Analysis of Bank Telemarketing Data

Kaushik Holla Vaderhobli Madhava Krishna,
Niyati Vikas Chopra, and Kaavya Gowthaman

Summary:

The dataset is based on a marketing campaign conducted by a banking institution. It contains client information collected via phone calls (Moro, Cortez, and Rita 2014). It was downloaded from Kaggle. The goal of the campaign was to get the clients to subscribe to a term deposit. The goal of this project is to predict whether a client would subscribe to a term deposit or not based on the input variable.
The input variables are divided into four categories:

1. **Bank client data**: This contains personal information about the client, like their age, education, job, marital status, if they have a housing loan, are they defaulters, and if they have a personal loan.
2. **Previous contact data**: This contains data about any previous contact the institution had with the customer. This includes the type of contact, day, month and time of contact and duration of the contact.
3. **Social and economic context attributes**: This includes social and economic factors at that time. These include employment variation rate, consumer price index, consumer confidence index, Euribor rate and a quarterly indicator of the number of employees.
4. **Other attributes**: These include number of previous contacts for this client in this campaign, number of days passed since last contact, number of contacts before this campaign for this client, and outcome of those contacts.

The value that is to be predicted is the 'y' variable in our dataset. It has the value 'yes' if the client subscribes to a term deposit, or 'no' if the client does not subscribe to a term deposit.
There are a total of 20 input variables and our goal is to fit classification models on our dataset to predict the value of 'y'. Classification is a machine learning technique that involves predicting the value of a categorical variable. In the case of this dataset, our categorical variable has two values, yes and no. Variables are analyzed and those variables that have a strong relation with the 'y' variable are chosen to fit the best possible model for the data set. Logistic Regression, Support Vector Classifiers, K Nearest Neighbors and Random Forest Classifier were fit onto this dataset for prediction. Model comparison was done using the F1 score obtained after generating the confusion matrix, a matrix that contains the values predicted as true positives, true negatives, false positives, and false negatives.

Methods:

The problem being solved is a classification problem, I.e., to check if a customer will subscribe to term deposit or not. The project is split into four categories, Exploratory Data Analysis, Initial Data Analysis and Under-sampling, Feature Engineering, and Machine Learning.

Part 1. Exploratory Data Analysis:

Exploratory Data Analysis is a technique in which the dataset is analyzed and studied using visual techniques. Below are some of the major conclusions concurred after performing exploratory data analysis:

Fig. 1

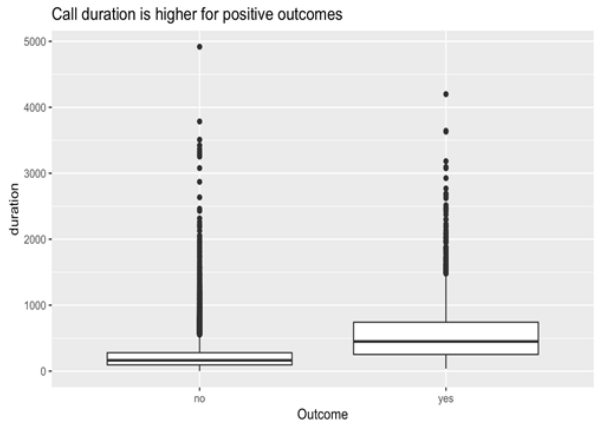


Fig. 2

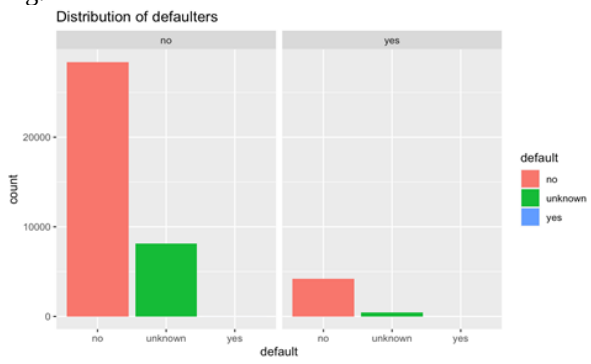


Fig. 3

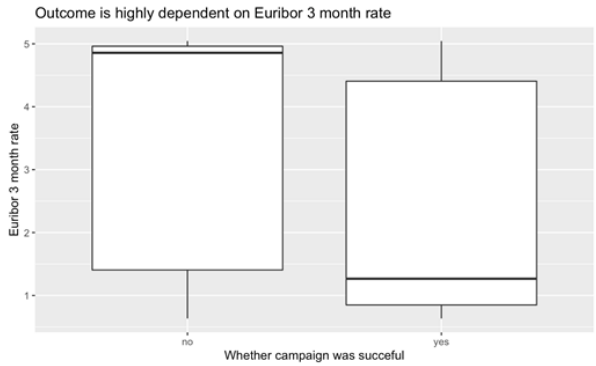
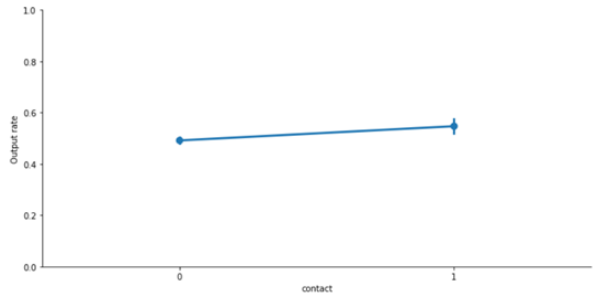


Fig. 4



In fig. 1, it was concluded that if the call duration is higher, it is likely that the customer will subscribe to a term deposit.
Fig. 2 represents the distribution of the people who were previously defaulters. Due to imbalance in data, no conclusion could be derived as to whether a defaulter would or would not subscribe to a term deposit.
From fig. 3, it can be assumed that the outcome of whether a person subscribes to a term deposit or not is highly dependent on the Euribor three-month rate. The Euribor three-month rate is the interest rate at which a panel of banks lend each other money with a maturity term of three months.
Fig. 4 concurs that if there was previous contact with the customer, then the probability that they subscribe to a term deposit is higher.

Part 2. Initial Data Analysis and under-sampling:
After investigating the data, it was found that the data was already in 'tidy' format. There were no missing values, and all row and column formats were appropriate. However, it was observed that the data was highly imbalanced I.e., out of 43128 instances, only 4096 belonged to minority class (customer subscribing to term deposit). Fitting any model on imbalanced data would have resulted in the model being bias towards the majority class (customer does not subscribe to term deposit).
To overcome this issue, under-sampling was performed. In under-sampling, only those instances of majority class with minimum average distance to 3 closest minority class examples are taken into consideration. After under-sampling, the ratio of majority class and minority was brought down to 1:1.

Part 3. Feature Engineering:

Feature engineering is the process of using data to construct explanatory variables and features, that can be used to train a predictive model. A machine learning model finds patterns in the data in an automated fashion. These pattens provide insights which are used to make decisions or predictions.

To improve the performance of such algorithms, feature engineering is beneficial. Feature Engineering is a data preparation process. Engineering and selecting the correct features for a model will not only significantly improve its predictive power, but also offer the flexibility to use fewer complex models that are faster to run and more easily understood.

The feature engineering methods that we employed were:

1. Correlation table: A correlation table between all the features and output variable in our data was constructed. Two things that need to be seen from the table are: if there exists a high positive/negative correlation between our features and the output variable, and if there exists a high positive/negative correlation among the features in the table.

