Marketing for Term Deposit

Summary – The report presents a framework on how to choose the best model for predicting conversion of potential customers for term deposit. This analysis will develop a better understanding of marketing campaign strategy which will lead to effectively targeting potential customers.

I. INTRODUCTION

The dataset (Bank Marketing Data Set) contains various attributes describing an existing customer of the bank like age, job, education etc. We try to do a feature engineering using our prior notions regarding the data.

We try to find the optimal model on modified dataset (from feature engineering) and try to answer the following questions through our analysis:

- What factors affect the conversion for term deposit?
- Which model explains the relationship in the dataset best?
- What metrics should we judge the model upon?

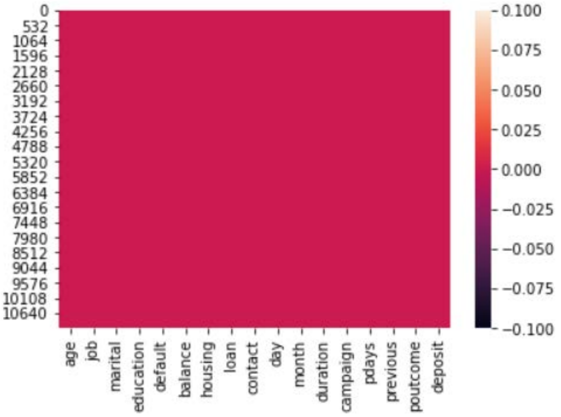
II. ANALYSIS

a) Dataset Summary

The data was loaded from the csv file. There are over 10000 observations in the dataset. There are 15 explanatory (independent) variables with dependent variable being a binary variable containing the response of success or failure of selling a term deposit. There are 10 categorical variables and 5 continuous variables.

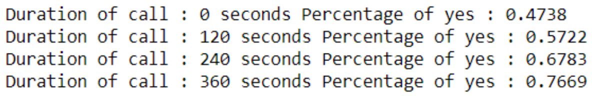
b) Exploratory Data Analysis

First, we check whether any of the columns are missing any data. We create a heatmap to check it to visualize it.



From the graph, it can be observed that there are no missing values.

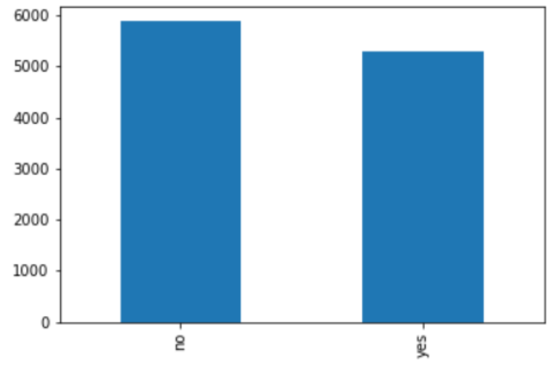
Then, we focus our analysis on 'duration' column. We study the trend of duration of ratio of conversion with duration.



```
Duration of call : 0 seconds Percentage of yes : 0.4738
Duration of call : 120 seconds Percentage of yes : 0.5722
Duration of call : 240 seconds Percentage of yes : 0.6783
Duration of call : 360 seconds Percentage of yes : 0.7669
```

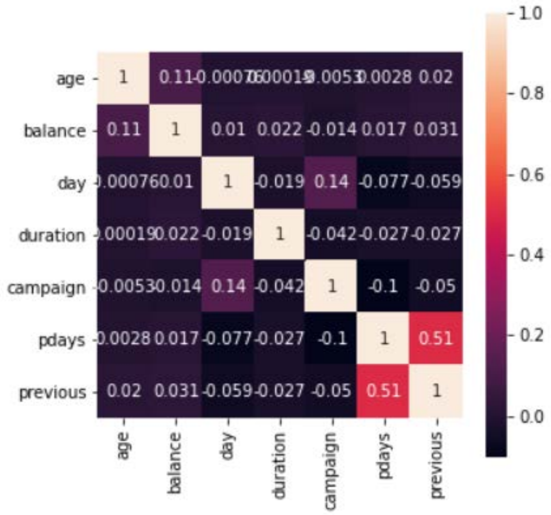
It can be observed duration is highly correlated with output variable, yet we cannot predict the call's duration before the call with the customer. If we include duration in the model, it will lead to erroneous predictions. So, we remove 'duration' variable from our analysis.

We check for balance in dependent variable else we need to adjust the balance using oversampling or undersampling techniques.



