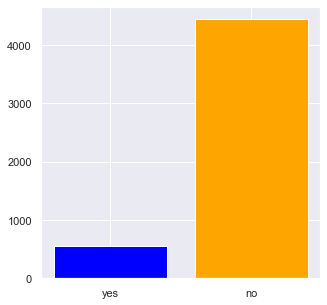
**FOUNDATIONAL BUSINESS ANALYTICS**

**COURSEWORK**

**SECTION-A: SUMMARIZATION**

The dataset provided contains 5000 records of customers contacted before for marketing of a similar financial product launched by another banking enterprise. The dataset includes information about the demographics of the contacted customers like their age, marital status, education, employment, credit information and loans disbursed in addition to previous contact details like the mode of communication, duration of calls, frequency of contacting before the campaign and their response to the same. The target class for the dataset (y) is a categorical variable which records if a sale was made to the customer successfully.

(For comparison between the target class responses for each feature, refer to the graphs attached in the appendix.)

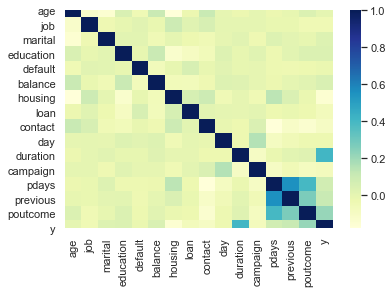
This sample (N=5000) has recorded 559 successful sales to customers, most of them being married professionals either working as managers or technicians and who on an average had completed their secondary education. With a mean bank balance of 1793 pounds (s=3245.47) and no sanctioned loans on an average, the customers who responded positively had been contacted about 64 days (s=119.72) after last contact with an average call duration of 544 minutes (s=3766.6).

The marketing campaign failed to target 4451 customers out of 5000. This set of customers were mostly married professionals with blue-collar jobs with a mean bank balance of 1281 pounds (s=3155.56). On an average, they had no defaulting credit or personal loans disbursed to them. But in contrast to the customers who responded positively, most of them had a house loan sanctioned to them and had been contacted about 38 days (s=101.29) after last contact with an average call duration of 219 minutes (s=202.13).

The dataset summary statistics is recorded below:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Age** | | **Balance** | | **Day** | | **Duration** | | **Campaign** | | **Pdays** | | **Previous** | |
| **Result** | **No** | **Yes** | **No** | **Yes** | **No** | **Yes** | **No** | **Yes** | **No** | **Yes** | **No** | **Yes** | **No** | **Yes** |
| Count | 4441 | 559 | 4441 | 559 | 4441 | 559 | 4441 | 559 | 4441 | 559 | 4441 | 559 | 4441 | 559 |
| Mean | 41 | 41 | 1281 | 1793 | 16 | 15 | 219 | 544 | 3 | 2 | 38 | 64 | 0.5 | 1.1 |
| Std | 10 | 13 | 3156 | 3246 | 8.3 | 8.7 | 202 | 377 | 3 | 2 | 101 | 114 | 1.7 | 2.6 |
| Min | 19 | 18 | -8019 | -1042 | 1 | 1 | 1 | 49 | 1 | 1 | -1 | -1 | 0 | 0 |
| 25% | 33 | 31 | 53 | 195 | 8 | 7 | 94 | 255 | 1 | 1 | -1 | -1 | 0 | 0 |
| 50% | 39 | 38 | 405 | 706 | 16 | 15 | 162 | 423 | 2 | 2 | -1 | -1 | 0 | 0 |
| 75% | 48 | 50 | 1289 | 2169 | 21 | 22 | 277 | 750 | 3 | 2 | -1 | 96 | 0 | 1 |
| Max | 83 | 92 | 71188 | 45248 | 31 | 31 | 2053 | 2231 | 51 | 21 | 850 | 784 | 30 | 29 |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Job** | | **Marital** | | **Education** | | **Default** | | **Housing** | | **Loan** | | **Contact** | |
| **Result** | **No** | **Yes** | **No** | **Yes** | **No** | **Yes** | **No** | **Yes** | **No** | **Yes** | **No** | **Yes** | **No** | **Yes** |
| Count | 4441 | 559 | 4441 | 559 | 4441 | 559 | 4441 | 559 | 4441 | 559 | 4441 | 559 | 4441 | 559 |
| Unique | 12 | 12 | 3 | 3 | 4 | 4 | 2 | 2 | 2 | 2 | 2 | 2 | 3 | 3 |
| Top | blue-collar | mgmt | marr-ied | marr-ied | secon-dary | secon-dary | no | no | yes | no | no | no | cell | cell |
| Freq | 974 | 133 | 2714 | 290 | 2325 | 271 | 4349 | 552 | 2604 | 358 | 3669 | 508 | 2758 | 454 |