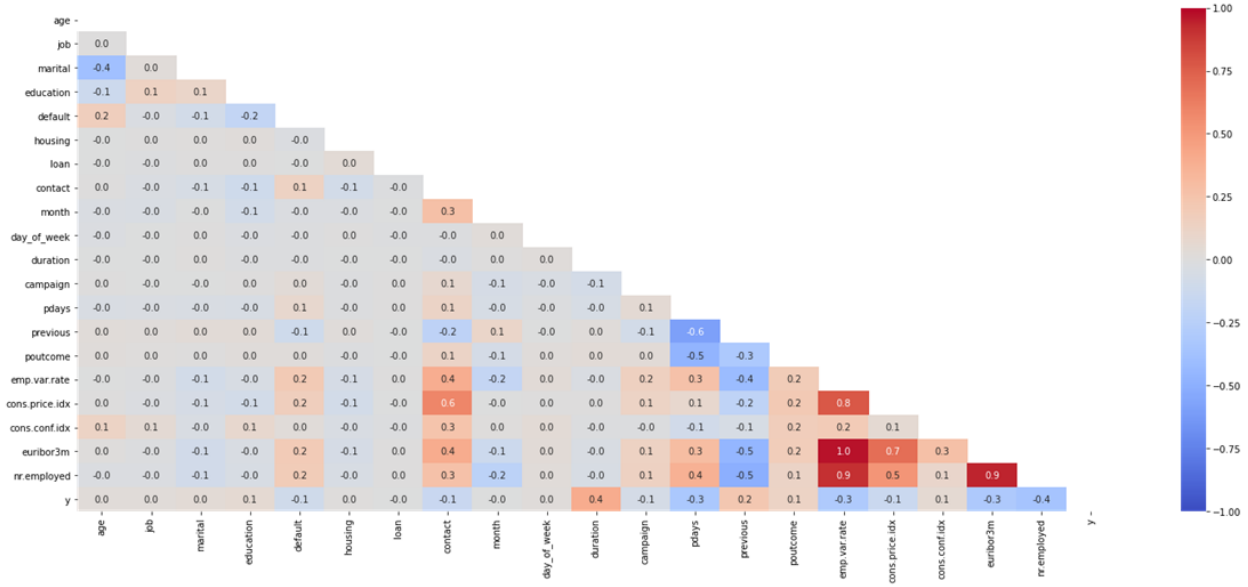


Fig. 5: The correlation table

Features having high positive/negative correlation with output variable suggest that they are good features as they vary with the output. Hence, they capture most of the variation in the dataset. It is important that features have less correlation amongst themselves. If two features are highly correlated, it leads to overfitting. To avoid overfitting, highly correlated features are identified and dropped.

2. Categorical plots: This is the second type of feature engineering method that was implemented. Here, the dots represent the output rate (probability of the customer subscribing to a term deposit) and the vertical line represents the error rate. An ideal graph is one with less error rate and good output rate. If the error rate for the features is high, then dropping them can be considered.

Fig. 8

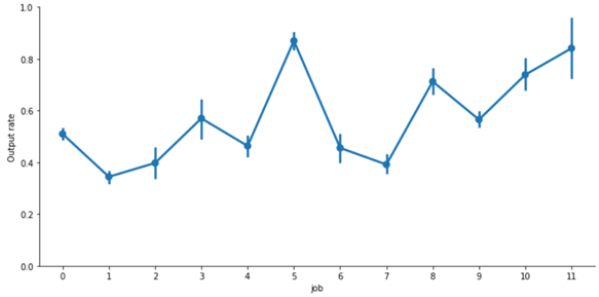


Fig. 9

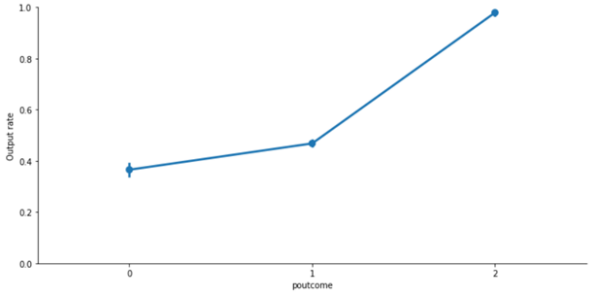


Fig. 6 and fig. 7 are some examples of categorical plots

3. Statistical Significance Test: Statistical significance tests were performed on all the features. The Null and Alternate Hypothesis for the model is described below.

- H0(Null Hypothesis): The feature is not statistically significant with respect to the output
- H1(Alternate Hypothesis): The feature is statistically significant with respect to the output

P-value for all the features with respect to the output was calculated. The alpha value is set to 0.05. This means if the p-value for any feature is greater than 0.05, then it fails to reject null hypothesis hence we can consider dropping this feature. These were the p-values obtained:

Feature	P-Value
Age	0.551
Job	0.495
Marital	0.00604
Education	0.000896
Default	2.47e-13
Housing	0.719
Loan	0.319
Contact	6.08e-10
Month	0.839
Day of Week	0.000613
Duration	4.38e-19
Campaign	0.466
pdays	2.25e-66
previous	2.04e-19
poutcome	1.37e-56
Emp.var.rate	1.09e-10
Cons.price.idx	8.1e-05
Cons.conf.idx	1.83e-05
Euribor3m	1.24e-19
Nr.employed	3.27e-38

Fig. 8 The p-values

As observed, the p-values for Age, Job, Marital status, Housing, Loan, Month, and Campaign are very high. These features can be dropped as they are not statistically significant.

Part 4. Machine Learning:

The main goal behind fitting multiple models, is to find the most accurate fitting model on the dataset. To perform model comparison, precision and recall are calculated from the generated confusion matrices. These are then used to calculate the F1 score of the model. Precision measures out of all the customers that subscribed to term deposit, how many were predicted correctly, and recall measures out of all the customers that the model predicted to subscribe to a term deposit, how many were correct. F1 score is the harmonic mean of precision and recall.

