Table of Contents

Introduction

EuroCom is a telecommunication company that currently aiming at boosting sales of its services among customers, particularly focusing on up-selling (upgrade service) and cross-selling (purchase add-on service). This report first analyses the data captured by EuroCom, which contains various data metrics of its existing customers. Then, three prediction models were built based on the variables provided, followed by the explanation and evaluation for each model. Lastly, the most suitable model will be selected, and recommendations based on this model will be provided.

Data Exploration

Data Cleaning

To produce an accurate analysis of the data and better prediction models, we first used python to complete the data cleaning. The dataset has 10000 records in total, but there are also abnormal values such as null or negative values in between.

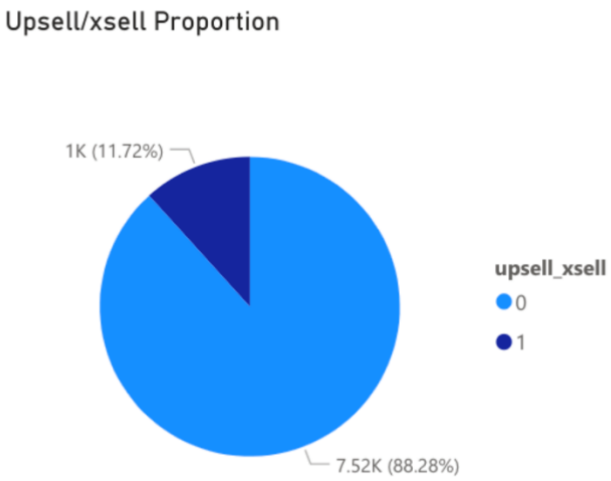
Firstly, we changed the dtype from object to float so that we can remove the negative values using the comparison operator. Then, all the records with negative values in 'data_device_age' and 'bill_data_usg_m03' were removed as they shouldn't be negative. The reason for removing them rather than changing them to positive is that the number of records with negative values is not high compared to the whole dataset so it might not greatly affect the result. Then we removed all the records with null values. After data cleaning, the dataset decreased to 8000+ records, then we conducted basic data analytics with power bi.

```
In [1]: #import dataset
        import pandas as pd
        df = pd.read_csv("Customer_Data.csv")

In [ ]: df.dtypes

In [ ]: df['data_device_age']=df['data_device_age'].str.replace(',','')
        df['data_device_age']

In [ ]: df['bill_data_usg_m03']=df['bill_data_usg_m03'].str.replace(',','')
        df['bill_data_usg_m03']

In [ ]: df['data_device_age'] = df['data_device_age'].astype(float)

In [ ]: df['bill_data_usg_m03'] = df['bill_data_usg_m03'].astype(float)

In [ ]: df.dtypes

In [ ]: Customer1 = df.drop(df.index[df['data_device_age'] < 0])
        Customer1

In [ ]: Customer2 = Customer1.drop(Customer1.index[Customer1['bill_data_usg_m03'] < 0])
        Customer2

In [13]: Customer_Cleaned= Customer2.dropna()
         Customer_Cleaned
         Customer_Cleaned.to_csv('./Customer_Cleaned.csv')
```

