

Assignment Title: Group Assignment

Due Date: ____11/11/2022___________________________________

We declare that this assessment item is our own work, except where acknowledged, and has not been submitted for academic credit elsewhere, and acknowledge that the assessor of this item may, for the purpose of assessing this item:

- Reproduce this assessment item and provide a copy to another member of the University; and/or,
- Communicate a copy of this assessment item to a plagiarism checking service (which may then retain a copy of the assessment item on its database for the purpose of future plagiarism checking).

We certify that we have read and understood the University Rules in respect of Student Academic Misconduct.

zID	Name	Signed	Date
z5359975______	Vinay Venkatesh___	__Vinay Venkatesh__	_10/11/2022___
z5308435_______	Rithe Zaman______	R.Zaman _________	_10/11/2022_____
z5363955______	Jessica Quach_____	_Jessica Quach_____	10/11/2022_
z5341683______	Huong Nguyen____	___________________	_10/11/2022___
z5254871_______	Judith Wei________	__Judith Wei______	10/11/2022____
z5359025_______	Hoang (Tram)_Vo_	___Tram Vo______	___10/11/2022_

This assignment cover sheet is ***valid only with your signature (could be typed due to covid)****. The cover sheet is to be provided with any print or digital submission of the assignment. Failure to include a signed cover sheet will lead to a 10% penalty of the total marks available for the assignment. No marks will be released until a signed cover sheet is received*

Executive Summary

This report highlights the likelihood of customers who upgrade their services (upselling) and those who additionally purchase services (cross-selling). Through this, Eurocom will be able to target this market audience and create more business processes to optimize profits. The primary aim of this research project was to provide evidence and predictive model sets which act like a backbone for this target variable. The secondary aim was to accurately comprehend which variables would be apt for predictors within the data dictionary provided.

Primary Problem	***Statistics***
Identifying EuroCom's Customer Demographics	*10,000 customers*
Identifying EuroCom's Customer's behaviors and preferences influence the purchase	• *Customer ID* • *Upsell/xsell* • *Handset Manufacturer* • *Billed Data Usage*
What influences the likelihood of an individual choosing up-selling or x-selling?	• *Demographics* • *Customer Behaviour* • *Bill Data Usage*

Initially data exploration amongst the three main influential variables was performed to determine correlation amongst variables and generate a deeper understanding amongst the task in entirety. A series of comparative analysis was performed through several bar charts to derive specific locations, sales channels, handset retailers, bill data usage amounts. Predictive modelling was then performed by categorising sections into data cleaning process, variables selection process, fitting variables in predictive model, evalusating the model, and preferable model. There are five predictive models that were considered: Logistic regression, Decision Tree Classifier, Gradient Boosting Classifier, K-Nearest-Neighbors and Random Forest Classifier.

Component	***Key Findings***
Data Exploration and Predictive Modelling	*Demographics* • *Great Lakes, South, Mid Atlantic, and the top two states were TX and CA* *Customer Behaviour* • *Sales Channel: Branded Third Party Retail, Indirect, and Retail* • *Handset: Samsung, LG, and Apple* *Bill Data Usage* • *Lower Bill Data Usage*
Recommendations	*Locations* • *Location-based Marketing* *Sales Channels* • *Employee Training* *Handsets* • *Advertising* • *Include More Cross-selling Options*

Table of Contents

About EuroCom

A telecommunication company typically offers local and long-distance communication services, varying network services and cellular/mobile communications. Telecommunications is split into three categories: wireless, wireline and other services. (Turk and Montes, 1995)

EuroCom is an American based telecommunications company that specifically focuses on wireless services, offering mobile phone communication services to its customers. Customers have the selection of the following:

- Different mobile phone brands to purchase, depending on that, there will be battery/storage options for each device
- Network and data inclusions (i.e. 4G/5G networks)
- Length of time the account will be active
- Call and messaging inclusions

The current upselling services that are offered include upgrading options for:

- Internation call inclusions
- International roaming services

Whereas the current cross-selling services offered by EuroCom includes:

- Warranty inclusions
- Other additional support

EuroCom has effectively communicated their priorities, where they aim to identify amongst their new customers who are likely to be upsold and cross-sold with the current services they are offering. This will be achieved by EuroCom working alongside our consultancy firm called DataSoc, where we will successfully understand their customer's needs and be able to target the customers who can be upsold and cross-sold.