Before fitting any models, label encoding was performed on the dataset. This converted any categorical variables to numerical variables to eliminate problems while model fitting. Then, a standard scalar is applied to the data set to get all the values onto the same scale for better performance and evaluation of the models.

Finally, the following variables are considered for model fitting: education, default, contact, day of week, duration, pdays, previous, poutcome, emp.var.rate,cons.price.idx, and cons.conf.idx. The dataset was also split into a training set (80%) and a testing set (20%).

Four machine learning classification models were fit on our data and this is how they compared in performance:

1. Logistic Regression: It is considered one of the simplest and most effective models for classification. A logit function or log of odds function is applied to the data. Odds are the ratio of an event happening versus the probability of the event not happening. This model gave an F1 score of 0.82.
2. Support Vector Classifier: This model uses the data dimensions to construct a threshold 'hyperplane' in the data space. Based on which side of the hyperplane the points lie, it is classified to the appropriate output class. This model gave an F1 score of 0.80.
3. K Nearest Neighbours: This model makes use of the subspace to compare every point to its 'k' nearest neighbors. Then chooses the maximum of the outputs of the k nearest neighbors to classify the current data point. On comparing the calculated error rate with a range of values of k, the least error rate was obtained when k was equal to 27. This model was fit and gave an F1 score of 0.79.
4. Random Forest Classifier: The dataset is bagged into various smaller datasets (with repetition) and each of these is passed through multiple decision trees. The maximum of the decisions of the decision trees is taken as the final answer. This model gave us an F1 score of 0.80.

Thus, logistic regression was the best fit model on the dataset.

Results:

Based on the initial analysis of the models, logistic regression had the highest F1 score. Let's look at the confusion matrices of all the models for comparison:

Fig. 9: Logistic Regression

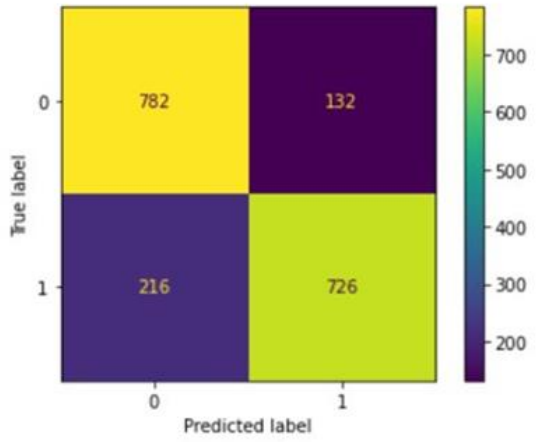


Fig. 10: Support Vector Classifier

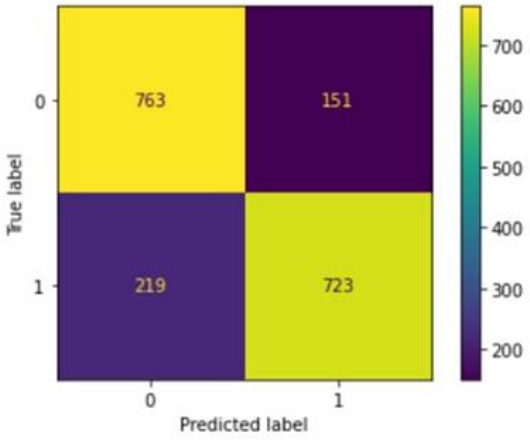


Fig. 11: K Nearest Neighbors

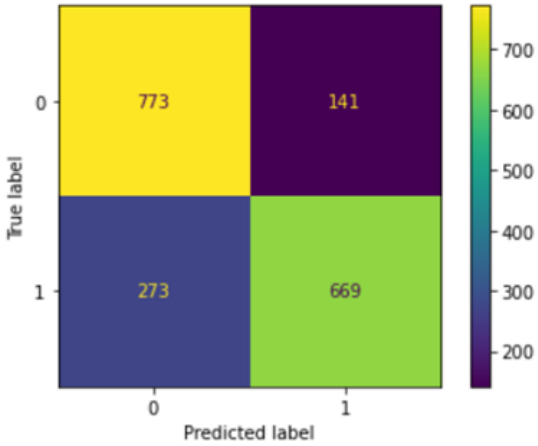
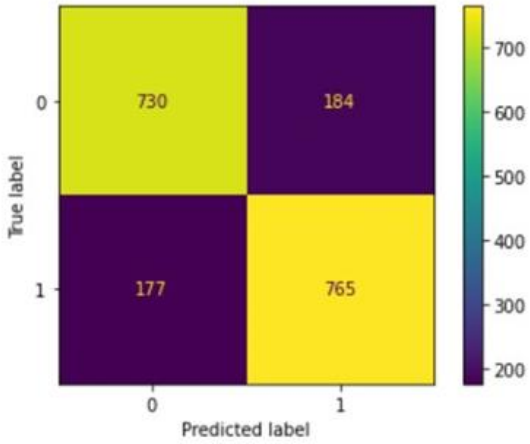