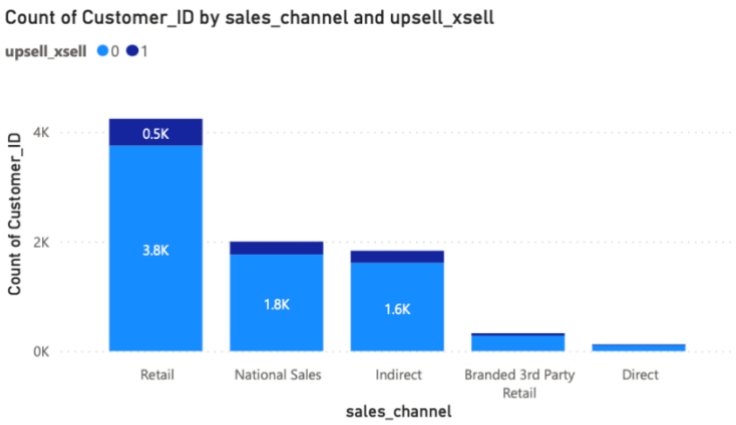
General Analysis

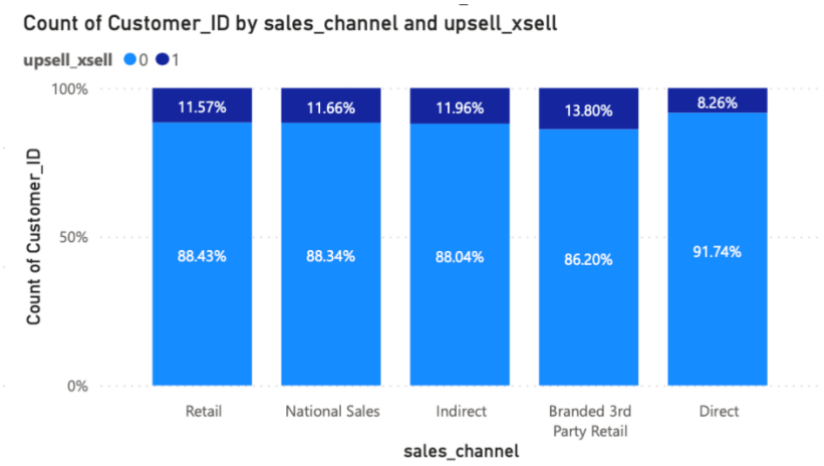
The dataset contains some basic data about the customers and the whether the customer bought/upgrade the service. We conducted general data analytics with these data and summarised some key information that might be useful.

Firstly, we explored the basic data regarding the customers including total sales and sales channels. The sales of EuroCom were relatively low, by looking at the total sales, only 11.72% of the customers have upgraded/purchased services.

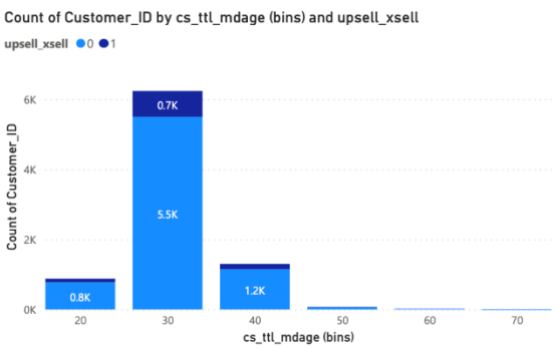


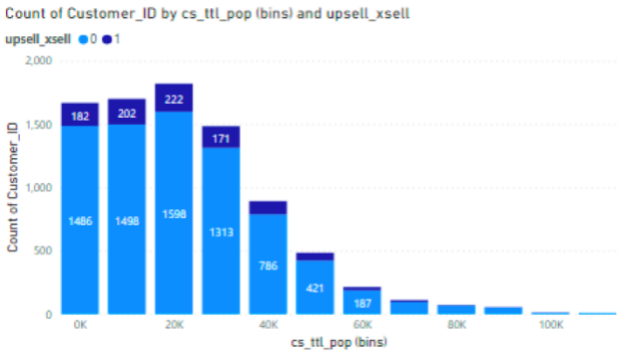
The major sales channel for the company is retail, followed by national sales, indirect, branded 3rd party retail and direct. But if looking at the ratio of sales in different channels, branded 3rd party retail shows a higher percentage of sales. This might be an opportunity for the company.

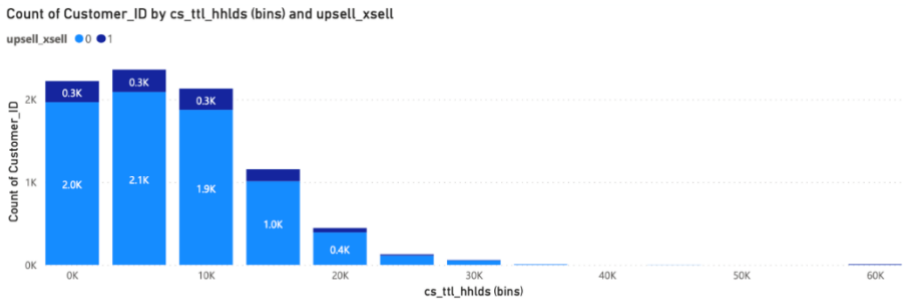


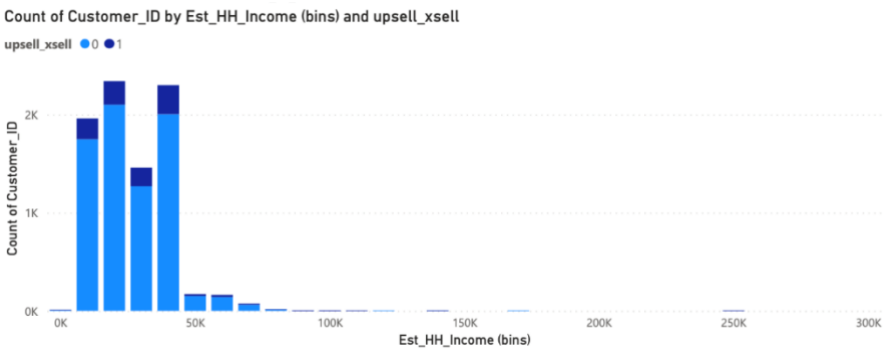


Secondly, we explored the neighborhood data of customers. The graphs below show that most of the customers are living in relatively young neighborhoods, with median ages around 20 – 40, and this group of customers also contributes to the most sales. The total population and households in the neighborhoods also show that most of the customers are not from huge neighborhoods. The estimated income of the customers also shows that most of the customers are with an estimated income of lower than 50k, which means most of the customers are from low-income households. Based on these graphs, we can assume that younger generations with lower income and living in small to middle size neighborhoods are the current customer group.