Problem Definitions

The goal of EuroCom, a telecommunications firm, is to consistently increase sales from its current customers. EuroCom faces a challenge in transforming a mass of existing customer data into useful insights. To help them enhance upselling and cross-selling of their services, DataSo consulting has identified three key business challenges that may be addressed through data exploration and predictive modelling.

- **Key Issue #1:** Identifying EuroCom's Customer Demographics
- **Key Issue #2:** Identifying EuroCom's Customer's behaviors and preferences influence the purchase
- **Key Issue #3:** What influences the likelihood of an individual to choose up-sold or x-sold?

Data Exploration and Research

Issue 1: Demographics

Are customers more likely to be upsold or x-sold depending on their location?

For the first issue, we prepared to explore the correlation between customer location and the success rates of being upsold or cross-sold. Through research, we hypothesised whether certain locations of greater population density were more likely to be upsold/cross-sold, or vice versa, with lower density population regions or states.

Locations

We first breakdown the question by identifying where the customers are predominantly located, more specifically the region and the state. Our results showcase that the top three regions that are upsold/cross-sold is the following: Great Lakes, South, Mid Atlantic, and the top two states were TX and CA.

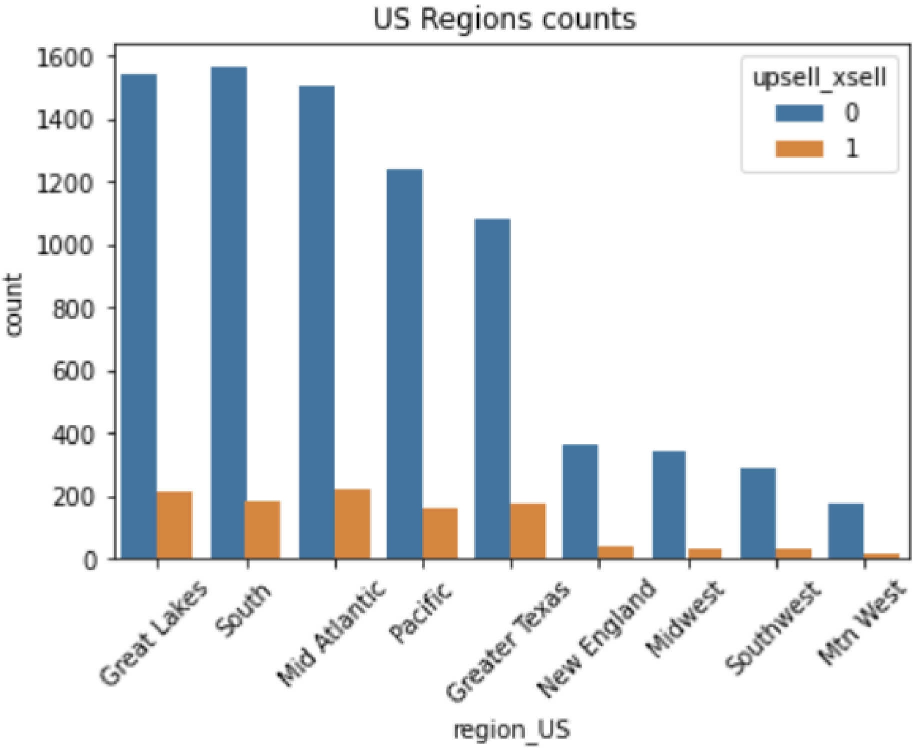

Figure 1. Number customers who chose upselling/cross-selling by Region in US

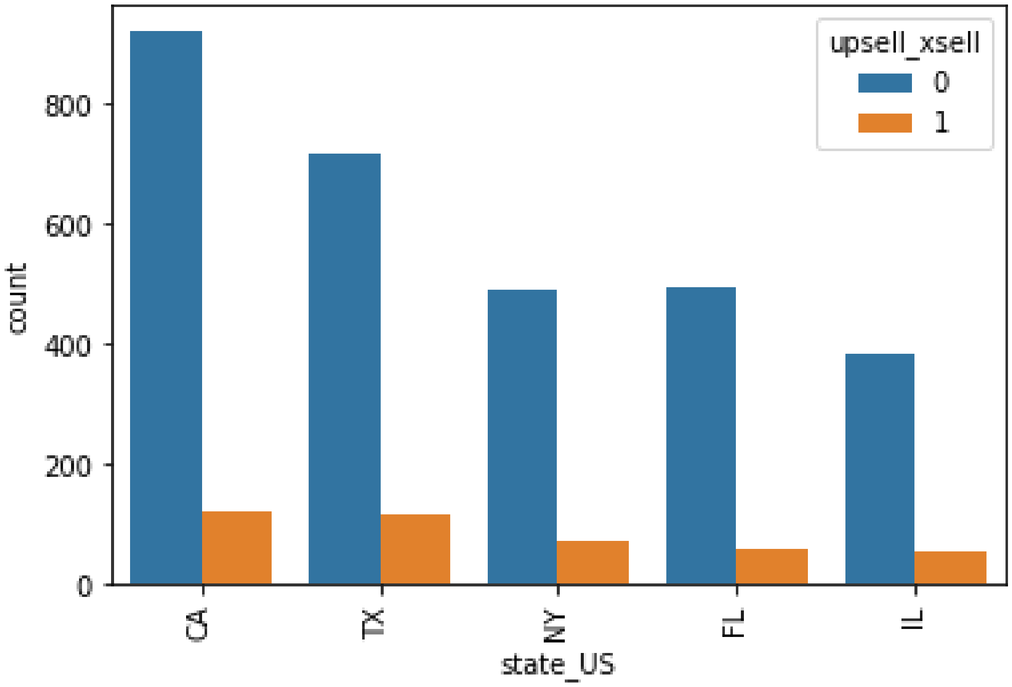