As observed, there are almost equal number of positive and negative responses, so sampling techniques are not required.

We study the correlation among the independent variable.

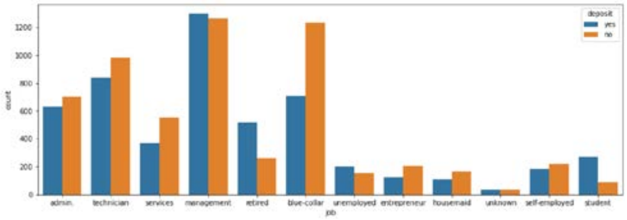


As observed, there is not any significant correlation among the variables except pdays and previous which is expected. '*Pdays*' represents number of days that has passed since last contact which is related with '*Previous*' representing number of contacts performed before this campaign for the client.

c) Feature Engineering

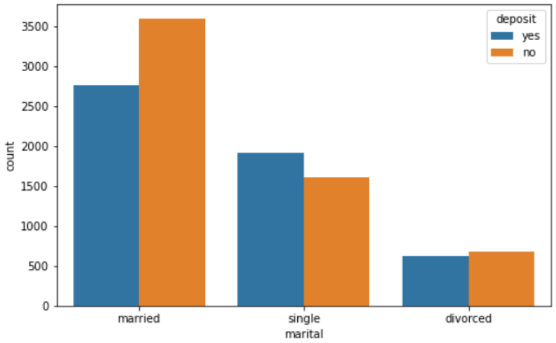
We try to explain which features affect successful conversion of a term deposit offering.

Job – Job of the client



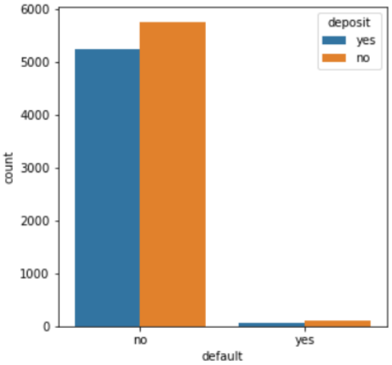
We can observe that the lowest rate of conversion is for clients holding a blue-collar job whereas the highest rate of conversion is for students and retired. Another interesting observation is the conversion of unemployed is more than rejection. It seems counterintuitive at first, as unemployed people will likely have less disposable income. It can be explained by the fact, unemployed will likely save more due to their future uncertainty for not having a job. So, they are more likely to sign-up for term deposits.

Marital Status – Marital Status of Client



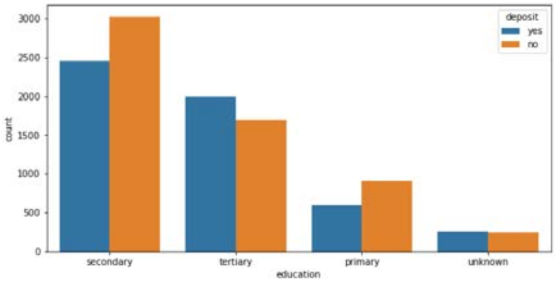
Single clients should be targeted as they are less likely to have other financial obligations.

Default – Whether the client has defaulted



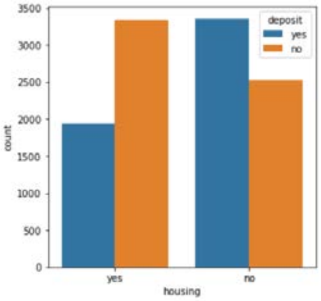
No conclusion can be found as defaults in datasets are very low.

Education – Education level of Client



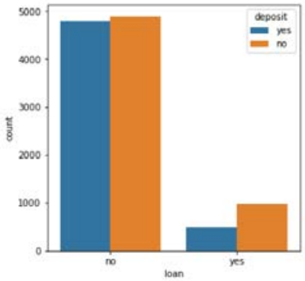
It is generally observed that pay is better as one gets more specialized knowledge. So, clients having tertiary education are more likely to earn more and in turn, will be more likely to sign-up for term deposit.

Housing – Whether the client has a housing loan



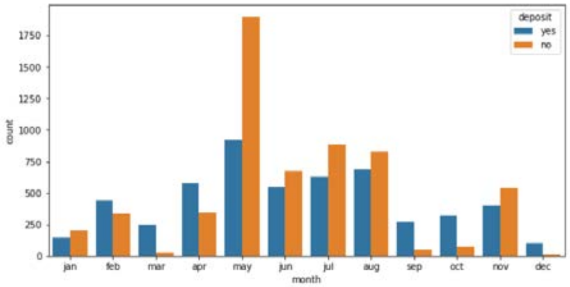
Given an extra disposable income, a person with housing loan will more likely pay the loan than to take a term deposit.