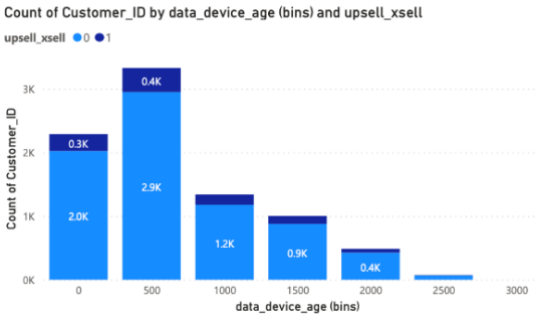


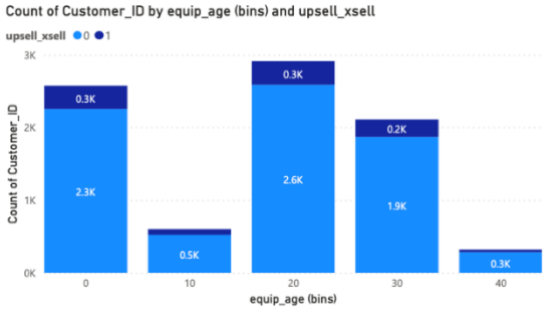




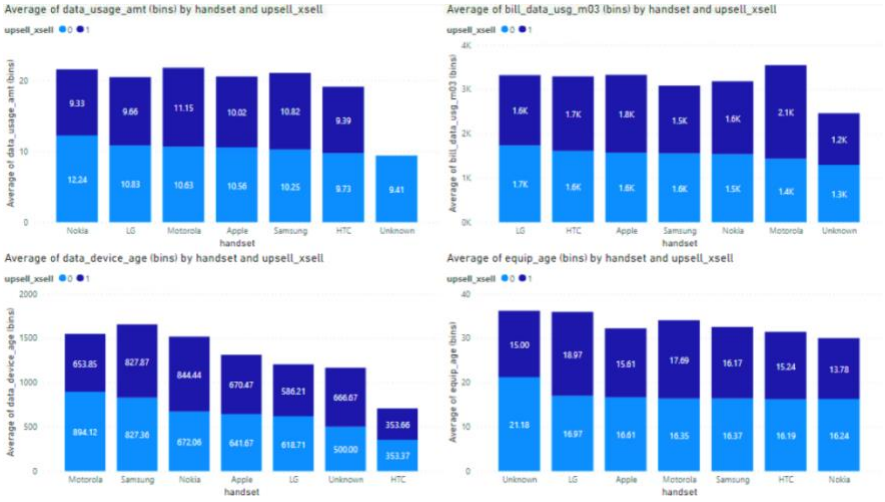


Lastly, we investigated the device on plan and handset age. Most customers have their devices on the plan for 0 – 500 and this group of customers is also more likely to purchase the services. The relationship between device age and sales shows that customers with new phones or customers who have own the phone for a relatively long time are more likely to purchase services.



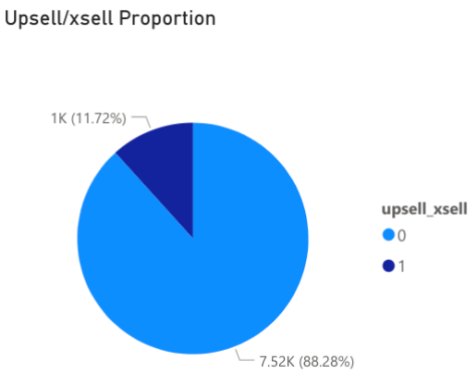


There are also some other data not mentioned above, as the result seem to show little or no difference across different groups. For instance, the gender percentage of the neighborhoods where the customer lives are relatively equal at around 50%, which doesn't seem to affect sales. Apart from this, the data usage, devices on plan, and device age are similar between customers who have or have not made purchases and customers using different handsets.