Figure 2. Top 5 States in US has the high number of customer who chose upselling/cross-selling

The double bar chart displayed in Figure 1 demonstrates the breakdown of each individual region and the count when upsold/cross-sold is equal to 0 (were not upsold/cross-sold) and to 1 (were upsold/cross-sold). Despite the overwhelming number of customers not being upsold/cross-sold, the regions Great Lakes, South and Mid Atlantic still display as the top three locations to have upselling/cross-selling occur.

To further assess the location of where customers were being upsold/cross-sold the most, we decided to investigate the specific states and found that the highest amount of upsold/cross-sold customers were in California (CA), which is in the Pacific region of the United States. This could indicate that other states in the Pacific were predominantly low in upsell_xsell. Despite this, we discovered the next highest upsell_xsell state was Texas (TX), which is in the South – one of the second highest regions with upsold/cross-sold customers.

Issue 2: Customer Behaviour

Does customer preferences influence their purchasing behaviour (to be upsold or x-sold)?

With the second section, we intended to investigate the impact of consumer characteristics on the success rates of being upsold or cross-sold. The purchase characteristics, including sales channel and handset time, allow us to identify the kind of buyers who are more likely to be upsold.

Sales Channel

The Bar chart in figure 3 below depicts the total count of customers depending on their sales channel. In this finding, the percentage of overall upsold for each category are:

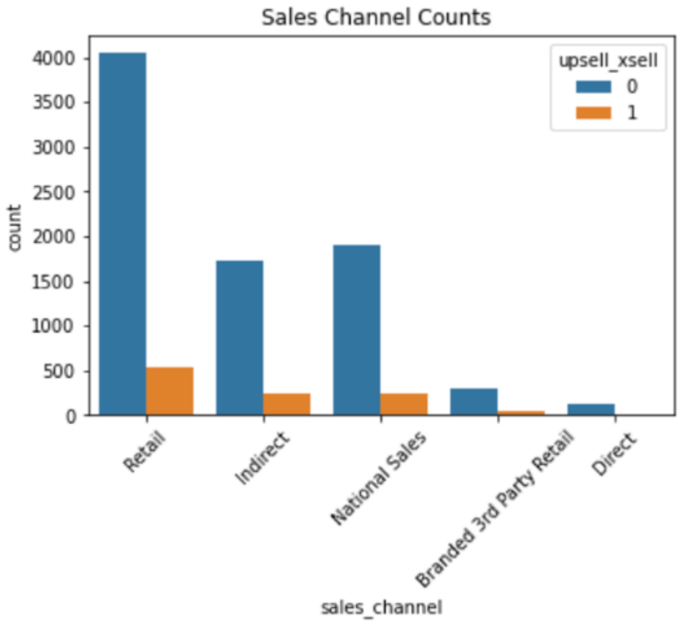

Figure 3: Sales channel Counts

Sales channel	**Upsold percentage**
Branded third party retail	13.89%
Indirect	12.02%
Retail	11.57%
Direct	9.70%
National sales	7.83%

Figure 4: Upsold percentage in Sales Channel

In a descending order of the most successful upsold, coming first is retail, although they also have the most successful upsells, the volume is somewhat significantly larger that indirect sales channels. Retail sales channel allow customers to build deeper relationships with customers by suggesting premium upsells that eventually delivers more revenue as well as making the customer feels like they had the better deal through their relationship with the retailers (Mehta & Balakumar, 2021). Furthermore, results reveal that the upselling in the company's indirect channel are positively and significantly impacted by supplying after-sales services through the direct channel. For national sales, business have more control over product distribution when dealing with single channel, they can create strict sales transaction to create a 'luxurious' market, therefore, creating less chance for customers to upsell or cross-sell. Whereas direct sales channel has lower volume in terms of upsell due to its low sales revenue.

Handset

The bar chart represented in the figure 5 illuminates the total_number of customers based on the number of handsets bought within the six main retailers. Figure 6 denotes the upsold percentage for each retailer.

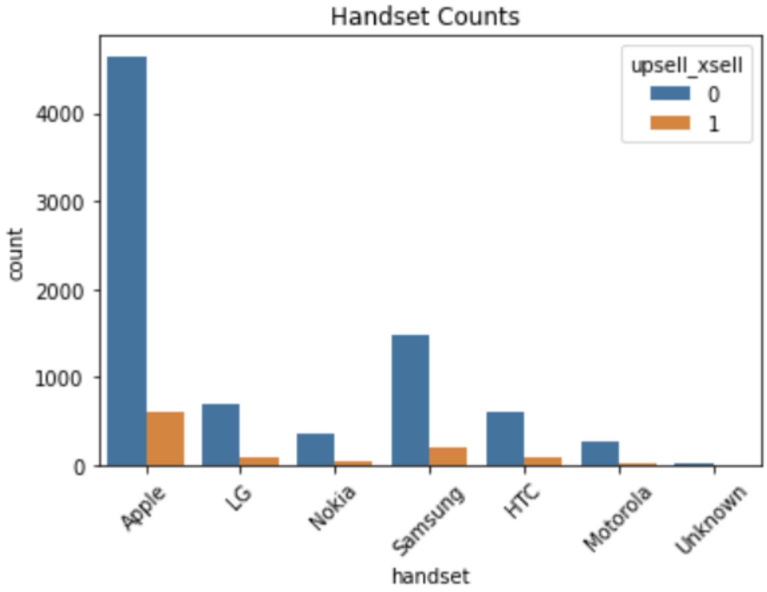

Figure 5: Handset counts

Retailer	Upsold Percentage
Apple	11.48%
Samsung	12.11%
LG	12.20%
HTC	13.10%
Nokia	10.90%
Motorola	9.74%
Unknown	25%