Using a heatmap, the Pearson correlation between the pairs of input features and of each feature with respect to the outcome can be visualized. On studying the heatmap generated for the same, it can be successfully established that most pairs of input features are independent of each other. There is a moderate positive correlation between only one pair of features (Pdays-Previous (r=0.554)).

It can also be observed that the outcome variable (y) has its strongest correlation with duration (r=0.408, which still however falls under the weak correlation umbrella).

From a business point of view, this particular correlation would establish a relationship between the outcome of the sales pitch during the previous campaign and the prior number of contacts to the customer. Every other input feature pair mostly has extremely weak correlation. The absence of a stronger correlation between Poutcome and the other features can probably be attributed to the majority of unknown values dominating the dataset for this particular input feature. After observing the heatmap, it can be concluded that the outcome variable shares an extremely weak correlation with every other input feature.

In summary, it can be observed that a vast majority of unknown values dominate the dataset for the Poutcome feature. Customers who respond positively to the campaign have a significantly longer call duration on an average. Additionally, they also have a higher bank balance on an average and have been contacted more as compared to the set of people who responded negatively to the campaign. The most significant observation is that a majority of the customers who responded positively to the campaign do not have a housing loan whereas the customers that the campaign failed to target are more likely to have a housing loan.

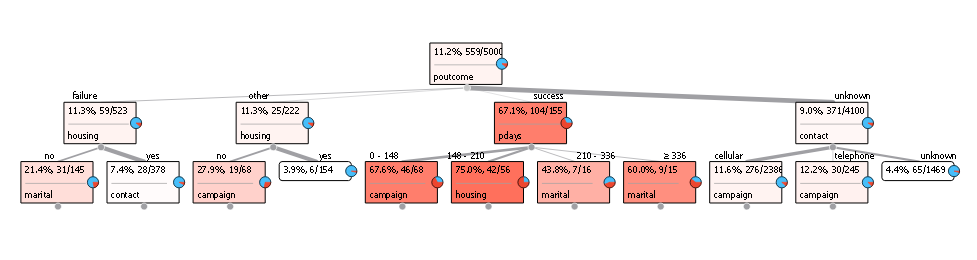
**SECTION-B: EXPLORATION**

The tree constructed using all the features ends up in leaf nodes that in turn, overfit the data because of the irrelevance of some of the input features. To tackle this curse of dimensionality, backward selection of features proved most effective. A tree with the full feature set was first produced and then the base accuracy, precision and recall were noted. Then the impact of removing each feature individually was observed. The feature, the removal of which led to maximum improvement of the performance metrics, was then dropped. Multiple iterations of this process were then carried out to reach an optimum trade-off point between classification accuracy, precision and recall. Feature selection not only optimizes the performance of the model but also plays a crucial role in reducing the complexity of the tree structure. Gini decrease index is an efficient coefficient to rank features as per their relevance to the dataset. However, splitting solely on the basis of high Gini decrease affects the precision of the model. Therefore, information gain ratio is also taken into consideration while understanding the relevance of the input features with respect to the target class.