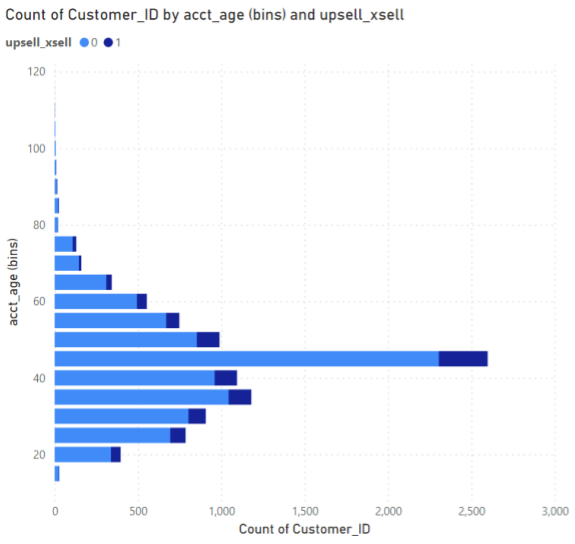
Business Issues

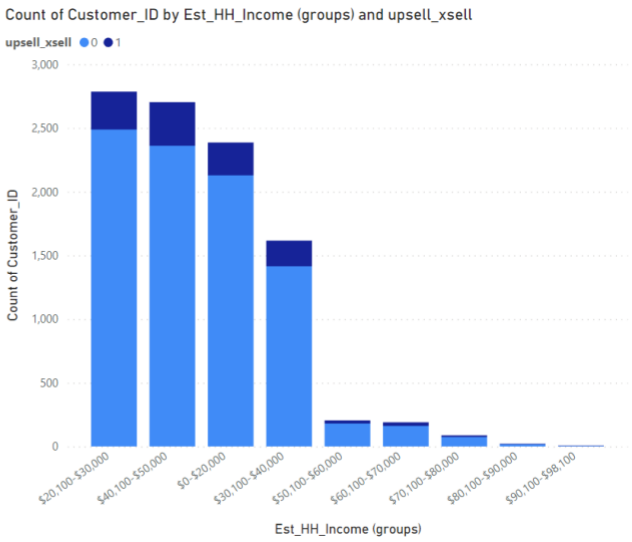
EuroCom currently has three main business issues that need to be solved urgently. The first commercial problem, and the most important one, is that very few customers choose to purchase up-selling (upgrade service) and cross-selling (purchase add on service). Out of the 10,000 customer data provided by the enterprise, only 1,174 customers chose to purchase the service, and the pie chart below shows the small percentage of these customers. This means that the business is not profitable and the customers are not interested in the current products or the marketing methods used by EuroCom. The enterprise needs to optimize its service or adopt more targeted marketing strategies to attract more customers to its services.

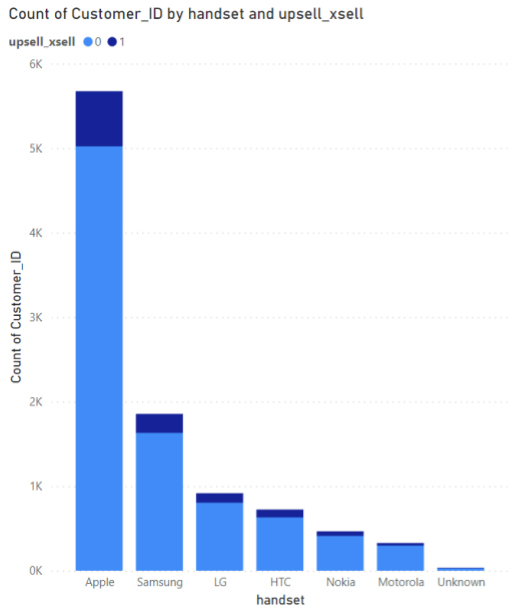


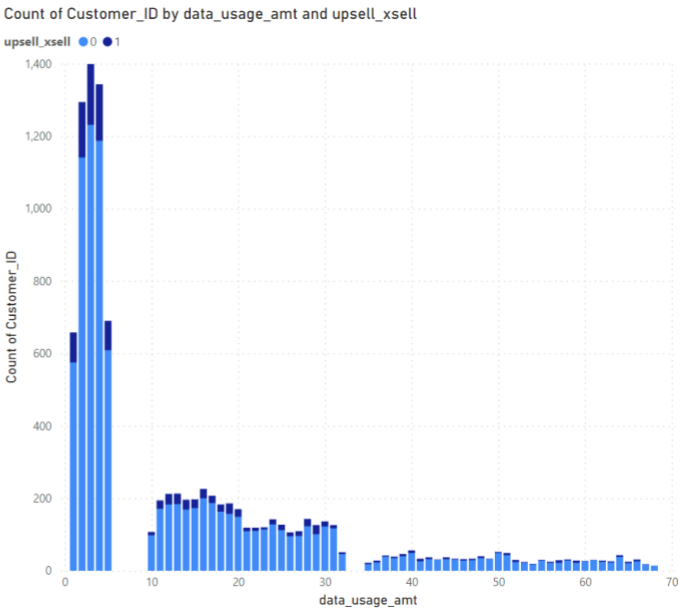
The second business issue is the ineffectiveness of the promotional strategies currently adopted by EuroCom. The problem can be divided into two fields: the lack of targeted marketing strategies for different users and the narrow sales channels at present. In order to see if EuroCom is currently using different marketing strategies (if more promotions are

being offered to groups with more customers), the following four bar charts were created to reflect the correlation between the number of upsells and the length of user account registration, estimated revenue, device brand and data usage amount. We believe that if a significantly high percentage of upsell is purchased by the group with a high number of customers, then the enterprise has adopted an effective marketing strategy, and vice versa, the strategy is poor.



