Predicting Success of Portuguese Bank Marketing Campaign

BY: Yunjie Wu, Reem Almijmaj, Ahmed Alsalim

1. Problem definition

Many financial institutions and banks are having difficulty measuring the impact of their marketing effectiveness. Portuguese Bank is one of many banks facing this problem. This is driven by the lack of understanding of customers' needs and efficient marketing strategies. Without a marketing strategy in place, Portuguese Bank will fail to increase its revenue. In this project, we built different machine learning models that takes client features, such as demographics and loan as inputs and predict if the client will subscribe to a term deposit or not as output.

2. Description of background:

This paper [1] analyzed the features in order to identify the bank customers who would positively respond to a new product offering when using a mass media marketing (like radio and television) campaign. The author collected data from Parsian bank and used SVM for classification purposes. The model could help the bank more efficiently select the positively respond customers and improve the efficiency. Another study [2] uses the CRISP-DM methodology, which is a tuning method applied to Data Mining, to improve the performance of the prediction. The author applies SVM, decision tree, and naïve Bayes models to the data collected from the Portuguese banking institute, and the 3 -iteration methodology improves the efficiency of the bank marketing campaigns. In [3], the author used 4 data mining models (Linear Regression, Decision Tree, Neural Network and SVM) to help the managers to choose the next willing-to-subscribe customer. Using the semi-automated feature selection procedure and applying sensitivity analysis, the paper selected 22 relevant and explainable features and Neural networks, which performed the best in AUC and ALIFT metrics as the final optimal model.

3. Project idea:

Nowadays, the financial industry can be challenging for marketers. Therefore, it is crucial for banks to enhance their marketing strategies. Understanding customers' needs lead to increased potential and existing customer satisfaction. In this project, we would like to improve Portuguese Bank's direct phone call marketing campaigns that aim to promote term deposits among existing customers. Our group objective is to use the Portuguese marketing dataset to build different Machine Learning models such as DT, SVM and LR that classifies whether clients will subscribe for term deposits or not. Consequently, we will increase campaign efficiency by selecting high-value customers who would subscribe to the term deposit with fewer phone calls.

4. Description of a dataset:

The dataset is shared by the Portuguese bank institute, and it can be accessed and downloaded from the website: data-society/bank-marketing-data | Workspace | data.world [1]. This dataset is publicly available for research which means the public has a right to access and use the data, but a citation is required. Moreover, the preprocessed data set is in excel format, it has some duplicated values and inconsistency for some rows. It consists of 41188 records, each record represents an existing customer that the Portuguese bank reached via phone calls. In addition, the dataset consists of 21 columns such as Age, Loan, and poutcome: outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success') and so on , the output is y - has the client subscribed to a term deposit? (binary: 'yes','no')

5. Project plan:

After acquiring data, python programming language was used as a main programing language, specifically the Pandas module, to explore and analyze the data. The aim is to know our attributes, data types for example if it is categorical, numerical so we know how to deal with them. Furthermore, we used Seaborn and Matplotlib to visualize statistical graphics to better understand our data. Lastly, we used the decision tree algorithm, SVM, LA as our prediction models.

6. Teammates and work division:

1. Reem: Coding for the purpose of removing duplicate data, unknown tuple values, transforming categorical attributes into numerical attributes, and participating in writing the report.
2. Ahmed: Preparing dataset, treating missing data point, outlier treatment, analysis of the result, final report.
3. Wu: Data visualization, code and model the dataset and give the insight from the output.

7. Method Description and Analysis Result:

The first part of the project is completing data preprocessing and data exploration which include data cleaning, outlier treatment, data categorization, and data creation.

7.1 Missing Data:

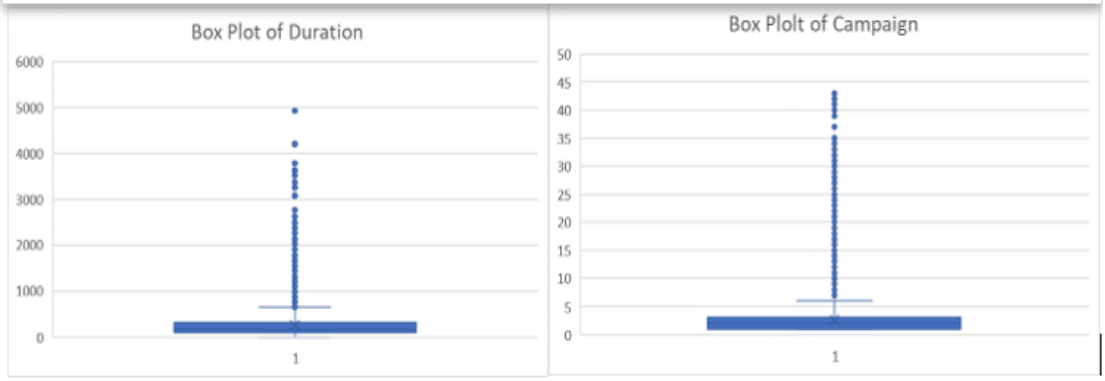
There is no missing value in this dataset. However, there are values like "unknown", and "others", which are useless. They can be treated as a missing value.

- In the column "job", there are 330 records showing "unknown", so we delete them from our dataset.
- In the column "default", there are 8445 records showing "unknown", which is a big number. And there are only 3 records showing "yes", so we decide that we will not use this column.
- In the column "housing", there are 985 records showing "unknown", so we removed them.

7.2 Outlier Treatment:

In the below box plots of columns “duration” and “campaign” (figure1.1), there are some “outliers” values. One method to treat outliers is removing them from the dataset so we applied it to our dataset to get a better analysis. The following formula was used to identify the outliers (Q3 + 1.5×IQR). This shows that values of duration beyond 647 and values of campaigns larger than 6 are outliers, so we deleted them. This led to the removal of 2856 records.

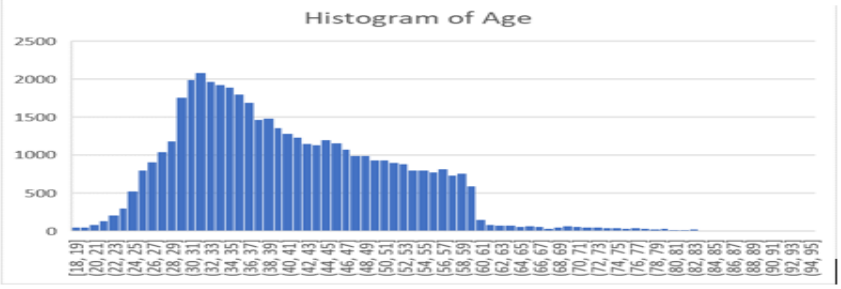
Figure 1.1: Box Plot of Duration and Campaign



7.3 Data Categorization:

As illustrated in the chart (figure 1.2), the participants are in the range of 30 to 59 years old. we have discretized the ages column into 4 groups. Considering those who are younger than 29 years are in the fundamental level of the companies, those who are between 30-45 years are in the middle level, who are between 45-59 years are in the high level, and who are 60 years or older are retired and use their own pensions.

Figure 1.2: Customers’ age distribution



7.4 Data Creation:

To show whether pdays influence customer decisions, a new variable has been created called “is_contact_before”, which has values 0 for those who haven’t been contacted before and 1 for those who have been contacted. It is shown that clients who have been contacted before are more willing to accept the loan. Further, the actual number of days may not be so affectable.

7.5 Feature Selection:

The dataset has 28545 of 31156 records whose “poutcome” equals “nonexistence”. We keep this column because if the customers don’t receive the contact before, the majority of them will not accept the loan. However, if the customer accepted the loan in the past campaign, he will be more likely to accept the loan at this time.

8 Data Visualization:

8.1 Age and Successful Loan

Figure 2.1: % Successful loan and % Age distribution By Age Group

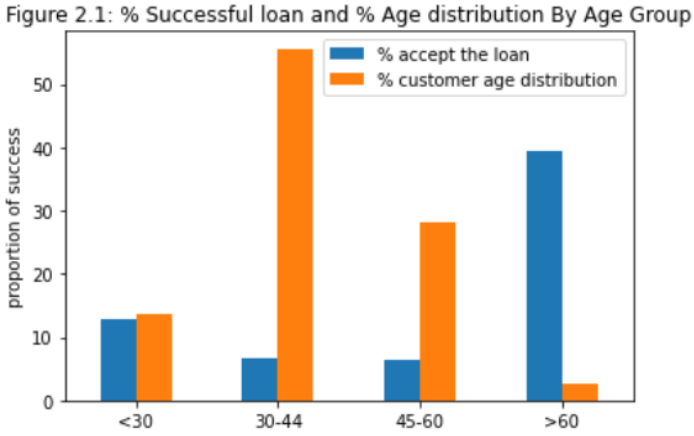


Figure 2.1 shows the proportion of the customers accepting the loan grouped by Age and the distribution of age group in the whole dataset. As our imagine, older people will be more likely to allocate the assets in the low risky investment tools, in this case our loan, which cause the percentage of the successful response of older people is much higher than that of middle-aged people. However, the distribution of the age group in the dataset suggests that the bank put more focus on the middle-aged people (age 30 – 60), which may potentially miss the chance in the 60+ customers.

8.2 Job and Successful Loan

Figure 2.2: % accept the loan By Job

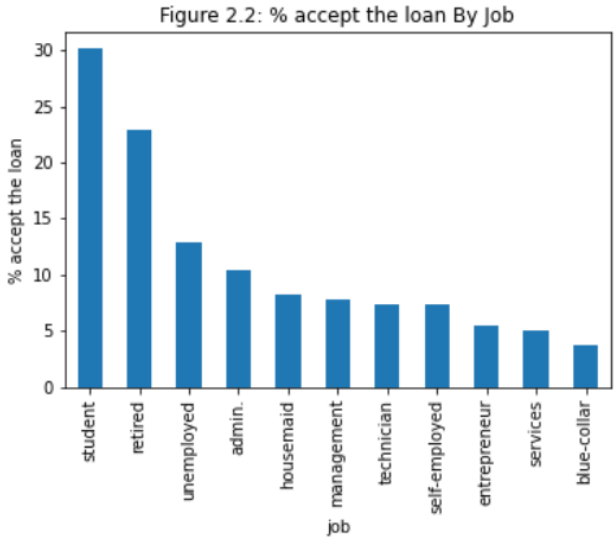
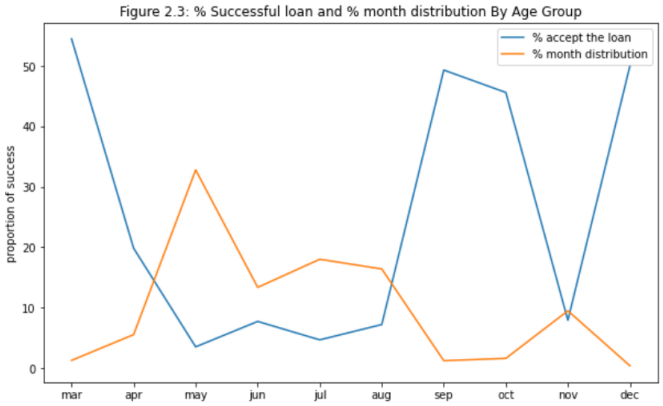


Figure 2.2 shows the percentages of successful response grouped by job. This figure shows the similar pattern as 5.1 shows. Since students are the majority of the age < 30 and people whose age > 60 are most likely retired from the jobs. The percentage of successful response of student and retired people is more than 50%, which indicates potential high profits from these people groups.

8.3 Month and Successful Loan



Due to some data collecting reasons, our dataset missed the data of Jan and Feb, so we draw the figure 2.3 to see whether the successful response is related to the month effect. The blue line shows the percentage of successful response over each month and the orange line shows the distribution of the calls over each month. As the figure shows, the bank made the phone mainly in the summer, from May to August. However, although there are few calls, the perception of successful responses is high in Mar, September October and December, which may potentially show that the response of loan may have the month effects and the bank may decide an inappropriate timing of marketing campaign.

9. Methodology and Result:

9.1 Build Model

In order to evaluate the performance of the models, we use stratified method to split the dataset into 80% training set and 20% test set. We will check the classification models' performance by accuracy, precision, recall, F1-score, ROC curve and AUROC score. We use the most popular and common algorithms with that is used in learning activities in the area of data mining. The model contains:

- Logistic Regression: LR is adaptive to use this probabilistic model to fit the data. It has been used because it is easy to implement, to explain, and efficient to train. Also, LR work better with linearly separable dataset.
- Decision Tree: DT has been used because it is easy to visualize, work well with categorical attributes, and will automatically select the needed features.
- Random Forest: RF is performing well with imbalanced dataset. Also, as the ensemble of the decision tree, Random Forest is not sensitive to outliers and may learn the high-dimension features of the dataset, so it may have a better performance.
- Support Vector Machine: SVM has been selected because it performs well with a clear margin of separation. Also, it performs fast prediction.
- Gaussian Naïve Bayes: a classification model based on the Bayes theorem. This model may not so apply to our dataset since some of our features are not so like in Bayes distribution, but it still could perform well in non-Bayes distribution dataset.

9.2 Model Performance

Receiver Operating Characteristic Curve (ROC) is a metric giving a graphical representation of the performance of the classification model. A well-performed model has AUC close to 1, which means it has a well measure of separability of success and fail. Once the AUC is almost 0, that means it has a worst measure. The ROC curves of five models are shown in Figure3.1, and the AUC Score of Gaussian Naïve Bayes is %75.62 highest score among these five models.

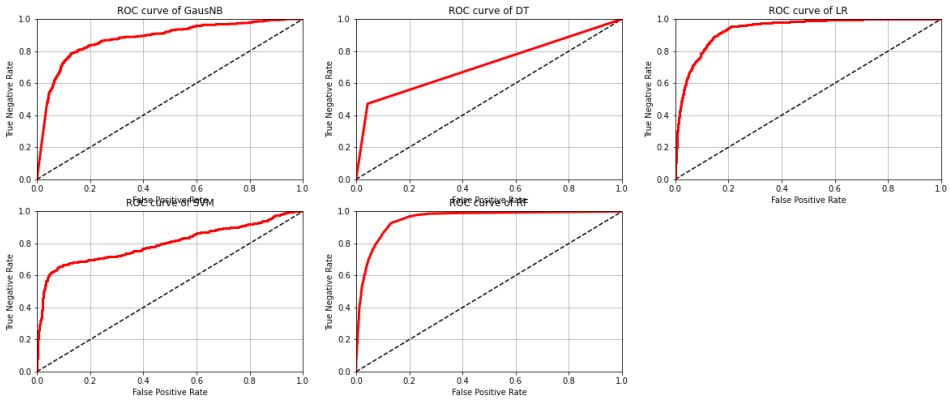


Figure 3.1

Evaluation Metrics of Classification Models on test set

Classification Models	accuracy	precision	recall	f1	roc_auc
GausNB	0.912556	0.475465	0.568826	0.517972	0.756164
DT	0.920749	0.521277	0.495951	0.508299	0.727473
LR	0.936633	0.676923	0.445344	0.537241	0.713104
SVM	0.929276	0.638132	0.331984	0.436751	0.657517
RF	0.939809	0.705521	0.465587	0.560976	0.724046

Table3.2a performance on test set

Evaluation Metrics of Classification Models on train set

Classification Models	accuracy	precision	recall	f1	roc_auc
GausNB	0.910337	0.463477	0.549645	0.502897	0.746211
DT	1.000000	1.000000	1.000000	1.000000	1.000000
LR	0.936546	0.677294	0.441236	0.534356	0.711164
SVM	0.927099	0.617828	0.305471	0.408814	0.644239
RF	0.999958	1.000000	0.999493	0.999747	0.999747

Table3.2b performance on train set

Table 3.2 shows the results of five classification models (Gaussian Naïve Bayes, Decision Tree, Logistic Regression, Support Vector Machine, and Random Forest) displayed with five evaluation metrics (Accuracy, Precision, Recall, F1-Score, ROC-AUC). From table 3.2a, we can see that Random Forest above all give the better results with %93.98 accuracy, %56.01 F1-score and %70.55 AUROC Score among five models compared to other models on the test set. Logistic Regression has almost similar metrics of Random Forest. The accuracy is %93.66, F1-score is %53.72 and precision is %67.69.

9.3 **Unbalanced Data**

Due to the highly unbalanced distribution of success/fail cases in the dataset, the Scores may be a little biased, and there is a need for stratified sampling or rebalancing to deal with this unbalanced structure of dataset. This leads to a good accuracy even on prediction of a near constant label however the F1-score becomes low.

In term of this, we try some methods to solve this problem. One is SMOTE algorithm, which is the popular used resampling method. It is an over-sampling method to deal with the unbalanced dataset. It will forge the class with less number in the dataset by searching the k-nearest point near each point of them. Then the algorithm will randomly produce the fake-samples between the line in the feature space. In our experiment, we will use SMOTE to resample the train set to get 70% train set and 30% test set, then after we create models, we calculate the performance on the test set.

Evaluation Metrics of Classification Models on over sampling test set

Classification Models	accuracy	precision	recall	f1	roc_auc
GausNB	0.910550	0.467305	0.593117	0.522748	0.766123
DT	0.919244	0.510891	0.522267	0.516517	0.738626
LR	0.908544	0.465086	0.714575	0.563448	0.820291
SVM	0.912724	0.478659	0.635628	0.546087	0.786649
RF	0.938137	0.662304	0.512146	0.577626	0.744318

Table 3.3a performance on over-sampling test set

Evaluation Metrics of Classification Models on over sampling train set

Classification Models	accuracy	precision	recall	f1	roc_auc
GausNB	0.836230	0.808554	0.594939	0.685490	0.767286
DT	1.000000	1.000000	1.000000	1.000000	1.000000
LR	0.864966	0.794802	0.741229	0.767081	0.829610
SVM	0.849179	0.812425	0.646502	0.720028	0.791268
RF	1.000000	1.000000	1.000000	1.000000	1.000000

table 3.3b performance on over-sampling train set

The result is shown in table 3.3, we can see that Random Forest still performed the best among five models in the test set, which get %66.23 precision and %57.76 f1 score. However, although we can see that the recall score is higher after the over-sampling, the f1-score did not increase a lot. This may be because most of the variables in the dataset are dummies and it is little useful to create the point between 0 and 1 in the feature space, which implies that SMOTE may not forge very informatic samples.

Also, we try Random under-sampling to the dataset, which means that we throw some failure cases to balance the dataset. This method may have the implied risk that it may drop some information from the training set and may worsen the performance at last. We deduct the failure cases and make the proportion of failure: success = 7:3. The result is shown in table 3.4. Still, the Random Forest get the better performance than others, which get %91.66 accuracy and %61.29 f1-score. This time, the recall scores on the test set increase a lot, at the expense of the precision score. The f1-score did not increase a lot, which means that the under-sampling may be have a little effective, but problem of unbalancing still exists.

Evaluation Metrics of Classification Models on under sampling test set

Classification Models	accuracy	precision	recall	f1	roc_auc
GausNB	0.906872	0.449275	0.564777	0.500448	0.751224
DT	0.884969	0.394565	0.734818	0.513437	0.816653
LR	0.904865	0.453646	0.742915	0.563315	0.831180
SVM	0.931115	0.600490	0.495951	0.543237	0.733122
RF	0.916569	0.496855	0.799595	0.612878	0.863348

Table 3.4a performance on under-sampling test set

Evaluation Metrics of Classification Models on under sampling train set

Classification Models	accuracy	precision	recall	f1	roc_auc
GausNB	0.828723	0.803584	0.567882	0.665479	0.754197
DT	1.000000	1.000000	1.000000	1.000000	1.000000
LR	0.862614	0.792031	0.735056	0.762480	0.826169
SVM	0.818997	0.875360	0.462513	0.605237	0.717144
RF	1.000000	1.000000	1.000000	1.000000	1.000000

table 3.4b performance on under-sampling train set

10. Observation and Conclusion

The main target of this report is to analyze the dataset from Portuguese bank in order to improve the effectiveness of the bank's telephone campaign. In the data visualization part, we found that age 60+ people, who are mostly retired people, and age 29- people, who are mostly students are the most responsive customers, which may need the bank to set more focus on them and make more calls and preferential to attract them. That may help the bank get more potential profits. Besides, the bank may change the major timing of call campaign to winter, such as September and October. The potential month effect of loan response may help the bank get more successful deposit.

As for the model selection, five algorithms (Gaussian Naïve Bayes, Decision Tree, Logistic Regression, SVM and Random Forest) were applied. Among them, the Random Forest Classifier has the advanced performance, therefore could be used as the final model to help accelerate the telemarking campaign opportunity and help the bank wasting less time on the potentially failure customers and increase the bank's earning capacity. However, due to the unbalanced structure of the dataset, the performance is not so excellent. We have tried SMOTE and Random under sampling methods to rebalance the dataset, but the performance is still not so ideal. Maybe there need more technologies and more accurate information to help us improve the performance of the models

11. Challenges

- We had difficulty with our dataset since it is imbalance, it affected the final result, we constantly tried to change some parameters to improve the overall result especially the Recall result. Still, not much of an improvement.
- Rebalancing the dataset using SMOTE and Random didn't improve the metrics results.
- Programming using Machine learning algorithms was intricate, especially with no background in using those algorithms.

Reference

[1] [Moro et al., 2014] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, In press, http://dx.doi.org/10.1016/j.dss.2014.03.001

[2] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62: 22-31, 2014.

[3] S. Hossein Javaheri, M. Mehdi Sepehri, and B. Teimourpour. Response modeling in direct marketing: A data mining-based approach for target selection. Data Mining Applications with R, Elsevier, 6: 153–178. 2014.