Loan – Whether the client has a loan

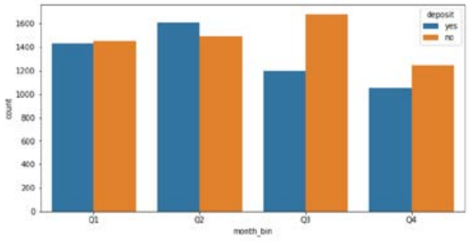


The same argument of housing loan can be applied in this case.

Month – Month of last contact



It can be observed that initial months (Jan- Apr) are favorable for marketing with middle months (May -Jul) being the worst for marketing. We can investigate further by putting the months into bins of 4 (dividing into quarters).

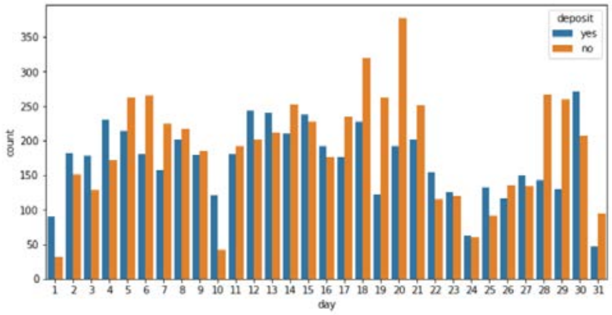


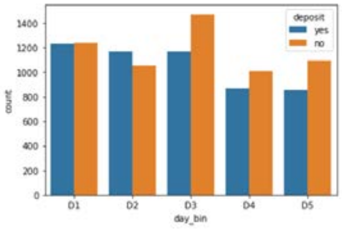
It seems that Q2 being more favorable than Q3 and Q4 with Q1 being neutral. It is still inconclusive as no trend is observed. It could be that month of May could be an outlier due to huge of number of calls in the month thus skewing the results.

So, we remove the column as it is not adding any information to the model.

Day – Day of last contact

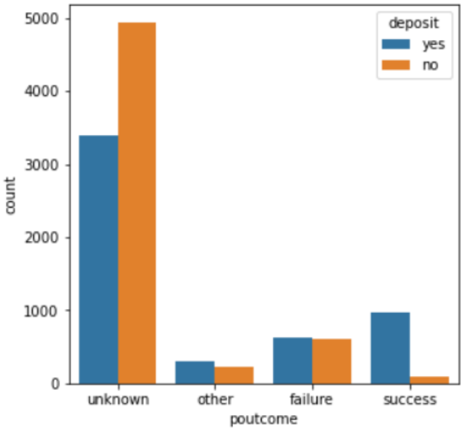
Like month, we do not observe any trend in day variable.





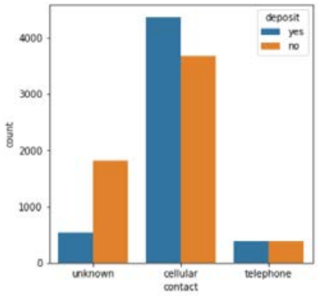
By dividing the days into 5 bins, we do not observe any trend in day variable. So, we remove it from our analysis.

Previous Outcome - Outcome of the previous marketing campaign



It can be observed that if previous outcome was a success, then there is a better chance of selling a term deposit. The unknown outcome could be resulting from the fact that the client is being contacted for the first time in this campaign.

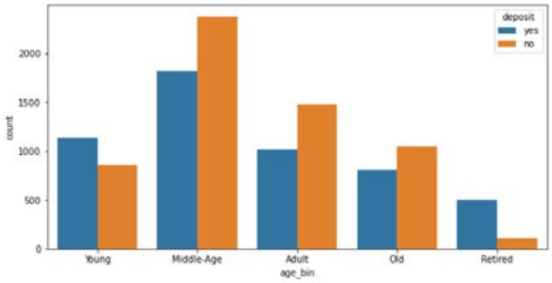
Contact Type – Mode through which contact was made



It can be observed that cellular contact leads to better conversion rate.