Figure 6: Upsold percentage in Handsets

Through comparative analysis, HTC's upsold percentage is the highest; disregarding the umbrella term 'Unknown'. This category holds the highest percentage; however, this includes several smaller companies with a significantly lower customer count. Hence, this portion of the dataset becomes invalid and unreliable. Through the same quantitative observations, LG and Samsung have the highest upsold percentage. Denoting that HTC, Samsung, and LG customers have the highest chance to upsell and/or cross-sell. A contradicting quantitative observation in relation to the volume of customers skews the qualitative conclusion. Samsung's customers are 0.63% more likely than Apple's customers to upsell/cross-sell; however, there are 193.17% more upsell customers in Apple than Samsung. This percentage increases drastically (HTC:Apple = 567.78% and LG:Apple = 526.42%). It's apparent that even though Samung, LG, and HTC have a higher upsell/cross-sell percentage; Apple, predominantly holds the most upsell customers. This is due to the significantly higher mode.

Issue 3: Bill Data Usage

Are customers who have a higher data/bill usage more likely to be upsold or x-sold?

Upon research on customer's amount of data usage, there was initial discussion of if higher data bill usage led to higher chances of customers being upsold/cross-sold, however instead we decided to identify the how much data was used by customers who were upsold/cross-sold the most to gain a better understanding of their purchasing behaviours.

To justify the reasoning why bill_data_usg_m03 was picked over data_usg_amt, was to observe the customer's behaviour over a smaller period rather than risk using outliers within the longer timeframe in data_usg_amt.

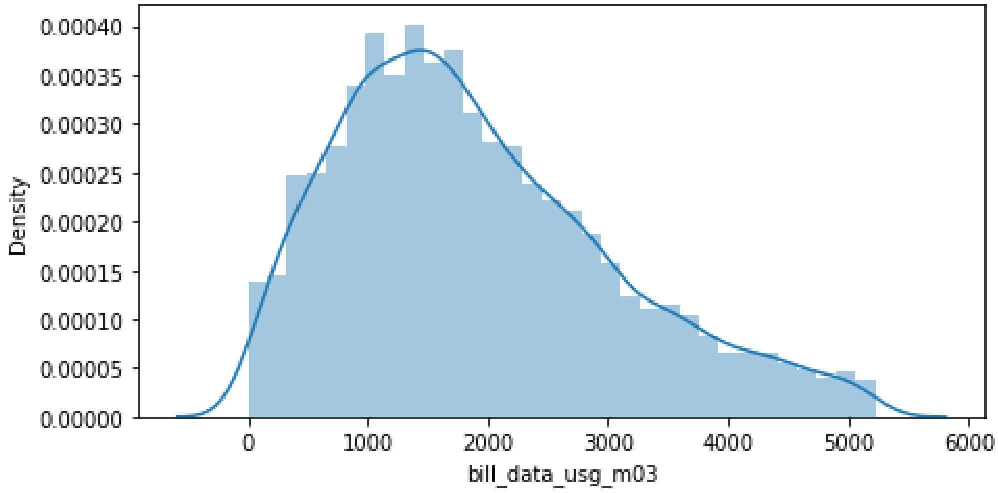

Figure 7: Density distribution in Bill data usage.

The figure above displays that the most billed data usage over the 3 months is within the range of 1000 to approximately 2000, which is not the highest billed data usage and this bar line chart indicates that it is negatively skewed. This gives an approximate range of where the amount of data usage lies.