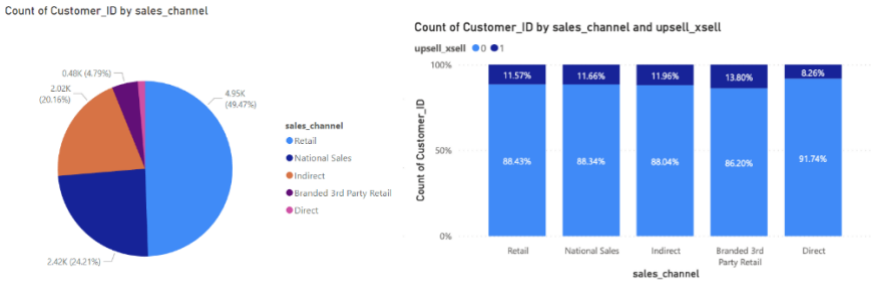






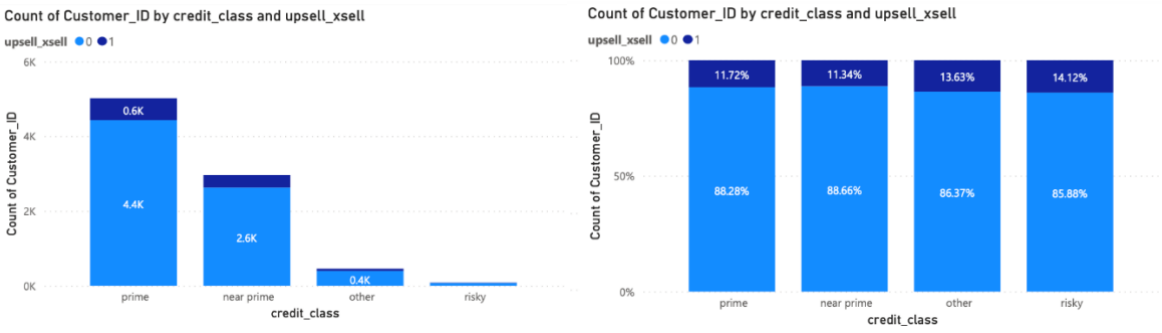
From the graphs above, it is clear that the proportion of customers buying services in the groups with a high number of customers does not significantly exceed that of the other groups, so we think that EuroCom is not using a targeted marketing strategy. The enterprise

should make more efforts to find its target customer groups and optimize marketing strategies for them.



The pie chart above shows the species sales channels currently used by EuroCom, and it can be seen that nearly half of the acquisition places are Retail. However, as can be seen from the bar chart on the right, Branded 3rd Party Retail sold a higher proportion of paid services. So, EuroCom should adjust its marketing strategy and focus more on other sales channels. In addition, the enterprise should also focus on broadening its sales channels. Online shopping is popular nowadays, so EuroCom can set up a website and offer its services online or use social media platforms like Twitter and Facebook.

The last business issue is about credit class classification in customer data – the classification lacks meaning. As we can see from the chart on the left below, EuroCom classifies customers into 4 credit tiers, and we speculate that credit classes are related to the probability of a customer purchasing – customers with a 'Prime' credit class are likely to have a higher probability of purchasing upsell services. However, the bar chart on the right shows that the proportion of purchases is similar across the different credit classes, so the association between credit and purchase likelihood is weak. In summary, we think that the meaning of this hierarchy is unclear and perhaps EuroCom needs a more meaningful user hierarchy based on additional information about the customers.



Logistic Regression Models

From the description of the data above, we already know the distribution of the data in the database and the basic information about each variable. Therefore, we decided to select some of the variables in the data and build a model. We wanted to build a reliable model that we could use to predict the company's customer buying behavior.

First, we decided to choose a logistic regression model as the model we would build. The reason for choosing logistic regression instead of linear regression is that our goal is to predict customer buying behavior, and this dichotomous problem has only 0 and 1 possibilities (buy or not buy). This fits the scenario of using logistic regression models. The response variable in the linear model, on the other hand, can be any number. Therefore, under this problem, we first choose the logistic regression model.

The second step is the selection of variables, we first chose the broadest one. We placed all numeric type variables in the model. The main types of variables we selected were the following:

```
Out[269]: Unnamed: 0            int64
          Customer_ID           int64
          upsell_xsell          int64
          acct_age              int64
          credit_class         object
          sales_channel        object
          rfm_score             int64
          Est_HH_Income       float64
          region_US            object
          state_US             object
          cs_ttl_pop          float64
          cs_ttl_male         float64
          cs_ttl_female       float64
          cs_ttl_hhlds        float64
          cs_ttl_mdage        float64
          handset              object
          data_device_age     float64
          equip_age             int64
          bill_data_usg_m03   float64
          data_usage_amt        int64
          dtype: object
```

From this we built the first model:

upsell_xsell =1/(1+exp(-(-2.05781306-0.05846268*acct_age-
0.19019962*rfm_score+0.06599982*Est_HH_Income-0.09389049*cs_ttl_pop-