After the iterations, six features were selected based on the Gini decrease and information gain ratio value, keeping in mind that the performance of the system is improved. A notable observation is when a categorical feature like job is considered as an input feature, it helps in identifying more interpretable subpopulations. For instance, if a student is targeted, he is more likely to subscribe to the product in comparison to customers belonging to some other job category. Similarly, with the combination of poutcome and job, it can be observed that the house-maids, entrepreneurs, self-employed professionals and customers belonging to the ‘other’ job category for whom the campaign had previously failed, have not subscribed to the product. However, the inclusion of the job feature comes as a trade-off with the model performance. On using a decision tree to unpack and split across the dataset, it was observed that the primary split was done on the basis of poutcome. The rank widget on Orange assigns the highest Gini decrease value to poutcome i.e., splitting on this feature would reduce the impurity of the node most significantly.

Splitting based on poutcome produced 4 categories namely success, failure, unknown and other which were then split further on the basis of pdays, housing, contact and housing respectively to produce nodes with lesser impurity. In the subsequent splits, the marital, contact, housing, campaign and pdays features play an important role in generating the following nodes and leaves.

The high frequency of splits based on a feature implies that it plays a crucial role in identifying useful subpopulations in the dataset which is also evident from the target based statistical analysis in the previous section. For instance, most customers who responded positively to the campaign do not have a housing loan whereas the other set of customers are more likely to have a housing loan. This when combined with pdays (customers who respond positively on an average have been contacted after a longer time as compared to the rest) generates a node with considerably less impurity. Similarly, the combination of pdays-marital and contact-campaign are effective in generating purer nodes and important target subpopulations.

The visualization of the final tree model up to three levels is as follows:

To reiterate, not all features play an important role while splitting and producing a leaf node with desirable purity.

**SECTION-C: MODEL EVALUATION**

The three classifier models chosen for the modelling of the training dataset to test the effectiveness of their classification are Logistic Regression, Decision Tree and Naïve Bayes classifiers.

* Logistic Regression is a simple yet efficient classification model which doesn’t require scaling or tuning of the input features to produce a probabilistic output. After selection of relevant independent features for the model, its performance works as a strong baseline for the other models being assessed.
* Decision Trees are flexible classifiers that not only provide a highly interpretable visualization for data exploration but they also divide the dataset into target subpopulations for further analysis. The scope for comprehensive analysis that comes with this classifier makes it an obvious pick for historical training of the dataset.
* Naïve Bayes is a scalable classifier designed for large datasets. The high dimensionality and heavy imbalance of the dataset would theoretically suit the nature of this classifier. Overall, it is well designed to handle the independent nature of the features to a good degree while classifying.

The models were optimized using certain parameters and by training the dataset with the input features relevant to each model.

* **Logistic Regression:**

Feature selection using Backward selection and rank widget on orange:

Poutcome, job, prev, contact, education, pdays, campaign

Parameters:

1. Regularization Type: Ridge(L2) (To penalize the sum of the squared values of the coefficients)
2. Strength: C=1 (Cost strength of regularization)

* **Decision Tree:**

Feature selection using Backward selection and rank widget on orange:

Housing, poutcome, contact, campaign, pdays and marital

Parameters:

1. Minimum number of instances in leaves: 5 (This is to prevent the overfitting of data by the resulting tree and to minimize the effect of outliers that create rules only for the specific datapoints instead of the entire dataset)
2. Limit the maximal tree depth to: 20 (This is to ensure that the tree depth is defined)
3. Stop when majority reaches [%]: 95 (To stop splitting the nodes after the specified threshold of 95% for performance optimization)

* **Naïve Bayes:**

Feature selection using Backward selection and rank widget on orange:

Housing, poutcome, contact, campaign and marital

The focus while formulating the models is to maximize successful sale pitches which would be highly profitable and simultaneously, minimize the frequency of fruitless calls to customers to limit monetary expenses and conserve staff time. This translates to minimizing false positives and maximizing true positives. While formulating the models, minimizing the failure to target potential customers has also been considered, i.e., the number of false negatives.

The instances of the target class are extremely skewed i.e., the data in this class is imbalanced. Therefore, because of a majority of negative outcomes in the target class (y), classification accuracy, precision and balanced accuracy over AUC have been picked as the evaluation metrics. The metrics are in line with the focus of making maximum sale pitches without compromising heavily with the expenses related to staff time. The choice of the evaluation metrics has been justified further in the following section.