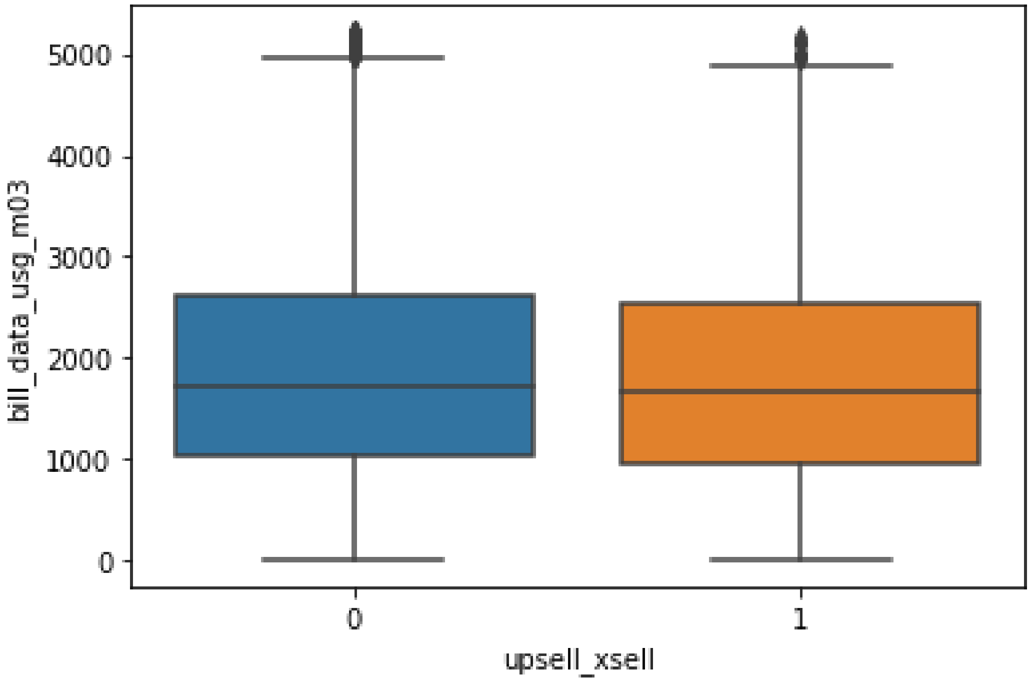

Figure 8: Boxplot of bill data usage.

In Figure 8, to gain a better understanding of the relationship between bill_data_usg_m03 and upsell_xsell we plotted a boxplot, which showed a small difference between the upsell_xsell = 0 and upsell_xsell = 1 for the data usage over 3 months. Through research, the billed data usage can assist in capturing the customers patterns regarding how they use data connectivity, alongside observation of customer case history and their video consumption. According to our research, the benchmark for the average amount of data used by an American consumer is 514GB per month (Baumgartner 2022), which is significantly more than the average of billed data usage of those who are being upsold/cross-sold which is approximately 1851MB.

Predictive Modeling

Data Cleaning Process

Before progressing to predictive modelling, many processes are performed to clean the data that was provided to make it more efficient for the machine. To minimise data bias, rows with missing values are deleted. Outliers are also eliminated using the interquartile range and outliers (IQR) techniques. Any negative values in columns containing negative values (bill data usg03 and data usage) were also deleted. Since the data frame provided was significantly unbalanced with 6081:799, Synthetic Minority Oversampling technique (SMOTE) was implemented to resample the dataset structure, yielding 6081 for upsell = 0, and 6081 for upsell = 1.

Variables Selection Process

Through utilising Recursive Feature Elimination (RFE), the feature selection algorithm configures features (x variables) in training datasets and decides the most relevant variables to predict our upsell_xsell (y variable). In the process, X variables were standardised in order to reduce the variables to a single scale without exaggerating the variations in their value range. Using 70% of our train and 30% of our test data through RFE, the predictive model concludes that the best combination of variables is incorporating 2 or 10 variables together giving the accuracy of 93.8%. Furthermore, to investigate which variables RFE have chosen, we utilised the make_classification function and in which resulting bill_data_usg_m03, and cs_tt_hhlds were the most relevant variables.

Fitting Variables in Predictive Model

The five models that were considered for the predictive model are Logistic regression, Decision Tree Classifier, Gradient Boosting Classifier, K-Nearest-Neighbors and Random Forest Classifier. The table below shows the result of fitting bill_data_usg_m03 and cs_tt_hhlds into the respective predictive model ranked by model accuracy.