Fig. 12: Random Forest Classifier



From the above confusion matrices, we can also calculate the accuracy of these models. Accuracy is the ratio of how many predictions were correct versus the total number of predictions.

- Logistic regression: 81.25%
- Support Vector Classifier: 80.06%
- K Nearest Neighbours: 77.69%
- Random Forest Classifier:80.54%

Therefore, Logistic Regression is the best fit model on our dataset with an accuracy of 81.25% and an F1 score of 0.82.

Discussion:

The obtained results show that logistic regression is a simple and very effective model for classification problems. As a result of the work done on this dataset, the banking institution carrying out the campaign will greatly benefit in assigning resources to the customers that have a probability of subscribing to a term deposit. This will save them time and effort and even money. The institution can now find out if a customer will subscribe to a term deposit based on the user's data as well as other available data, and social and economic factors.

These results can be used to make better informed decisions by only considering those customers whose data, when fed into the model, gives a positive outcome. Thus, based on the model results, they can decide whether to approach a customer or not. This will also benefit in not developing a negative experience for the customers by not approaching any customers that were predicted to not be interested in a term deposit.

Some improvements that can be made to improve the model are:

- Using better under sampling techniques to deal with the imbalance in the dataset, such as SMOT analysis
- Trying more complicated machine learning models, like neural networks
- Collecting more information about the customers to get more variables in our dataset
- Using a feedback-based approach to improve data accuracy and implement any user-based changes that can be made

Statements of Contribution:

While working on the project, the work was divided into three parts, I.e., initial and exploratory data analysis, feature engineering and machine learning. Each of the team members worked on one part and then verified the other two parts for more uniformity in the work.

1. Kaushik Holla Vaderhobli Madhava Krishna: Kaushik worked on the feature engineering part of the project. He worked on generating the correlation table, categorical plots, feature extraction, feature selection and performing hypothesis tests on each of the features. He also worked on under sampling the dataset to get a 1:1 ration between the minority and majority classes. He supervised the building of the plots for exploratory data analysis and label encoding as well as feature scaling.

2. Niyati Vikas Chopra: Niyati worked on machine learning. She carried out train-test split and model fitting. She then fit machine learning models, calculated error rate to get the appropriate number of k neighbors in the model, generated the confusion matrices and calculated the model evaluation metrics. She also supervised exploratory data analysis and label encoding and standard scaling.

3. Kaavya Gowthaman: Kaavya was responsible for identifying the problem, collecting all relevant information on the dataset, putting it into usable format, and performing initial as well as exploratory data analysis. She generated all relevant plots for analysis and derived conclusions (such as imbalance in the dataset, the types of features, etc.) She also verified the methods for feature selection and choosing the best model for the classification problem.

References:

Dataset: https://www.kaggle.com/henriqueyamahata/bank-marketing

Code references:

- https://scikit-learn.org/stable/modules/svm.html
- https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html
- https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html

Term references:

- Euribor 3 month rate: https://www.homefinance.nl/english/international-interest-rates/euribor-rates-3-months.asp#:~:text=The%203%20month%20Euribor%20interest,a%20maturity%20of%203%20months.&text=The%20first%20rate%20of%20every,like%20mortgages%20and%20savings%20accounts.
- Employment variation rate: https://www.scu.edu.au/staff/hr-services/hr-policies-procedures--guidelines/employment-variations/#:~:text=Employment%20variations%20occur%20when%20a,time%20or%20fixed%20term%20employee%3A&text=Increases%20or%20decreases%20their%20hours,a%20temporary%20or%20permanent%20basis
- Consumer price index: https://www.bls.gov/news.release/pdf/cpi.pdf
- Consumer confidence index: https://en.wikipedia.org/wiki/Consumer_confidence_index

Appendix:

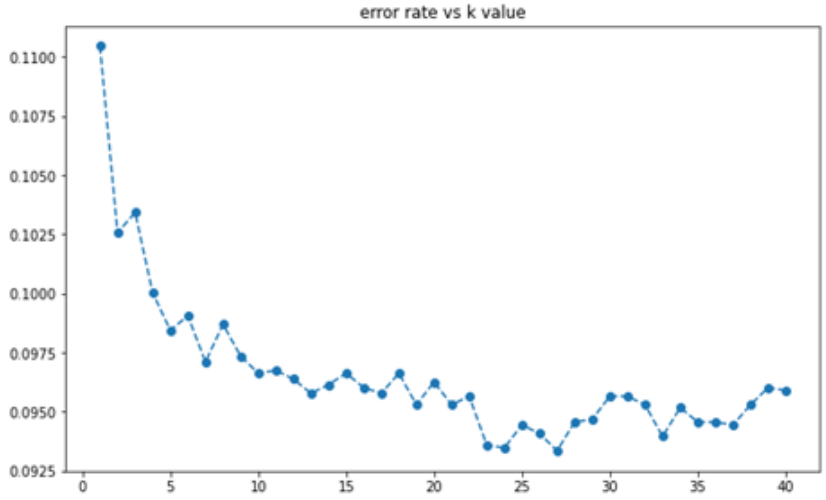
Positive Correlation: means that if feature A increases then feature B also increases or if feature A decreases then feature B also decreases. Both features move in tandem and they have a linear relationship.

Negative Correlation: means that if feature A increases then feature B decreases and vice versa.

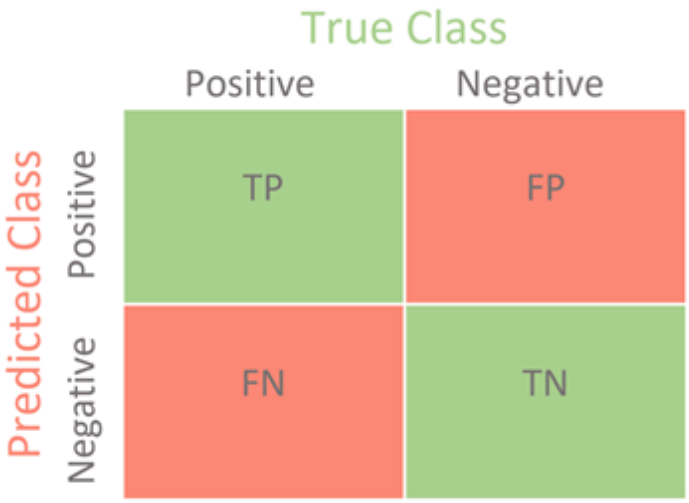
Code: https://github.com/kaushik-holla/Bank-Marketing

Other relevant plots:

1. Error rate vs K value for choosing best K value for K Nearest Neighbors.

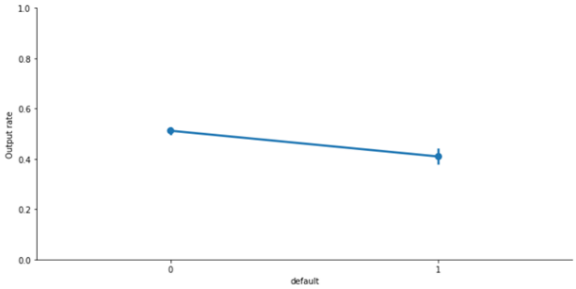


2. Standard confusion matrix:

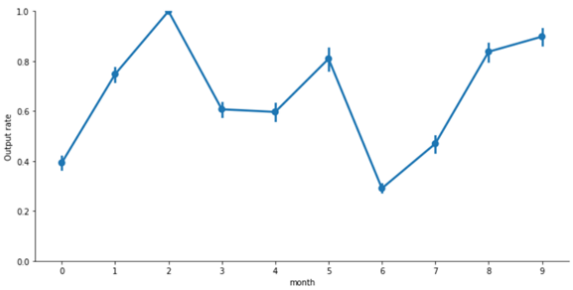


3. Some more categorical plots:

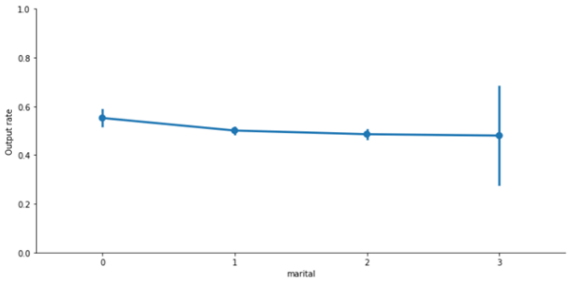
3.1 Whether the customer is a defaulter or not