The confusion matrices and performance measures for each classifier is attached below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Constant** | **Logistic Regression** | **Decision Tree** | **Naïve Bayes** |
| **Confusion Matrix** |  |  |  |  |
| **CA** | 0.888 | 0.900 | 0.899 | 0.899 |
| **Precision** | 0.000 | 0.721 | 0.712 | 0.682 |
| **Balanced Accuracy** | 0.000 | 0.583 | 0.575 | 0.586 |

As it can be observed from the confusion matrices of the three classifier models, even after maximum optimization, the number of false negatives is a considerably large number. However, the focus is to trade off the number of false positives and false negatives, minimizing both. However, it has been emphasized strongly in the message from the CEO that fruitless calls should be avoided whenever possible, i.e. the minimization of the number of false positives should be prioritized. Therefore, performance metrics which take our primary focus into account are precision and specificity. Our second focus is to minimize the number of customers lost due to a wrong prediction (false negatives). This is accounted for by the recall performance metric.

**SECTION-D: FINAL ASSESSMENT**

On comparing the models based on the selected evaluation metrics, the classifier that performs best considering classification accuracy, precision and balanced accuracy is Logistic Regression. With a classification accuracy of 0.9, precision of 0.721 and a balanced accuracy of 0.583, the model does the best job when it comes to making correct predictions about target customers.

A high classification accuracy implies that the model is well trained to make correct decisions on most occasions. From a business point of view, a high classification accuracy means that the model identifies the response of the target class to the campaign correctly on a majority of occasions. A high precision value indicates that if the target customer has been predicted to respond positively to the campaign, the likelihood of the campaign ending up as a failed attempt is relatively low. Balanced accuracy is the average of sensitivity and specificity. A high sensitivity implies that the customers likely to respond positively are predicted more accurately i.e. fewer number of potential customers are classified as not interested. Whereas a high specificity implies that the non-target class is identified correctly with low chances of false prediction.

Avoiding unnecessary calls (False positives) has been the primary focus while picking the winning classifier. Minimizing the number of potential customers misclassified as not interested is obviously our second motive. So, keeping both motives in mind and after trading off between the performance measures for each classifier, logistic regression is chosen as the winning classification model with the best performance measures.

**SECTION-E: MODEL IMPLEMENTATION**

The winning classifier model of logistic regression is attached in the folder as Final\_Classifier.ows. Using this workflow, the logistic regression widget is first trained using the training dataset lixpm28\_csv which is also attached in the folder. The model is then used to make predictions from the test dataset using the Predictions widget on Orange.

The test data is to be loaded first using the file widget named Test Data. The discretize widget is configured on the basis of 10 intervals of equal frequency. Ensure before testing that the campaign feature is left numeric for best performance. Select the same input features as chosen while training the model i.e. poutcome, job, prev, contact, education, pdays and campaign. The Predictions widget needs no further configuration. The different accuracy measures on the Predictions widget evaluate our overall performance of the model when it comes to making predictions based on the test dataset. This can further be visualized using the Confusion Matrix widget linked to Predictions. Since logistic regression makes the classification based on a probabilistic output, one can use the Show Probabilities option on the Predictions widget to analyze the predictions made. A probability more than 0.5 is classified as yes by the model.