Predictive Model	ROC AUC		Precision	Recall	f1-score
Random Forest Classifier	0.9775	0	1.00	0.96	0.98
Accuracy = 97.73%		1	0.96	1.00	0.98
Decision Tree Classifier	0.9209	0	1.00	0.85	0.91

Accuracy = 92.30%		1	0.86	1.00	0.92
K-Nearest Neighbours Accuracy = 77.28%	0.7786	0	0.91	0.62	0.74
		1	0.70	0.94	0.80
Gradient Boosting Classifier Accuracy = 56.81%	0.5675	0	0.58	0.54	0.56
		1	0.56	0.60	0.58
Logistic regression Accuracy = 52.48%	0.5072	0	0.52	0.40	0.45
		1	0.50	0.61	0.55

Figure 9: Fitting bill_data_usg_m03 and cs_ttl_hllds in five different predictive models

Evaluating the Model

An excellent Receiver operating characteristic (ROC) curve model is one in which the threshold score indicates that the true positive rate (TPR) is approaching one while the false positive rate (FPR) is as low as feasible. Evidently, The Random Forest classifier has the largest Area Under the Curve (AUC) of all five prediction models, indicating that the model outperforms the others in differentiating between positive and negative classifications.

Preferable Model

Random Forest Classifier is preferred the most out of the five models. Its ability to employ aggregating to increase predicted performance and minimise over-fitting by training several decision tree classifiers on different sub-samples of the data, enable it to excel in ROC AUC, prevision, Recall and F1-score. Additionally, compared to the second-best model, Decision Tree Classifier, Random Forest classifier does not only rely on the feature importance that is taken from a single decision tree. Furthermore, Random Forest classifiers can generalise the data in a more effective and efficient manner, thus making its randomised feature selection demonstrate higher accuracy compared to Decision Tree Classifiers.

Evaluation Research Questions with Predictive Model

Thus, to further evaluate our research questions, we fit the chosen variables with bill data usage to see how they fit in one another relating to the bill data usage. Through utilising our chosen model of Random Forest Classifier, all predictive models of top three States, Sales Channel and Handsets are in the range of high eighty accuracy, and ROC AUC.

States

Fitted with bill data usage	*ROC AUC*		*Precision*	*Recall*	*F1-score*
Texas	*0.8658*	*0*	*0.95*	*0.77*	*0.85*
		1	*0.80*	*0.96*	*0.87*
California	*0.8689*	*0*	*0.96*	*0.77*	*0.85*
		1	*0.80*	*0.97*	*0.88*
New York	*0.8629*	*0*	*0.96*	*0.76*	*0.85*
		1	*0.80*	*0.96*	*0.87*

Sales Channel

Fitted with bill data usage	*ROC AUC*		*Precision*	*Recall*	*F1-score*
Branded third party retailer	*0.8515*	*0*	*0.95*	*0.74*	*0.83*
		1	*0.78*	*0.96*	*0.86*
Indirect	*0.8831*	*0*	*0.97*	*0.79*	*0.87*
		1	*0.82*	*0.9*	*0.89*

Retail	0.8831	0	0.97	0.79	0.87
		1	0.82	0.98	0.89

Handsets

Fitted with bill data usage	ROC AUC		Precision	Recall	F1-score
Apple	0.8878	0	0.98	0.79	0.88
		1	0.82	0.98	0.89
Samsung	0.8777	0	0.97	0.78	0.86
		1	0.81	0.97	0.89
LG	0.8622	0	0.96	0.76	0.85
		1	0.79	0.97	0.87

Recommendations

Locations

Customers nowadays are particularly attentive to the advertisement they receive from businesses; with personalisation, Eurocom can address the demands of each customer while also creating a stronger relationship with them. In our explorative and predictive model, states with dense populations, such as California, Texas, and New York, demonstrate a substantial association between bill data consumption and the likelihood of being upsold. These states are among the most populous in the United States, yet they are also among the most sophisticated and developed due to their strong economic footing. In an ever-changing economic climate, consumers in these states are more likely to use more bill data to satisfy their everyday demands. As a result, Eurocom should employ location-based marketing to launch additional franchises to satisfy the demand of consumers in those states, particularly in the Great Lakes, South, and Mid-Atlantic regions. The higher the population, the more feasible it is for an economy to expand, and therefore the greater the need for modern telecommunication networks in Eurocom's favour.