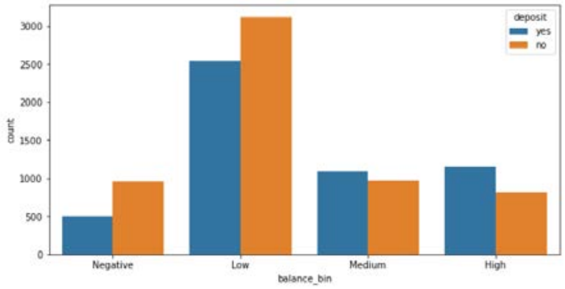
Remark – Following variables are continuous in nature, so we divide them into bins to study the variation.

Age – Age of Client



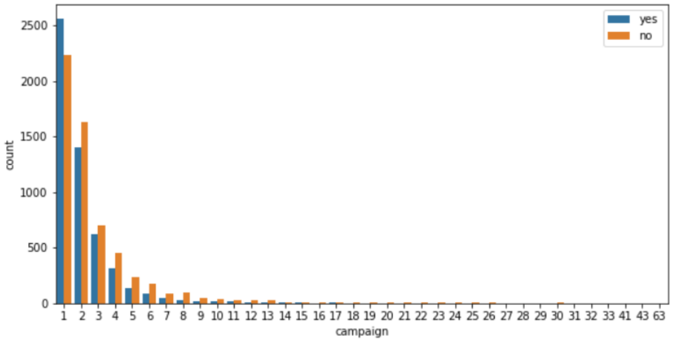
It is in-line with earlier observation from job that students(young) and retired clients are more likely to sign-up for term deposit.

Balance – Balance in account of client with bank

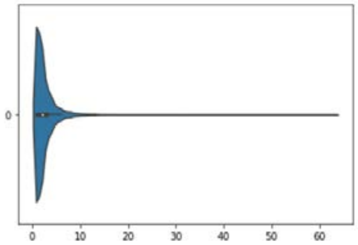


As mentioned earlier, more the disposable income, better the chances of conversion. More the balance, more the disposable income and better the conversion rate.

Campaign - Number of contacts performed during this campaign



As calls are increased, the ratio of conversion falls. Fortunately, usually clients are contacted only once during a marketing campaign.



It is highly concentrated around 1. So, we can infer that usually a customer is contacted once in the campaign.

Pdays–Days passed since last contact with client

Previous - Number of contacts with client before this campaign

Both of variables represent the same information. Also, we observed the high correlation between these two, validating our hypothesis. Thus, we drop ‘Previous’ from analysis.

d) Data Preparation

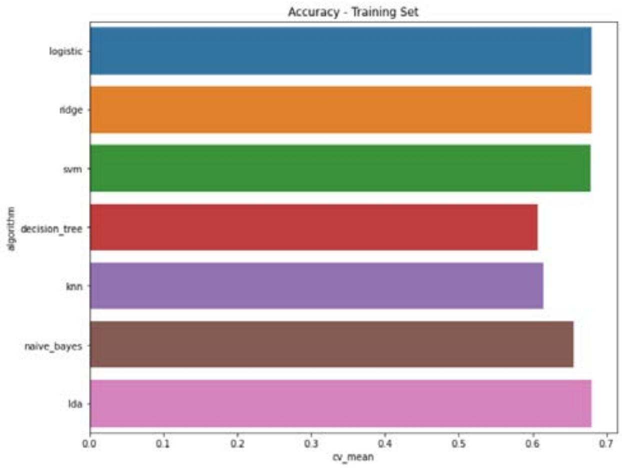
For feeding the data to the model, we remove the unnecessary columns stemming from results of feature engineering. The categorical variables were converted into binary variables by introducing slack variables. For some columns, we decreased the number of categories by cascading them into other categories.

The outliers were removed from the continuous variables as they might lead to erroneous results. A 10-fold cross validation was performed to bring randomness into the sampling. The data set was divided into training and testing dataset with ratio 0.75. Various statistical models were fitted on this modified dataset.

III. RESULTS AND DISCUSSION

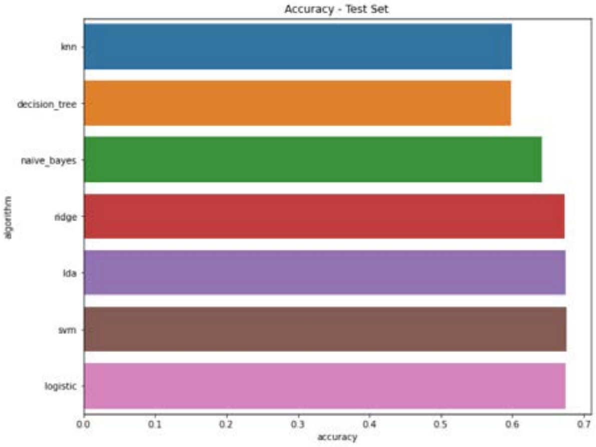
We fitted various statistical models to capture the relationship and predict the dependent variable (conversion of term-deposit). We used Grid Search to fine-tune the hyperparameters for each model. The best performing model was trained using training dataset and then, this trained model was used to predict the testing dataset and performance metrics were gathered.

Accuracy results on training dataset:



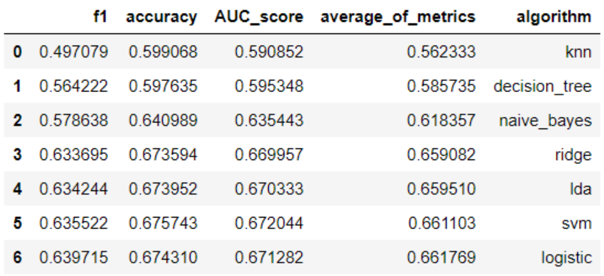
Though accuracy on training dataset do not give an indication on performance of the model, by comparing it to results of testing dataset, we can infer the fitting of dataset (overfitting and underfitting) and suitably adjust the hyperparameters to avoid it.

We judge the accuracy of predictions using test set. We get the following results:



It is observed that logistic regression performs the best measured by accuracy. But accuracy represents the ratio of correct positive predictions to total positive predictions. So, it neglects the negative predictions results, therefore leading to incorrect conclusions for best performing model.

So, we take F1 score and AUC as additional performance metrics to judge the performance of the models. We get the following results:



	f1	accuracy	AUC_score	average_of_metrics	algorithm
0	0.497079	0.599068	0.590852	0.562333	knn
1	0.564222	0.597635	0.595348	0.585735	decision_tree
2	0.578638	0.640989	0.635443	0.618357	naive_bayes
3	0.633695	0.673594	0.669957	0.659082	ridge
4	0.634244	0.673952	0.670333	0.659510	lda
5	0.635522	0.675743	0.672044	0.661103	svm
6	0.639715	0.674310	0.671282	0.661769	logistic

We can observe that Logistic Classification performs consistently across all the metrics, thus leading to conclusion that Logistic Classification is best performing model for the dataset.

Another important metric which is usually not considered would time taken to train the model. As the dataset is for 1 marketing campaign, we need to frequently update the model, by training the model over and over by adding new data from most recent campaigns. These are results that we get for time-taken for training:

Algorithm	Time-taken (in seconds)
Logistic	4.89
Ridge	0.652
SVM	2.23
Decision Tree	9.11
KNN	31
Naïve Bayes	0.015
LDA	0.046

It is observed that Logistic Classification takes around 4.89s to fit the training dataset of 7500 samples (75% of total dataset of 10k samples). It also observed that SVM takes almost half the time for training and LDA takes almost 2 orders of magnitude lesser time for training the same number of sample and almost performing the same as logistic classification across all the metrics. So, it can be argued that LDA is better choice for optimal model, due to assumption that the dataset would increase exponentially, and model should ideally take least time to train while not sacrificing performance for time.

We try to find which variables are being contribute the most to prediction by the best performing models. We take top 3 performing models – Logistic Classification, Support Vector Machine (SVM) and Linear Discriminant Analysis (LDA).

Logistic	SVM	LDA
housing	housing	housing
poutcome_success	campaign	poutcome_success
campaign	poutcome_success	campaign
poutcome_failure	poutcome_other	contact_unknown
poutcome_other	poutcome_failure	poutcome_other

Across, all the best performing models the most important factors are **housing**, **campaign** and **poutcome**.

III. CONCLUSION

This analysis serves as template for those marketing strategists who can use data to plan their subsequent marketing campaign and improve the effectiveness of their marketing campaign. The main findings from the analysis of bank marketing dataset are:

- We find clients having housing loan, whether the client was made contact previously and previous outcome of contact made were the key drivers of conversion.
- Experimenting with various model, we can conclude that LDA is best model for the use case.
- Usually, only accuracy is considered as performance metric. We considered additional metrics like F1 score and AUC to judge the model in a holistic way. Besides the statistical metrics**, time taken** to train the model was also considered due to ever-increasing data and updating the model to capture latest trends.

IV. REFERENCES

- https://archive.ics.uci.edu/ml/datasets/Bank+Marketing

V. APPENDIX

Data file, code and its output files are attached in the zip.