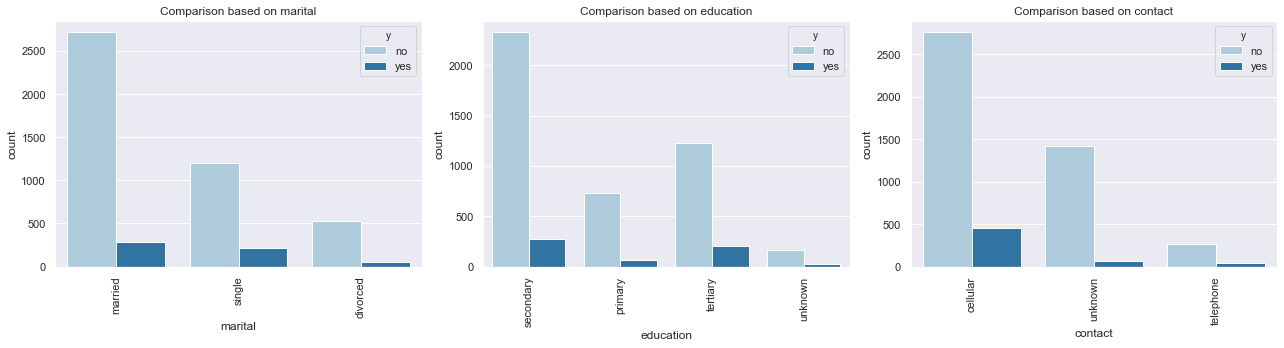
**SECTION-F: BUSINESS CASE RECOMMENDATIONS**

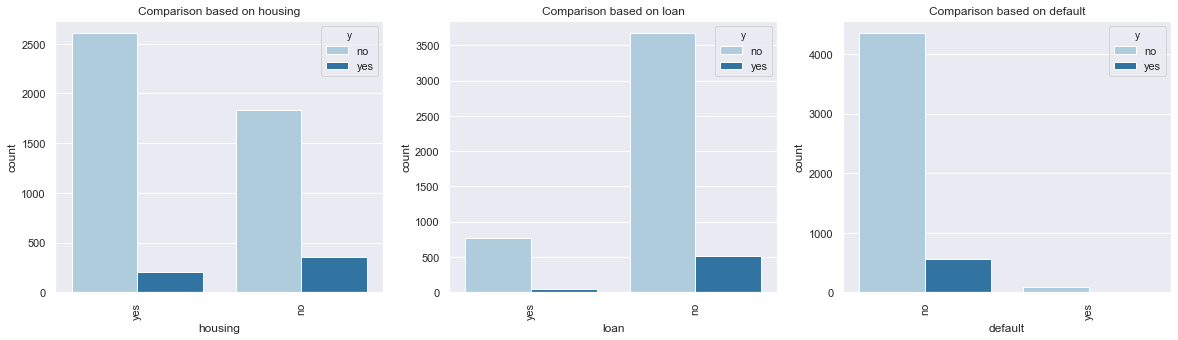
The N/LAB Platinum deposit is a profitable financial product if marketed to appropriate population subsets. Since our focus is to minimize the expenses from fruitless calls, the N/LAB enterprise will benefit a lot from marketing the product to the subpopulations that have a tendency to respond positively to such campaigns. Based on the unpacking of the dataset, for instance, it can be observed that if a student is targeted, he is more likely to subscribe to the product in comparison to customers belonging to some other job category where there is a relatively higher chance of negative response. Similarly, targeting house-maids, entrepreneurs, self-employed professionals and customers belonging to the ‘other’ job category is not advisable since based on the records, the campaign is extremely likely to fail while targeting these customers. It is important to understand that, on unpacking the dataset using the combination of features with a decision tree, few more useful subpopulations can be identified which will streamline targeted marketing further.

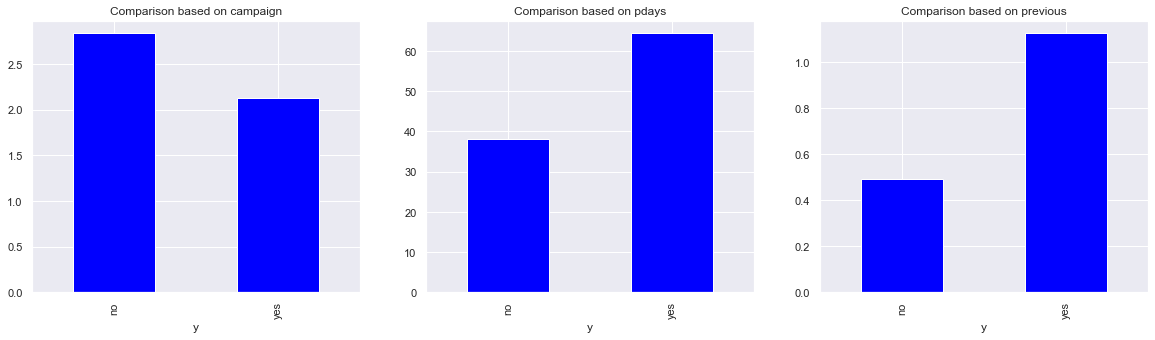
Secondly, it can be observed from the second statistical summary table that, people that the campaign fails to target have a lower Pdays mean than the ones who respond positively. This in a business context, implies that too frequent contacts might result in a customer losing interest in the product. Thirdly, from analyzing the target class response on the basis of their loan sanctions, it can be concluded that it is risky to target customers who have a loan disbursed to them because it is unlikely to engage them in a product requiring a considerable financial investment. Further, aligned with our summary statistics, it can also be inferred that customers with a higher bank balance are more likely to respond positively. But at the same time, vice versa is not true.