Sales Channels

In our research section, we explore the possibility of different types of sales channels that might have effects on the consumers behaviour when are being upsold or cross sold. It is concluded the types of sales channels including Third Branded Retailers, Indirectly and Retails has a higher influence on the customers to be upsold. The three sales channels share a common characteristic of being able to persuade customers in person enabling the employees to build connections on the site. Thus, we recommend that Eurocom should implement trainings for employees to identify the types of customers to cater them through their preferences. A more intensive training program will enable Eurocom to strengthen employee's skill bridging them to a higher knowledge to generates more sales for Eurocom. Furthermore, customer experience is also a crucial element increasing the percentage of consumers to be upsold or cross-sold. Eurocom should initiate more offers more incentives for consumers such as testing out quality of a product or having coupons and discounts. This will enable consumers to be more actively involved in Eurocom and more likely will become a loyal customer.

Handsets

Our data analysis has identified that the top mobile handsets that are purchased by customers are Apple, LG and Samsung. Additional research has shown that the reason for this is because these are well-known brands which enable customers to trust their devices more than others and are more likely to purchase their marketed devices. Also, these devices come with multiple upgrades like battery capacity and storage amount, where customers re-evaluate their initial decision and are most likely to purchase upgraded versions of the device they wished to purchase (i.e. their initial decision was to purchase an iPhone 14, but they decided to buy an iPhone 14 Pro). Customers are made to believe they are making the appropriate decision of buying an upgraded mobile handset, although brands like Apple are simply utilisng their marketing/selling techniques to upsell their customers. Alongside that, all devices have the option of adding data services like 4G/5G, and the data analysis has shown that most customers are inclined to purchase such services with their handset. For this reason, we recommend EuroCom to continue advertising their upgrade options for each device, and also enhancing the way they advertise certain data network services (i.e. unlimited 5G service for $30 a month). Additionally, EuroCom should include more cross-selling options for the outlined brands, like Bluetooth devices that customers may add onto their purchase. This is because as these brands already have the highest purchases, they are more inclined to add a Bluetooth device to their purchase. Whereas other brands in the dataset like Nokia does not support Bluetooth connectivity so such services do not need to be added to those devices.

Appendix

It is assumed that removing missing values using the *dropna()* function the datasets shows 8.38% null value in the given data. It is essential that data wrangling happens prior to applying machine learning algorithm as it affects the performance and accuracy. Secondly, outliers are removed; it takes up 19.73% of the remaining data in case of factors such as measurement error, data entry error, processing errors as well as poor sampling. Lastly, negative values are also removed which took up 7.3% of the data. It is indicated that '1' is for upselling and '0' for no upselling, so a negative value in this data prediction is simply an error that must be discarded.

The limitations of the dataset for machine learning algorithms is that it is unbalanced, around 80% of the whole data indicates that there was no upsell or cross-sell being made, therefore it is exceedingly difficult to work with our limited data and the lack of variables that can be used to model the prediction.

References

Baumgartner, J, 2022, *Average data consumption eclipses half a terabyte per month - OpenVault | Light Reading* 2022, accessed 10 November 2022, < https://www.lightreading.com/cable-tech/average-data-consumption-eclipses-half-terabyte-per-month---openvault/d/d-id/775689#:~:text=Average%20data%20consumption%20eclipses%20half%20a%20terabyte%20per%20month%20%E2%80%93%20OpenVault&text=Consumer%20broadband%20consumption%20in%20North,a%20new%20study%20from%20OpenVault>

Mehta, R. & Balakumar, K., 2021. Redesigning after-sales service: Impact on incumbent product distribution channels. *Journal of Retailing and Consumer Services*, *58*, p.102279.

Turk, I. and Montes, S. (1995). *THE U.S. TELECOMMUNICATIONS SERVICES INDUSTRY Assessing Competitive Advantage*. [online] Available at: https://www.commerce.gov/sites/default/files/media/files/2018/the_u.s._telecommunications_services_industry_assessing_competitive_advantage.pdf