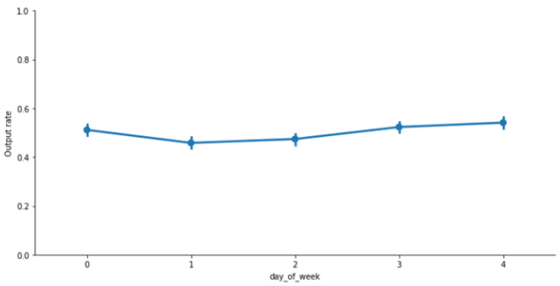
3.2 Month of contact



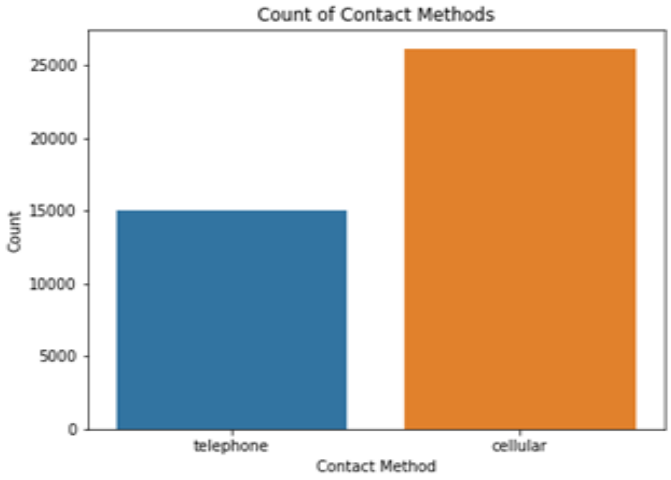
3.3 Marital status of customer

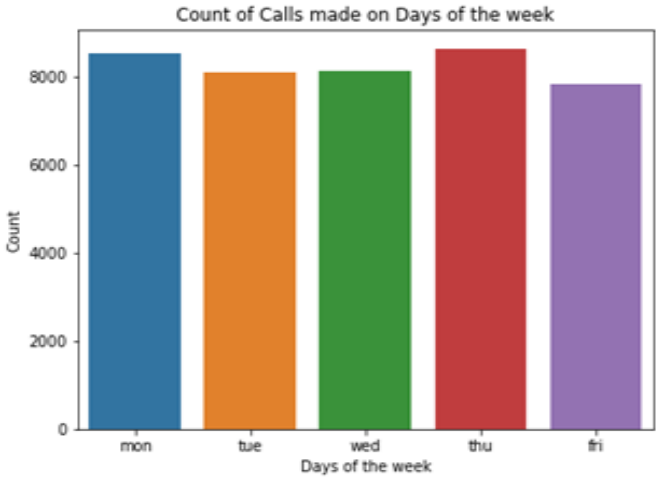


3.4 Day of the week contacted