0.01389292*cs_ttl_male+ 0.01389292*cs_ttl_female+0.11171283*cs_ttl_hhlds-0.01856212*cs_ttl_mdage+0.00891108*data_device_age-0.05608108*equip_age+0.15172219*bill_data_usg_m03-0.01184414*data_usage_amt)))

In this model, we found the accuracy rate is 0.8744. And when looking at the Confusion matrix we found 0 on both false negative and true negative.

```
 [[1490    0]
 [ 214    0]]

True Negative(TN) =  1490

True Positive(TP) =  0

False Positives(FP) =  0

False Negatives(FN) =  214
```

Before explaining why we will have this situation, we need to briefly define false negative and true negative (Cheung 2022):

True Positive: True Positives occur when we predict an observation belongs to a certain class and the observation actually belongs to that class.

False Positive: False Positives occur when we predict an observation belongs to a certain class but the observation actually does not belong to that class. This type of error is called Type I error.

This situation is caused by the fact that our prediction results are all 0 (customers do not buy our product. Does this mean that our model is wrong? We look at the distribution of the whole data and find that customers who do not buy are the majority of the overall customers

reaching 88.28%. So, it is understandable that the model built in this situation tends to give a result of 0. Similarly, this model yields an AUC of 0.500 which is also due to the specificity of the data, we will explain more at the end of this section.

Then we tried to build more models to see if we could simplify the variables and get the same or higher accuracy with fewer variables. Here we used Rstudio, in which we used stepwise regression, and by looking at AIC. We found the optimal few variables for which we could build a logistic regression:

upsell_xsell=1/(1+exp(-(-1.650-4.561*acct_age-2.413*rfm_score+
6.397*Est_HH_Income+7.514*bill_data_usg_m03)))

When we performed machine learning on the above model through python we found. Accuracy, Confusion matrix, and AUC are consistent with the first model. In other words, the second model allows us to obtain the same results as the previous model with fewer variables.

So far, we think the second model is better. Of course, all this is based on the case of using logistic regression to build the model. Since we also mentioned above that both False Negative and True Negative are 0, although this is acceptable, it also means that logistic regression may not achieve optimal choices for explaining our question, so we need to use other methods.

Decision Tree

In addition to logistic regression, we have built a decision tree model as an alternative approach to classify binary outcomes. Instead of creating equations, decision tree sets up rules with selected predictors to make predictions. We conduct a decision tree model and try to figure out if decision tree model can outperform logistic regression in this case.

For the selection of variables, we start by putting all the numerical variables into the model (figure 1), which are the same variables in our logistic regression model:

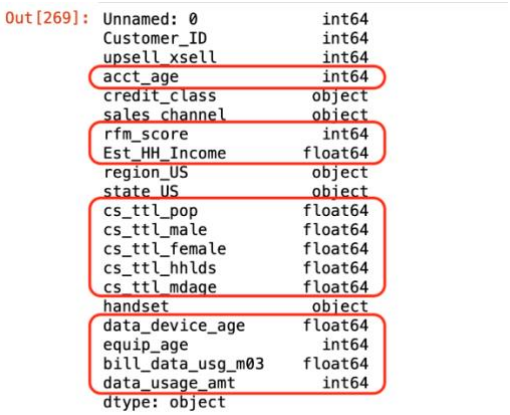


Figure 1. variable selection

With the input of these variables, the accuracy rate of the model is 0.7858, which is much lower than our logistic regression model, while the AUC is similar between the two models. As shown in Figure 2, the result of Confusion matrix indicates that the decision tree model still has limitation due to the distribution of the raw datasets, which has been explained in our logistic regression model. As over 80% of the customers in datasets do not buy upsell or cross-sell services, it is understandable that the results of the decision tree prediction are mostly 0.

However, due to different algorithms and logic from logistic regression, even with the limitation of raw datasets, the decision tree model can correctly predict customers who will buy the upsell or cross-sell product to a certain extent. In this case, the result has 29 correct predictions of customers who will buy the service, which is slightly better than logistic regression model.

```
Confusion matrix

 [[1310  180]
 [ 185   29]]

True Positives(TP) =  29

True Negatives(TN) =  1310

False Positives(FP) =  180

False Negatives(FN) =  185
```

Figure 2. Confusion matrix of the first Decision tree model

And therefore, we try to improve the accuracy of the decision tree model by applying different sets of variables. For the first trial, we removed some variable that has less influence on the model, which includes Handset Age, 3-month Avg Billed Data Usage, and Data Usage Amount. After removing these variables, we found that the true positive slightly increased from 29 to 32, and the AUC remained at the same level as the last model.

After reviewing the datasets dictionary and characteristics, we found that there are variables that have multicollinearity among each other, which are Neighbourhood Total Population, Neighbourhood Total Males, Neighbourhood Total Female, and Neighbourhood Total Households. The multicollinearity will affect the influence of each variable on the model, and it will eventually affect the accuracy of the model.

Thus, we removed the variables which have multicollinearity and keep only Neighbourhood Total Households among these 4 variables. And we got below Confusion matrix for the model. (Figure 3). With the new set of variables, the confusion matrix has been improved and the AUC of the second trial model has improved to 0.5247, which is the best result with the limitation of data distribution so far.

```
Confusion matrix

 [[1285  205]
 [ 174   40]]

True Positives(TP) =  40

True Negatives(TN) =  1285

False Positives(FP) =  205

False Negatives(FN) =  174
```

Figure 3. Confusion matrix of the final Decision tree model

Model optimization discussion

From the discussion of the two build model ways above, we find that even the optimal model so far only achieves 0.5247 on the AUC. which is obviously very low. Therefore, we will discuss some methods in this section and see if we can improve the AUC of the model.

1st Method - Stratified sampling

We randomly select 500 records with 'upsell_xsell' of 0, and 500 records with 'upsell_xsell of 1. This method solves the problem of too many 0's affecting the model as mentioned above. However, the disadvantage of this is that the accuracy of the model decreases due to the small amount of data available for training.

We rebuilt the logistic regression model using the data frame obtained by the above method. The result is that the AUC increases to 0.5424, which is an improvement from the previous 0.500, but not significant. However, the accuracy drops significantly to 0.5400, which is significantly different from the previous 0.8744. Therefore, we believe that Stratified sampling is not applicable to this case, so we made a second attempt.

2nd Method - Category ordinal encoder

In this method, we used the ordinal encoder for two category variables.

These two new variables are then put into the optimal models of decision tree and logistic regression. The following is a comparison of the variables before and after they are added.

credit_class	Encoders credit_class
prime	1
near prime	2
risky	3
other	4

handset	Encoders handset
apple	1
Samsung	2
HTC	3
LG	4
Motorola	5
Nokia	6
Unknow	7

Optimal Model Build by Decision Tree		
	Before add encoder variable	After add encoder variable

Accuracy	0.7776	0.7829
Precision	0.1633	0.1609
Recall	0.1869	0.1729
F1	0.1743	0.1667
AUC	0.5247	0.5217

Optimal Model Build by Logistic Regression		
	Before add encoder variable	After add encoder variable
Accuracy	0.8744	0.8744
Precision	0.0	0.0
Recall	0.0	0.0
F1	0.0	0.0
AUC	0.5000	0.5000

It can be seen that adding the category encoder variables does not increase the reliability of logistic regression model in this case. On the other hand, adding category encoder variables decreases some values in decision tree model.

Although both attempts did not make our model better, this process provides us with ideas for future research. In the future, we can not only change the variables of the model to increase the reliability but also make changes to the data.

Recommendations and conclusion

From the optimal logistic model we selected above, we can find 40 potential customers within the existing customer base. However, this number is obviously not enough and based on the current data, we cannot find more upsell/xsell customers for EuroCom by building a model. One of the most important issues is the positioning of the product/service offered by EuroCom. Based on the above data exploration and models, we did not find any variables that have a significant impact on the accuracy of the models. In addition, EuroCom also needs to deal with the quality of customer data. This is probably because many people try to create accounts and then quickly choose to abandon them. If we build models by using these data together with other customers' data, it will greatly reduce the accuracy of the forecast. Our suggestion is to use EuroCom's back-end data to distinguish and remove all the low quality, useless data, which will provide predictive accuracy in the next phase of analysis.

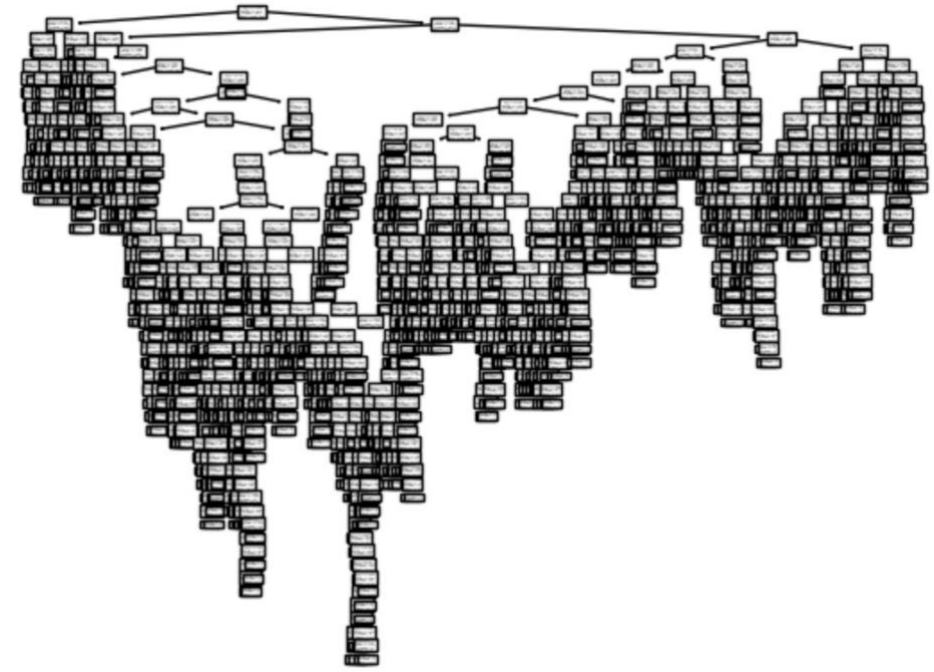
On the other hand, Eurocom's existing marketing strategy is ineffective, for example, there is no significant difference in sales to different credit classes as mentioned in the previous section, and in this case, Eurocom clearly needs to target its products more closely to the groups with the most purchasing power. Besides, Eurocom's official website is not visually or content-wise attractive to new customers because of its cluttered homepage and the overwhelming amount of textual information. （see appendix）We strongly recommend that the website be adjusted accordingly.

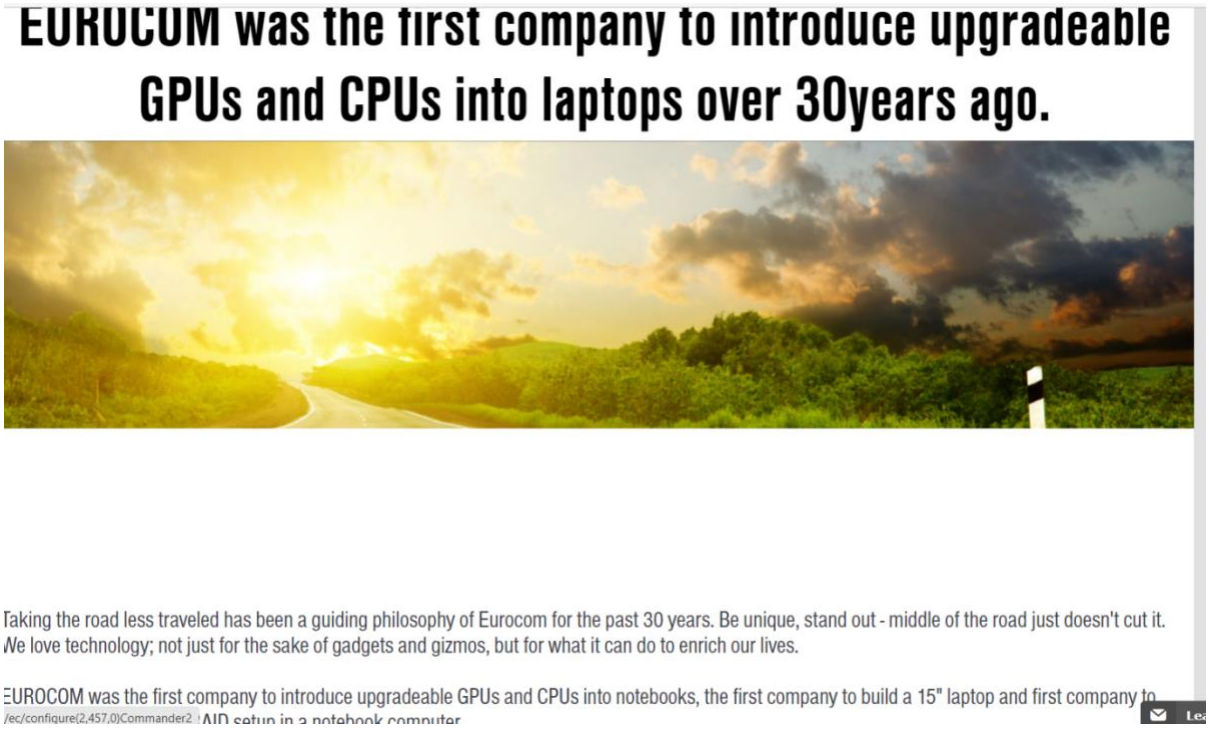
What's more, the current sales channel for Eurocom's products is mainly through retail, but only 11.72% of the customers have upgraded/purchased services. party retail had the highest sales success rate of 13.8%. Of course, due to the sample size, we are not sure that the success rate will be significantly improved by increasing the percentage but based on the current poor situation, it is definitely an approach worth trying.

The issues Eurocom facing are serious and need to be addressed urgently. However, we believe that with the new marketing and sales strategy in place, Eurocom can collect a new round of high-quality user data with more relative variables. At that point, building new models to forecast will have a greater chance of finding more potential customers.

Appendix

Decision tree outcome





Reference

Cheung, K.F. (2022), ed: Confused about confusion matrix, course material, programming for Data Analytics INFS5715, University of New South Wales. Available online at: https://edstem.org/au/courses/9841/lessons/26292/slides/202886.