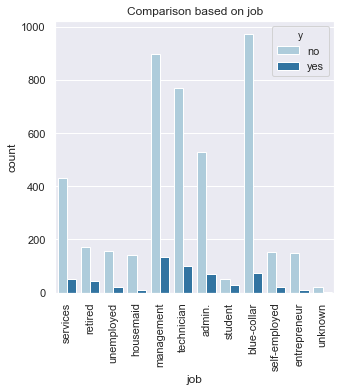
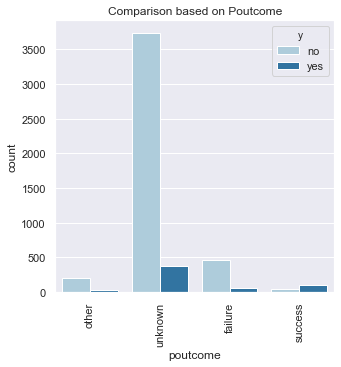
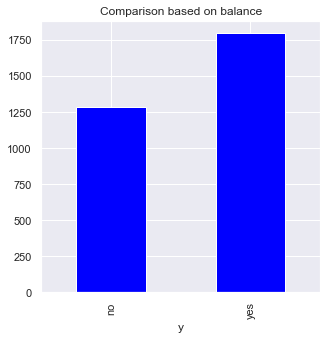
Most importantly, it can be observed that most instances of the poutcome feature are unknown values. A vast majority of the customers who have previously subscribed to a product respond positively to the campaign. If the dataset can be cleaned for this feature in particular, poutcome has the potential to greatly impact the efficiency of our current customer targeting strategy. With efficient targeting of potential customers, the firm can cut down the expenses incurred from fruitless calls and simultaneously maximize successful sales pitches with the predictions from the implemented model.

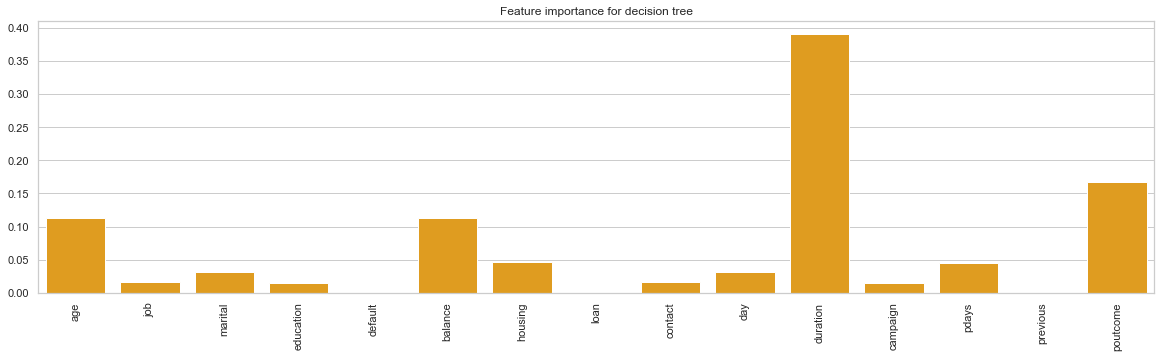
**APPENDIX**

****

****

****

** **

****