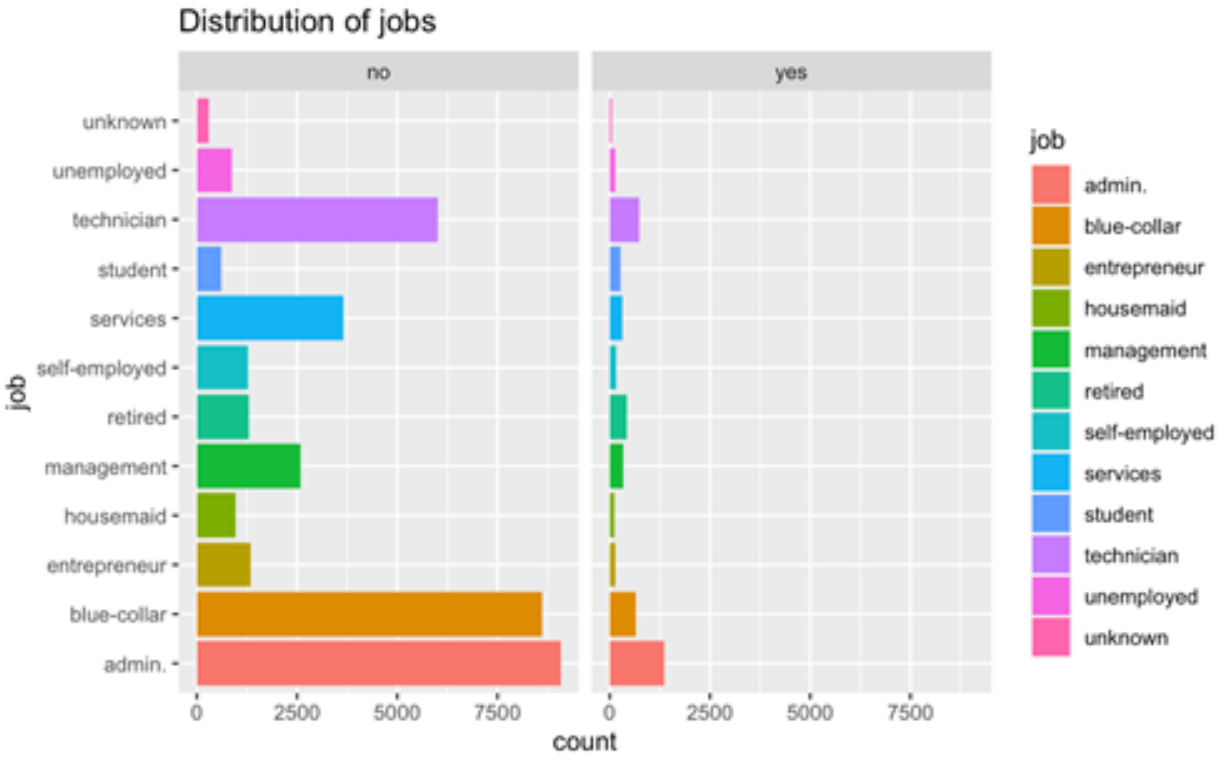


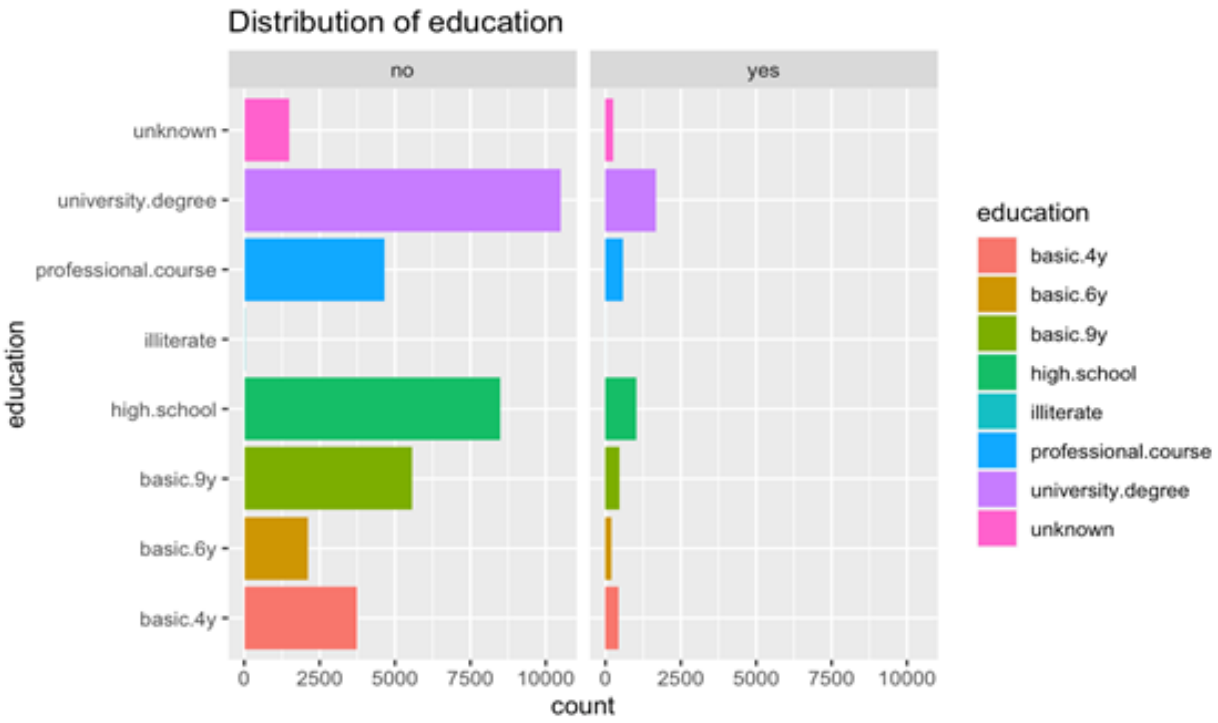
4. Exploratory data analysis on last contact data:





5. Exploratory data analysis on customer data:





6. Exploratory data analysis on socio-economic factors: