**Applied Machine Learning in Financial Modeling**

**Final Report**

Machine Learning in Bank Telemarketing Campaign

Author:

Bin Xu (bx2161)

Aishen Li (al3836)

Zichen Pan (zp2197)

Ruxin Shen (rs3878)

Xiaoxin Xu (xx2307)

Jinhao Zhang (jz3011)

**1. Introduction and Business Understanding**

Bank Marketing campaigns offers a powerful and effective way for companies to introduce their product and build a connection with their customers. Among them, telemarketing is one of the most direct and frequently used approaches in which salesperson reach potential customers through phone calls and introduce their products. Compared to other approaches such as advertising, telemarketing could reach more customers with less time cost. However, though the cost is comparatively lower, sending calls still requires much time and people if the target customer lists are very long. Therefore, how to accurately reach those target customers is still challenging. In the past, people might do simple filter for target customer using age, income, and gender regard to the product. For example, females are more likely to purchase lipstick than male customers, and male customers are more likely to purchase lighter than females. However, this method is not valid because it would lose many potential customers. To improve this and make telemarketing more efficient, we machine learning techniques come to the play.

Before building the machine learning models, we need to fully understand our business goal as well as the dataset itself. In telemarketing, industry experts divide contacts inbound and outbound depending on which side triggered the contact (client or contact center) with each case posing different challenges. For example, outbound calls are often considered to be more intrusive. In this paper, we focus on the previous situation(outbound). Machine learning techniques could improve the telemarketing by maximizing the customer lifetime value through the evaluation of available information and customer metrics, thus allowing us to build a more extended and tighter relationship in alignment with business demand. Also, it should be noted that the task of selecting the best set of clients who are more likely to subscribe to the product is considered an NP-hard Problem. Therefore, implementing which classification model and how many features are feed into the model will both significantly affect the final prediction result.

There are many matured classification models such as classical Logistic Regression (LR), Naïve Bayes(NB), Random Forest (RF), Neural Networks (NNs), Support Vector Machines (SVMs) and some most recent models such as XGBoost and LightGBM. Among these models, LR and NB models are usually used as evaluation baseline due to its advantage of easy implementation and understandable by human beings, while also giving good prediction results. By contrast, NN requires more computation power, and the model is more difficult to be interpreted by human beings. However, with very high prediction accuracy and hardware such as CPU and GPU getting cheaper, NN is becoming increasingly popular now. A similar issue happened with SVM that gives high prediction accuracy in high dimensional data, but the obtained model is hard to be understood by human beings. However, we can open these “black box” model can by using sensitive analysis, which allowed us to measure the importance and effect of particular input in the model output response. Compared to previously mentioned models, Random Forest, along with other tree-structured models outperformed the LR and NB in most of the cases with higher prediction accuracy. On top of that, the obtained tree-structured model is more readily open to be interpreted with tree visualizations instead of like a black box in NN and SVM. Therefore, a tree-structured model can potentially be an ideal model for our dataset, and we would compare it with other models in the model selection section.

In this paper, our thesis is to test the hypothesis ---- whether adding macroeconomic features, tope.g. social, economy to the overall dataset could improve the performance of the whole classification model. The model is trained to predict whether telemarketing phone calls target would make long-term deposits. We collect the dataset from a Portuguese retail bank during the year of 2008 and 2010 with a total of 45,211 phone contacts. We enriched these records by adding 18 social and economic influence features, which we downloaded from the Portuguese Republic statistical website. The added features include macroeconomic indicators such as Economic Sentiment Indicator, Active Population Rate, Unemployment Rate Indicator of Financial Stress. The merging of the two data sources results in a significant increase in the prediction accuracy. Conducted correlation analysis and implemented stepwise selection, we finally picked 5 features as an addition to the model. Since the dataset also includes the data during the 2008 financial crisis, we also did some feature analysis to see the impact of macroeconomics on people’s behavior in making financial decisions.

Furthermore, we are going to compare different models and evaluate its performance using different metrics such as accuracy, AUC-ROC curve, as well as the baseline.

We organized the paper as follows: Data Understanding presents the bank data and feature selection we implemented for macroeconomic features we added. Model Selection presents the models used in previous work and evaluation between different models with different metrics. Descriptive Statistics presents the data insights and value-added derived from the dataset and the model. Functions Type and Variables Relationship presents the statistical techniques applied in the model and the reason behind it. The last model performance evaluation part presents the final results supported by different metrics.

**2. Data Understanding**

**2.1 Client Data Set**

Our study focused on a real dataset of 45,211 records come from the UC-Irvine data repository. The dataset consists of direct marketing campaign information of a Portuguese banking institution. While people at the call center for this bank excuse phone calls to sell term deposit to a real person, the data was recorded to analyze potential customers and find values to improve the banks telemarketing performance. This dataset has its granularity at the personal level; each record contains the demographic information and statistics of the call, they are labeled with whether subscripted a term deposit with the banking institution.

The demographic information for each person in the dataset includes age, some categorical features such as job, marital status, education level, and some binary variables such as if the person has credit in default if the person has a housing loan if the person has a loan. Call information includes the type of contact, the day of week and month information for the last contact, duration of the last contact, and information for the last campaign. It contains the number of contacts in the last campaign and total previous contacts, days since the last contact, and the outcome from last campaign( whether the person subscripted a term deposit or non-exists if the person was not involved in the last campaign).

**2.2 Macroeconomic Data Set**

We have our macroeconomic dataset from Eurostat, which is the official statistics office for the European Union. We take 19 indicators that can represent the economic situation of Portugal from 2008 to 2010. After examination of the correlation with the original features at the person level, and empirical analysis of the relationship of the indicator on people’s behaviors of saving, we picked 5 features to add to the model.

**Payment - debt ratio**

The payment to debt ratios measures a residents ability to repay their monthly debt. Since people are more likely to subscribe to the relationship of the indicator ave excess capital after the debt payments. This ratio can contribute to our model to determine the likelihood of making a deposit.

**Unemployment rate**

This rate can imply the relationship between the indicator of society, which can be rather crucial for people to make deposit decisions.

**Active population rate**

This rate measures the percentage of the total population that are economically active, disregard of employment status; this would indicate the population that can make a deposit.

**Financing of residents**

This indicator measures how much people like to practice financing. We expect that people are less likely to subscript term deposit while they are likely to do financing. Thus we believe this factor would improve the predictive model.

**Economic Sentiment Indicator**

This indicator measures the confidence level of the economy, and thus affect the overall level of willingness to make term deposit.

We merged the 5 macroeconomic features into the original dataset at the year-month level. The complete dataset then has 20 features.

**2.3 Preparing data**

One problem with this dataset is that only 12% of the people have labeled success for subscribing the term deposit. As the goal of the model is to improve marketing performance, thus we want to improve the identification of the rare class as opposed to achieving the overall accuracy. Therefore, bagging and boosting techniques would be used to address this issue.

The previous study on this dataset (Moro et al. 2014) divided the record into training and test set by year. In our case, we would use the more traditional way to split the 2 sets. As proposed, we consider the economic condition has a rather significant impact on the model target, and those economic indicators are varying by time. Therefore, use a specific year as test set contradicts with our assumption, and we would instead randomly select 20% of the total data as the test set.

**2.4 Data Manipulation**

**2.4.1 Data Cleaning**

The dataset is about direct campaigns of a Portuguese banking institution based on phone calls. The original dataset has missing values in several features. First, there are missing values labeled as ‘unknown’ in the ‘job’ column. Empirically and psychologically, people tend to be more reluctant to disclose their job status if they are unemployed. So imputing the missing values with ‘unemployed’ makes sense.

For the same reason, ‘primary’ is chosen as the imputation for missing values in the ‘education’ column. However, the situation is quite different for column ‘poutcome,’ which is the outcome of a previous marketing campaign. We cannot determine the outcome without any sound evidence, so we change the ‘unknown’ value, which represents missing values, to ‘other.’ We stick to the truth and regard it as missing values but change its name to ‘other’ to be distinguishable from missing values. Finally, we transform the binary columns such as ‘default,’ ‘housing’ and ‘loan’ from string labels (‘yes’ or ‘no’) to numeric labels (‘1’ or ‘0’).

**2.4.2 Feature Engineering**

The most significant and core value-added idea to our model is that instead of building a predictive model with the variables in the original dataset, we add some macroeconomic variables to improve the accuracy of the classification and make the model more interpretable. Macroeconomic variables refer to the variables that reflect the ongoing situation and trend of the domestic economy in Portugal, such as the import and export of goods, unemployment rate.

The timeline of our original dataset starts from May 2008 to November 2010. We collected 19 macroeconomic variables by month from an official Portuguese economy website and joined the macro dataset with the raw dataset by month. Figure 14 is the pair correlation of the resulting dataset.

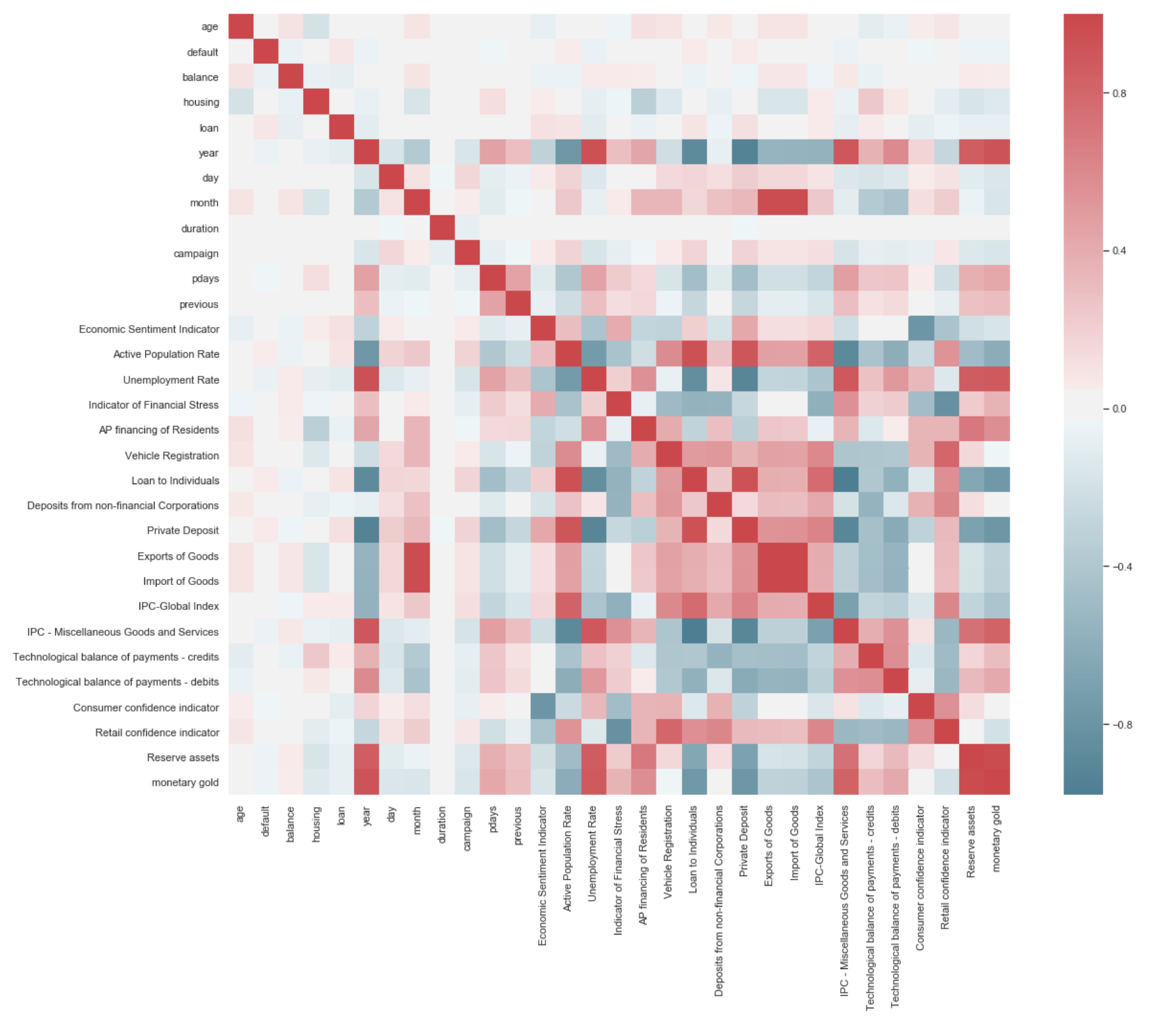


Figure. Pair correlation of the dataset with macroeconomic variables added

The figure above shows that the lower right part has more dark colors than that in other areas, which implies the high correlation between the new-added macroeconomic variables. For example, the unemployment rate is highly inversely proportional to private deposit, which makes sense in reality.

Due to the high correlation, we apply a f\_classif function to all the added macroeconomic variables and keep those with the highest score. f\_classif is a function for the classification task. For each one of the variables, we applied one way ANOVA F-test and those features with high scores means. The means of the groups are not all equal. Finally, five of the added variables are kept and added to the raw dataset. Figure 15 shows the correlation of the selected five macroeconomic variables and features in the raw dataset.

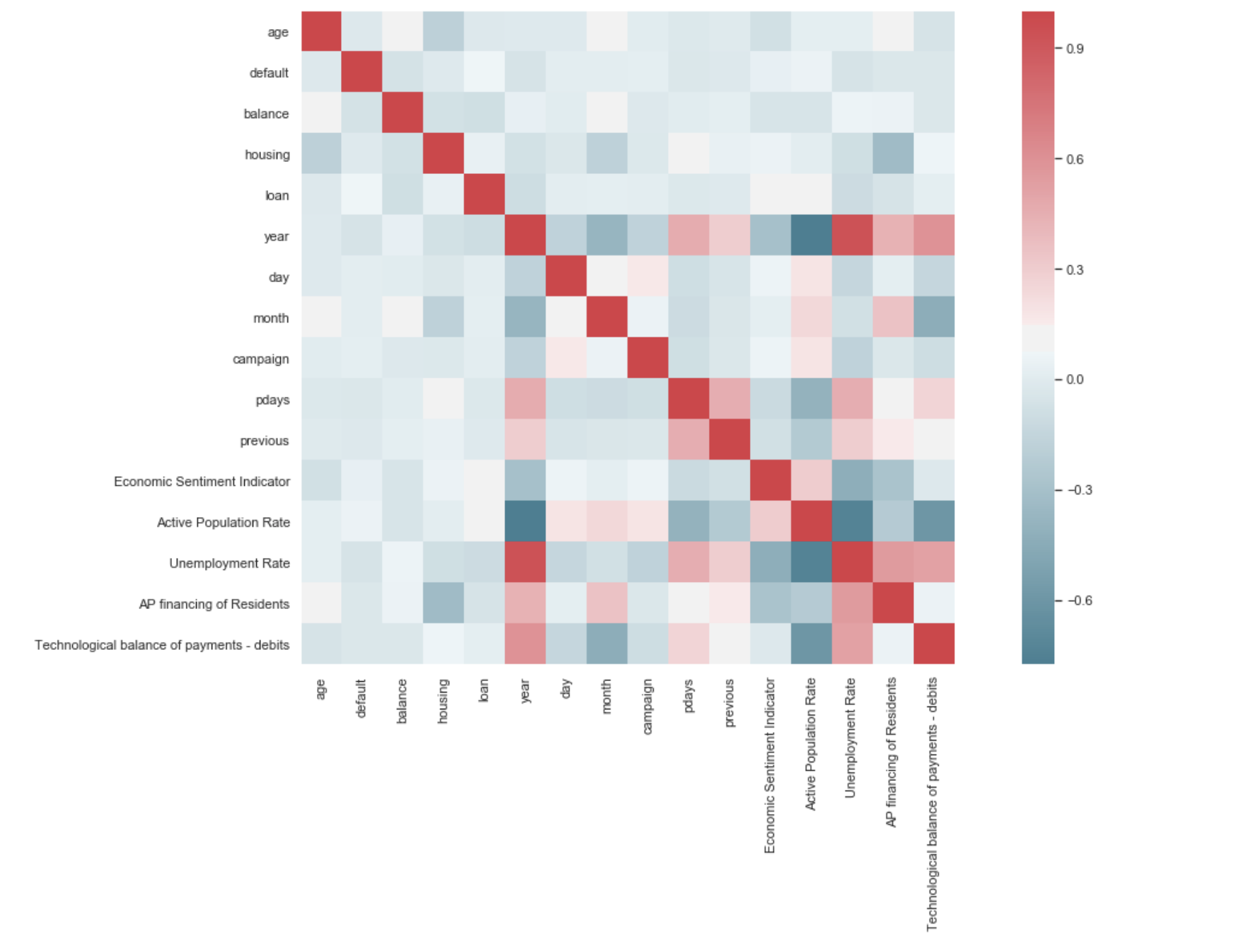


Figure. Pair correlation between final dataset variables

We can see from the figure above that the correlation between variables significantly decreased. The final five selected macroeconomic variables are Economic Sentiment Indicator, Active Population Rate, Unemployment Rate, AP financing of Residents, and Technology Balance of Payment - Debits. The five macroeconomic variables are much less correlated with each other and with the variables in raw dataset. What should be noticed is that the variable ‘year’ is highly correlated with the five selected macroeconomic variables, but it is understandable if we take their practical meaning into account. Finally, We transform the categorical variables to dummy variables with one hot encoder because empirically for tree models, dummy variables have a positive impact on the prediction metric. The variable ‘during’ is also ignored when building the model for its unique representation and effect on the final classification.

**3. Model Understanding**

**3.1 Model Selection**

To reiterate our goal - predict the term deposit using machine learning techniques, which means we are trying to forecast whether or not one client would subscribe to the bank deposit. It makes the problem become modeling a binary dependent variable, and thus, classification models should be considered.

When applying predictive modeling techniques to business, it’s crucial to accentuate the interpretability of the results. Black-box is undesired in business as it does not validate the results and lacks proof of concept. To avoid the issue, the paper uses multiple tree methods instead of deep learning neural network approach, which is popular and in demand lately. Moreover, deep learning neural networks require large training datasets and much computing resources. Given that our tree-based models already achieved relatively good performances, as discussed below, we do not consider the neural network as our top choices.

**3.1.1 Logistic regression**

Our baseline model is Logistic Regression, a widely used binary classification model which uses log-odd to predict the probability of each class. Data scientist usually uses Logistic Regression as a baseline model due to the simplicity of its algorithm and the accuracy it maintains. Logistic Regression serves as our line model due to its drawbacks - it is vulnerable to overfitting. The model can appear to have more predictive power than they do due to sampling bias. In reality, it is often the case that far more people do not subscribe to one specified deposit than those who do or the other way around. Logistic Regression model does not handle unbalanced datasets well. Realizing this issue, we attempt to apply tree-based algorithms. By incorporating bagging and boosting methods, these algorithms add randomness and help address issues such as data imbalance. They also achieve better performance and interpretability.

Furthermore, by taking considerations of macroeconomic measurements, we hope to improve our models and provide more business value-added to traditional banking problems. We are going to discuss Models’ performance metrics and evaluation criteria in detail in the following sections.

**3.1.2 Random Forest**

Random Forest is an ensemble method that comprises of the idea of “bagging” and random selection of features. It corrects for one decision tree’s habit of overfitting to training sets and improves stability and accuracy. Besides, it runs efficiently on large databases and generates an internal unbiased estimate of the generalization error as the forest building progresses. The Random Forest is also natural to interpret compared to other machine learning algorithms due to its tree structures, which helps data scientists tune other hyperparameters for model improvements.

One other significant advantage of RF is that it gives estimates of what variables are important in the classification. By obtaining the weights of importance of each feature in the model, we can tell which variable/variables contribute the most to prediction.

**3.1.3 XGBoost**

Compare to Random Forest, XGBoost utilizes gradients and performs optimization in function space rather than in parameter space. XGBoost is also a tree-based algorithm that builds trees one at a time, where each new tree helps to correct errors made by a previously trained tree. It starts with a weak classifier and adds weights to learn misclassified labels. In this way, boosting algorithms can obtain significant improvement building upon simple bagging techniques.

The model gives an excellent strategy to deal with unbalanced datasets by strengthening the impact of the positive class. XGBoost can be interpreted as one business value-add in this case - use the idea of boosting to learn our clients. Our clients are diverse, and using limited features to capture their decisions is challenging. Thus, one approach is to learn the ones that are easy to predict, and then focus on the misclassified ones to improve prediction.

Despite that XGBoost requires iterations to perform well, as most of the tree algorithms do, it does not add complexity to the model. It takes regularization so that the model is not overly complicated and lowers variance to avoid overfitting. In this project of a bank telemarketing campaign, we would like to use one model that has both excellent performance and less complexity.

**3.1.4 LightGBM**

LightGBM splits the tree leaf wise with the best fit whereas other boosting algorithms split the tree depth wise or level wise rather than leaf-wise. Consequently, when growing on the same leaf in Light GBM, the leaf-wise algorithm can reduce more loss than the level-wise algorithm and hence results in much better accuracy which can rarely be achieved by any of the existing boosting algorithms.

In terms of scalability, Light GBM works faster than Random Forest and XGBoost and demands low memory usage. It is because Light GMB supports parallel learning and handles categorical features by taking the input of feature names. It does not convert to one-hot coding. On the other hand, XGBoost cannot handle categorical features by itself; it only accepts numerical values similar to Random Forest. Therefore, one has to perform various encodings such as label encoding, mean encoding, or one-hot encoding before supplying categorical data to models.

In our data set, there are categorical features, such as job and education. Encoding these categorical features not only increases the feature dimensions and demands more memory usage but also slows down the training and testing procedure. The term deposit data we acquired was a sample provided by a Portuguese bank; if we want to scale it up into production, we have to analyze the required computing resources.

**3.3 Parameter Tuning**

Machine learning encompasses many experimentations, for instance, the tuning of “internal knobs” of a learning algorithm, hyperparameters. Especially for tree-based machine learning algorithms, there are many hyperparameters tuning involved. In the pursuit of a better performance, grid search cross-validation is exploited to search for the hyperparameter that achieves the best accuracy score. It is not easy to compare the performance of different algorithms by randomly setting the hyperparameters, because one may outperform the other with a different set of parameters. If the parameters are changed, the algorithm may perform worse. To solve this problem, we adopt Grid search CV automatically finds the best parameters stepwise instead of randomly select values of parameters.

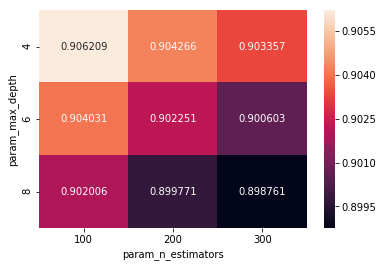


Figure. Grid Search Cross-Validation

This paper tries out different combinations of maximum depth, the maximum number of features, the minimum number of samples leaf, the minimum number of samples split, and the total number of estimators/trees. As a result, the best hyperparameter choice achieves a better accuracy result than the worst hyperparameter choice by 0.04.

**3.4 feature importance**

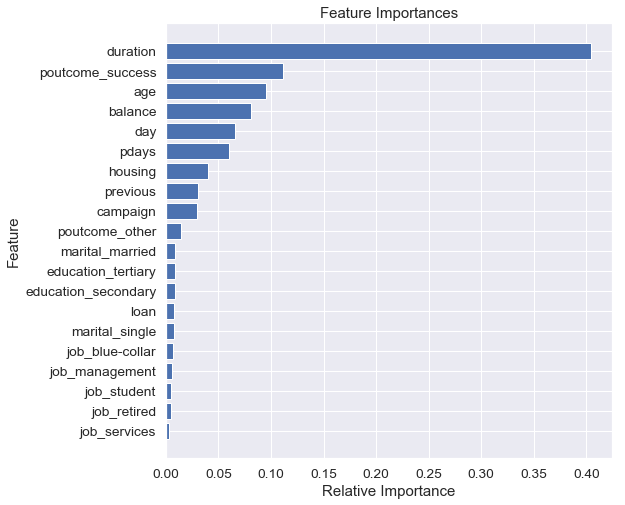


Figure. Feature Importance Ranking by LightBGM

Random Forest, XGboost, and LightBGM allow us to determine which features affect the model by ranking their importance. As shown, “duration” is dominant in this algorithm; the following are “poutcome\_success,” “age,” “balance,” This ranking intuitively informs us of what influences our clients’ choices: “duration” is dominant; age and bank account balance are crucial in the prediction as well. As a result, we decide to drop one feature of “duration” because it is too dominant according to the importance chart. Another reason to exclude this feature is that it is hard to obtain before calling the clients, which imposes difficulties in telemarketing improvements. Therefore, feature importance is valuable in the sense that managers and key stakeholders would be able to interpret the results and design strategies correspondingly without much knowledge of the model itself.

**3.5 Descriptive Statistics**

We explore the descriptive statistics of the data column by column. The main purpose of our descriptive statistics is to explore the distribution and range of each feature and lay the foundation for later analysis and interpretation. For continues numeric features, we explore the mean, standard deviation, range, and percentile. Also, we make up bins to plot the histogram to dig deeper into the distribution of the variable. For discrete numeric features, instead of plotting histogram, we drawbar plots to indicate the concentration of one or several particular values. For categorical features, we explore whether it is balanced overall its categories.

**age:**

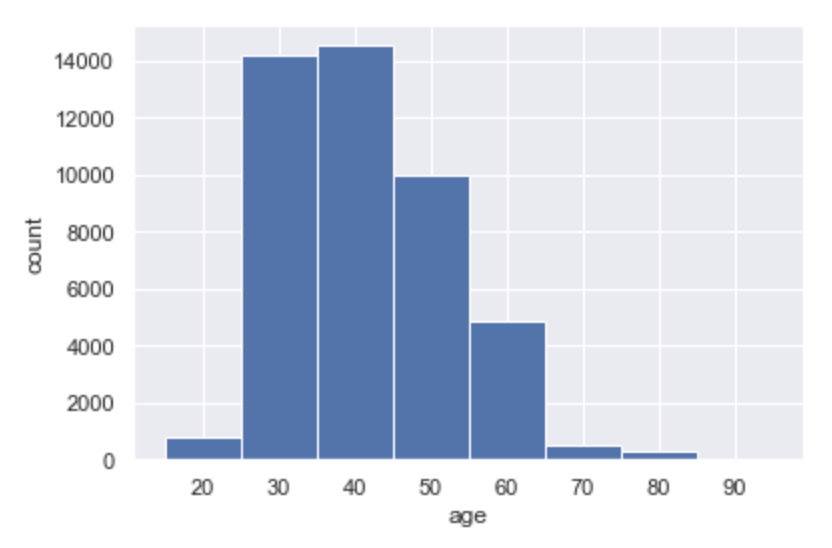


Figure. Histogram of age

Statistics of variable ‘age’ tells that the values of ‘age’ range from 18 to 95. For better analysis, we create bins of length 10 and finally plot the histogram. In the figure, the customers of the bank campaigns concentrate on the middle-aged from 30 to 60, which are all up to 10000. The distribution is approximately normally distributed with std 10 and mean 41.

**job:**

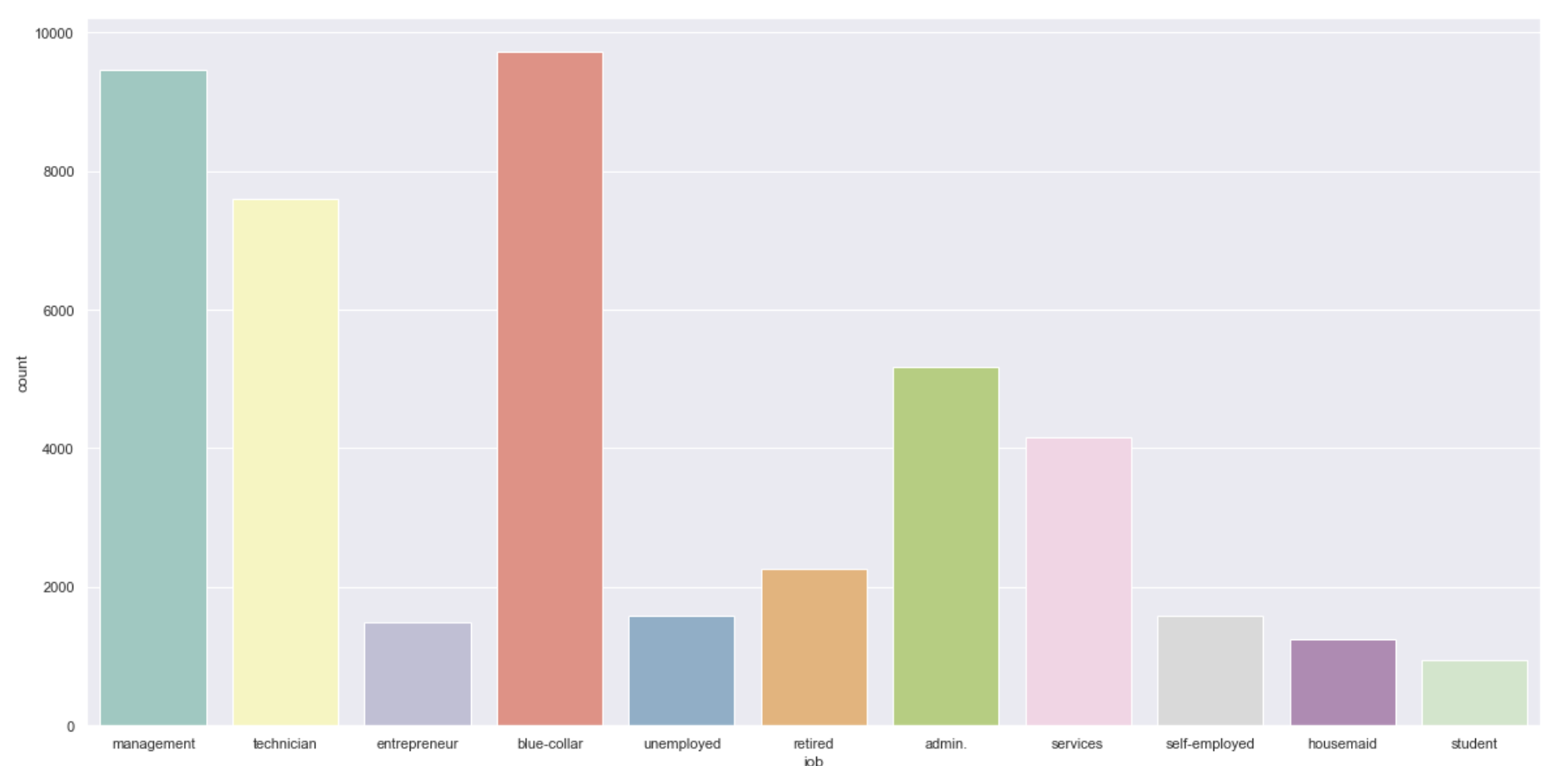


Figure. Count plot of job types

The figure above is the count plot of ‘job’ variable. The data covers nearly all of the job types appeared, considering we include more than 10 types of jobs. Most of the customers are blue-collar, or in the management position or technical field.

**marital:**

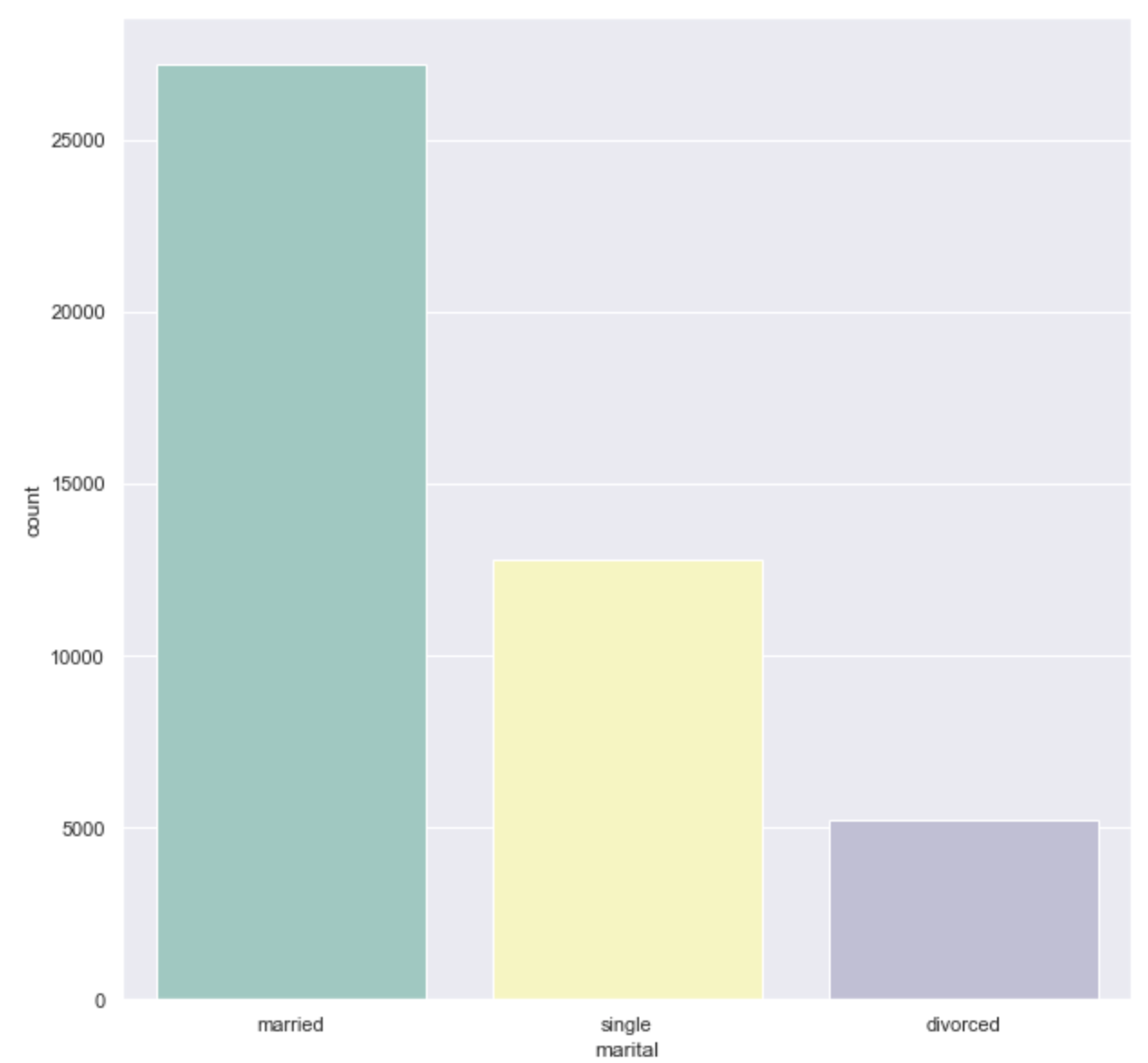


Figure. Count plot of marital

The figure above is the count plot of ‘marital’ which indicates that over half of the customers are married and there is enough data on all three categories in terms of marital.

**education:**

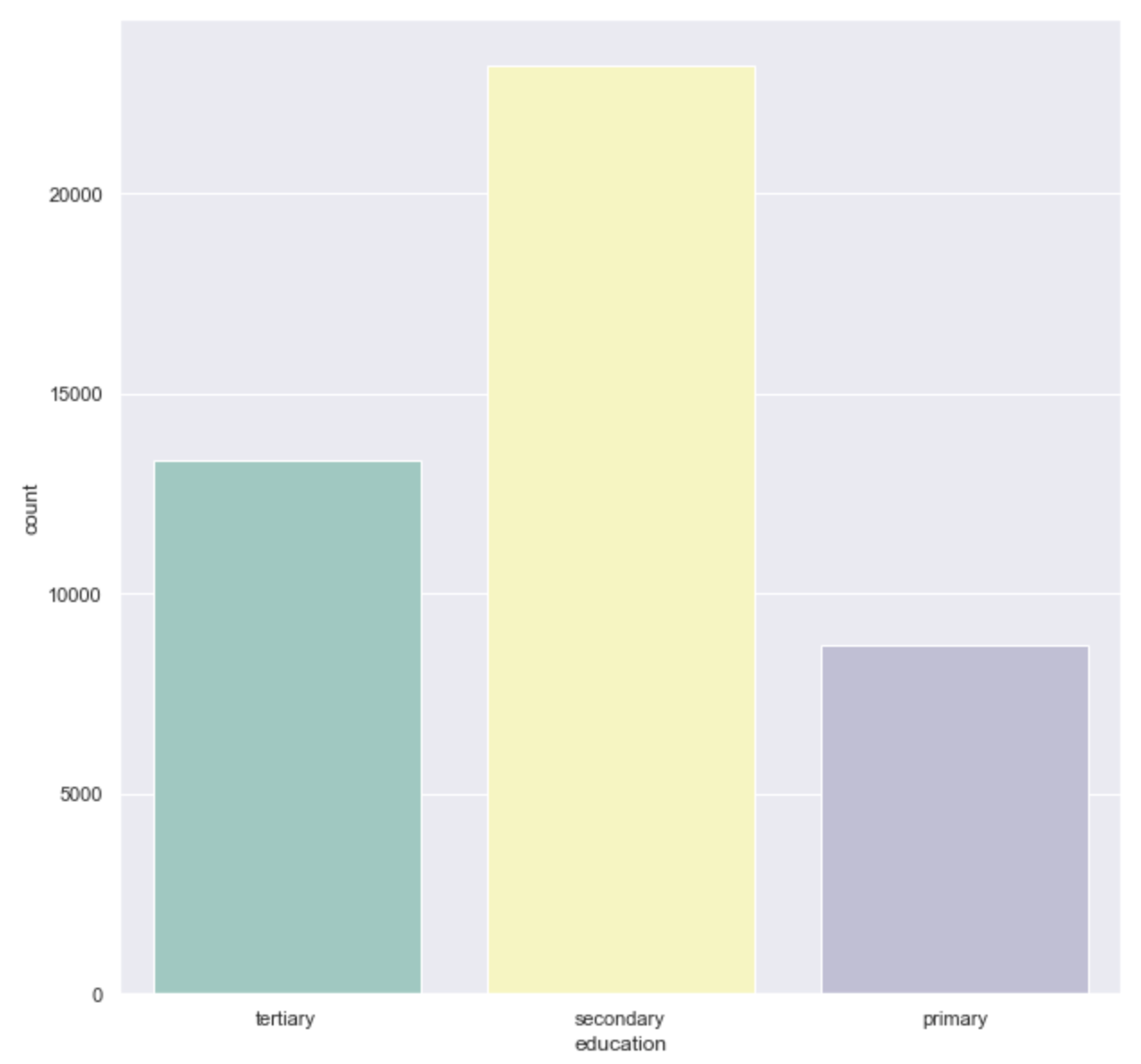


Figure. Count plot of education

The situation and distribution of education status are similar to that of marital. The figure above is the count plot of variable education. Customers with secondary education account for the majority.

**default:**

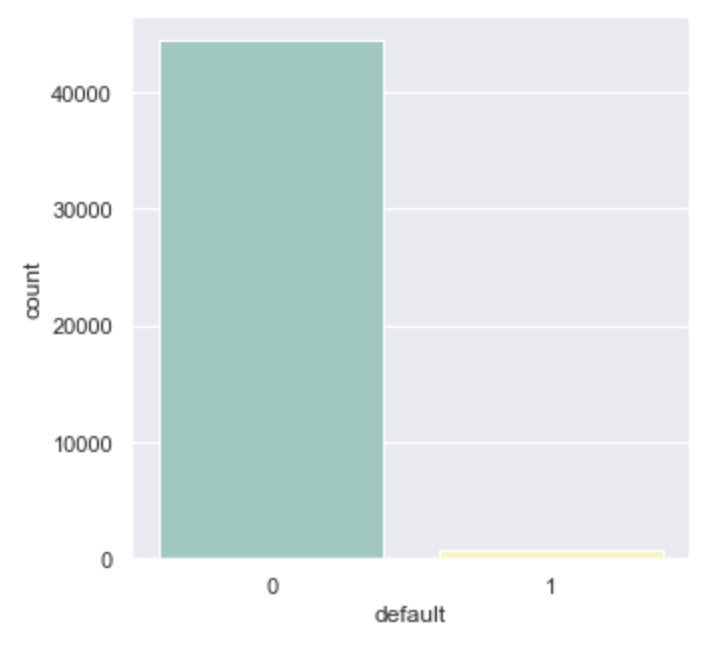


Figure. Count plot of default

The variable ‘default’ indicates if the customer has credit in default, with ‘1’ representing yes and ‘0’ representing no. The figure above is the count plot of variable ‘default.’ The data is quite imbalanced in terms of ‘default.’ This makes sense because the majority of customers would fulfill their financial obligations in reality.

**balance:**

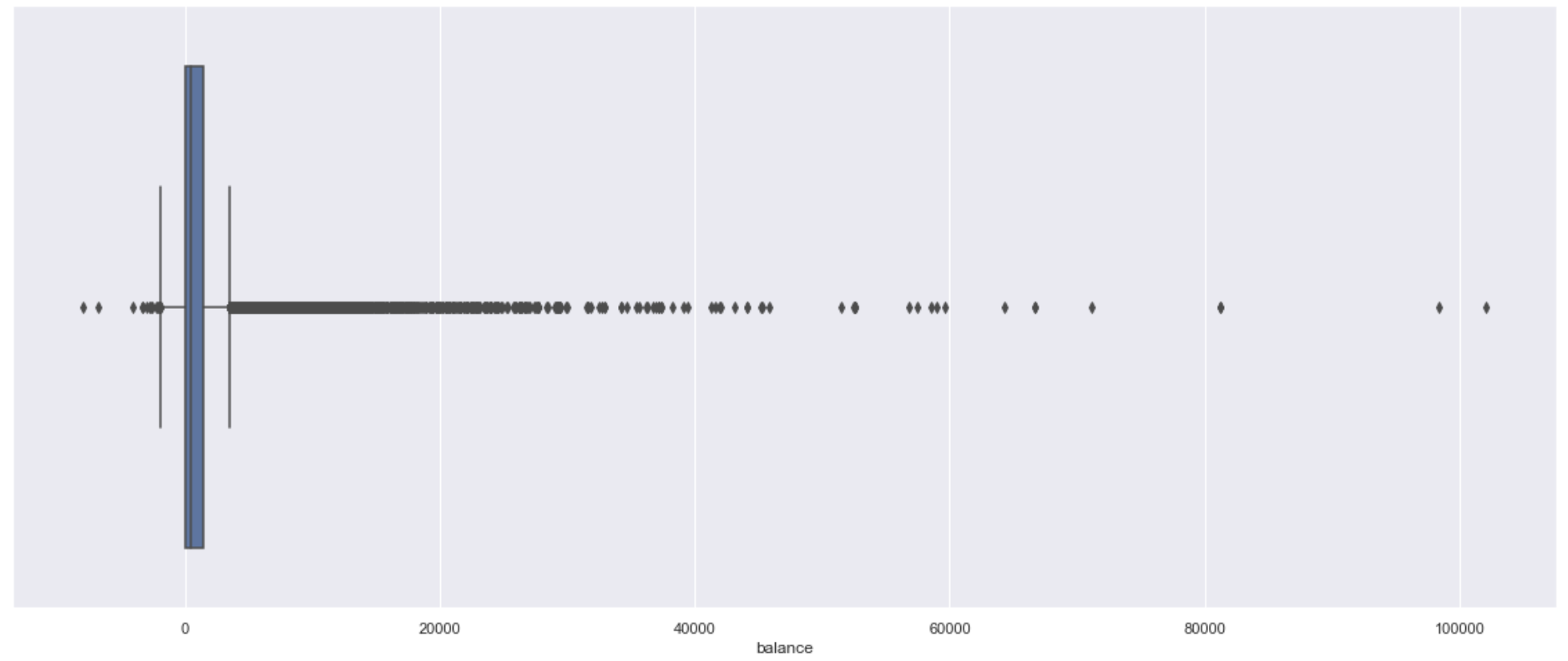


Figure. Boxplot of balance

The values of the variable ‘balance’ range from -8019 to 102127. The difference of median (50% percentile) and mean is significant considering the range, which means that the values of variable ‘balance’ are quite imbalanced. The figure above is the boxplot of ‘balance’ and proves the conclusion. There are many outliers which has a much higher balance than the rest. The majority of balance concentrates in the range from 0 to 1500. The empirically negative or positive balance could have a significant impact on the final decision of whether the customer subscribes a term deposit. This variable could be a significant one.

**housing:**

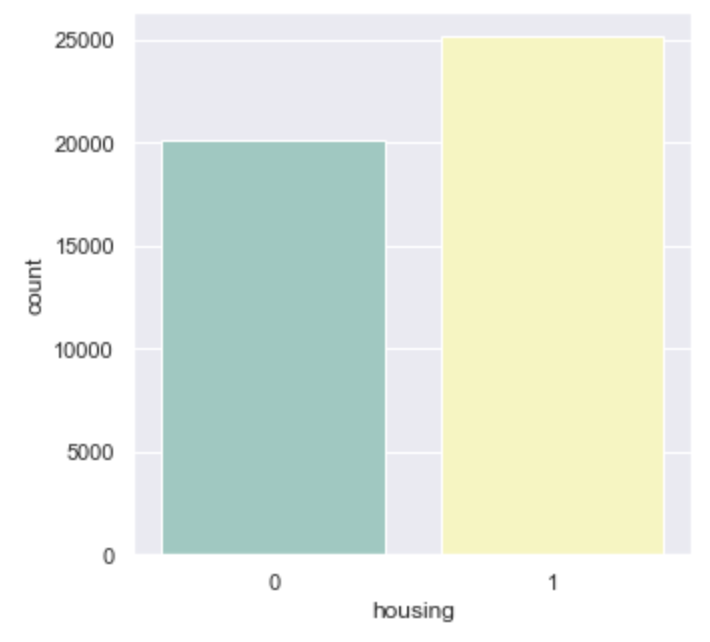


Figure. Count plot of housing

The figure above is the count plot of ‘housing.’ Variable ‘housing’ represents whether the customer has a housing loan. ‘1’ represents yes, and ‘0’ represents no. The figure indicates that the data is balanced in terms of variable ‘housing.’ This variable is empirically vital in that for customers with housing loan; they are less likely to prefer a term deposit unless they have enough balance. So the combination of this variable and balance may reveal something meaningful.

**loan:**

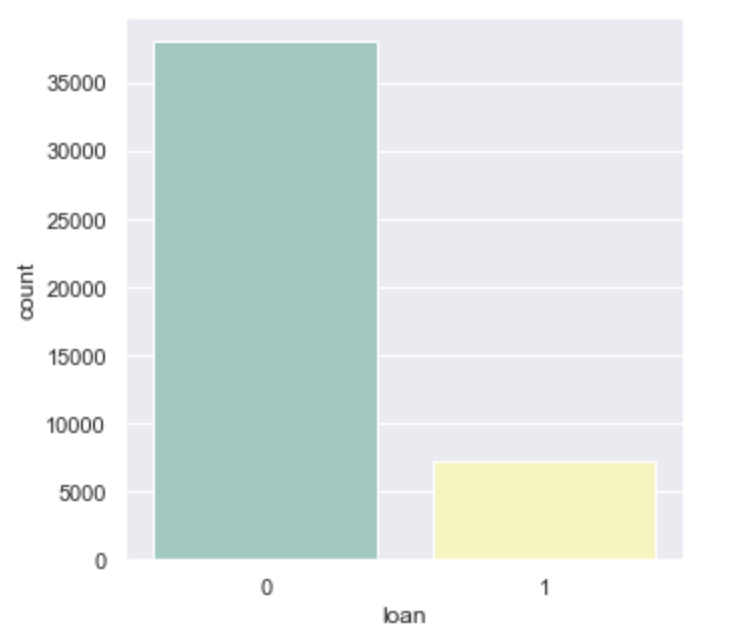


Figure. Count plot of loan

The ‘loan’ variable is a symbol of whether the customer has a personal loan with ‘1’ representing ‘yes’ and ‘0’ representing ‘no.’ The ‘loan’ variable is similar to the ‘housing’ variable, and thus, it is also determinant to the final result, especially when the customer does not possess enough balance.

**campaign:**

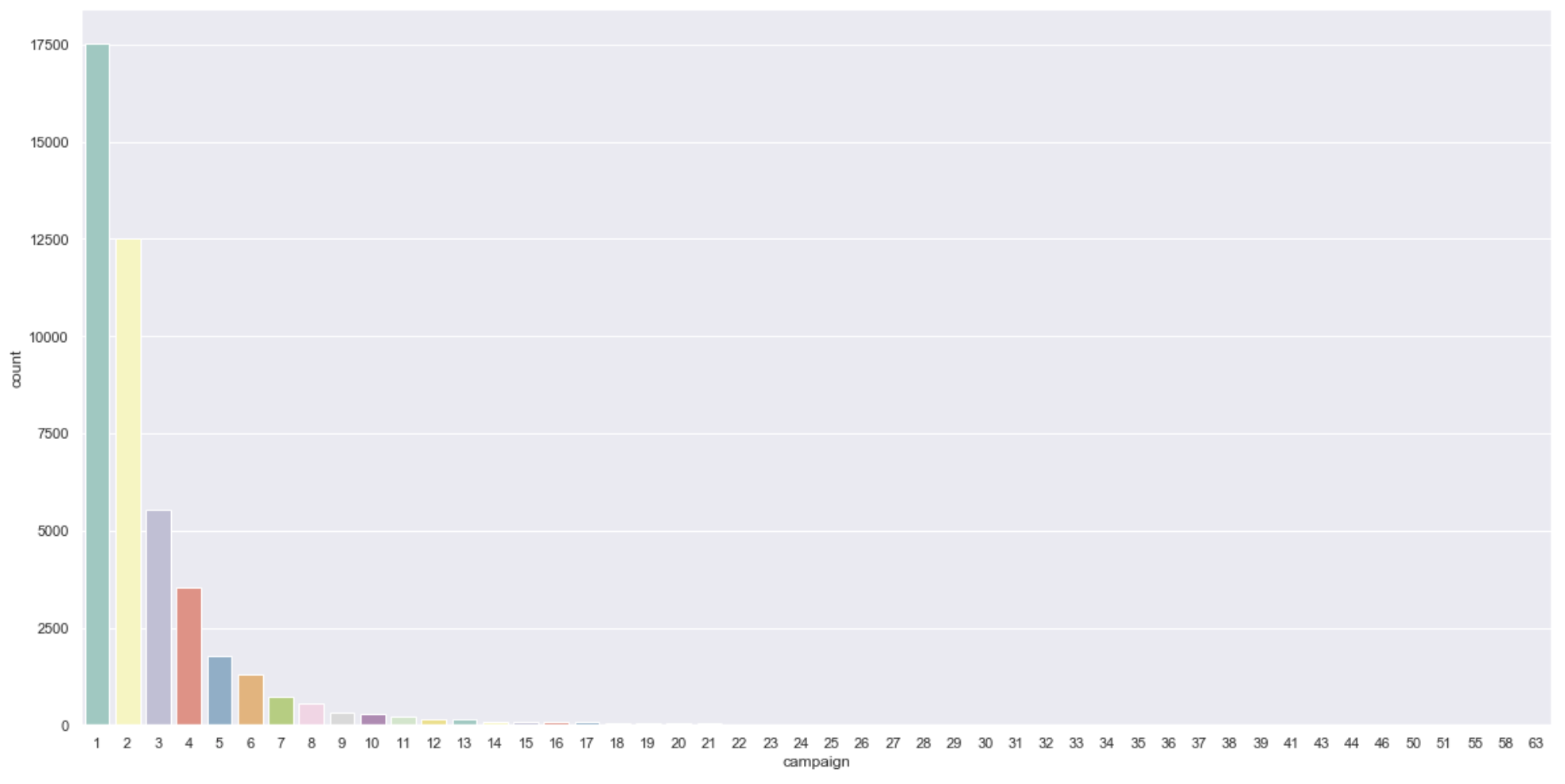


Figure. Count plot of campaign

The ‘campaign’ variable represents the number of contacts performed during this campaign and for this client. The number can be up to 63 but it rarely transcend 9. Most observations concentrate in range 1 to 5. The count plot above could testify the conclusion.

**pdays:**

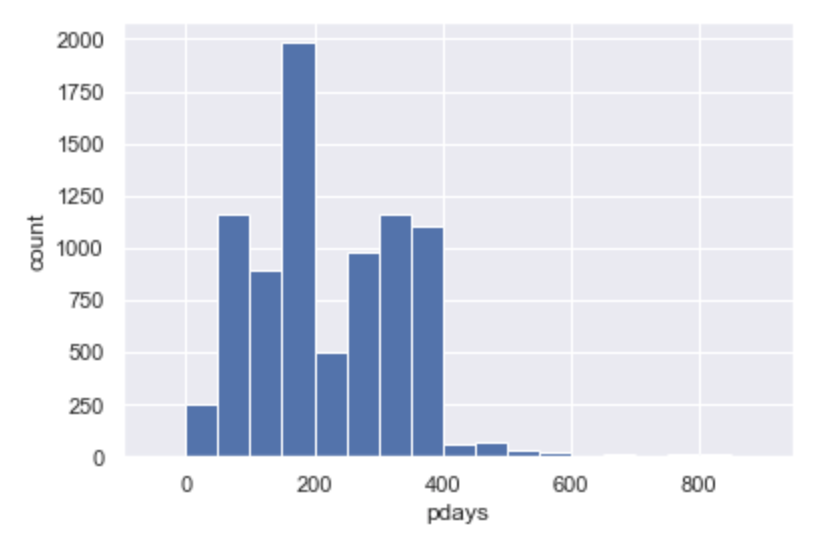


Figure. Histogram of pdays

Variable ‘pdays’ represents the number of days passed by after the client was last contacted from a previous campaign. And value of ‘-1’ means the client has never been contacted before. The dataset indicates that over 80 percent of the observations are clients that has never been contacted before, meaning the bank is digging new client resources which is good. But empirically if a client is contacted before, he or she will be more likely to accept a telemarketing campaign. The figure above shows the histogram of the pdays values with ‘-1’ value eliminated.

**previous:**

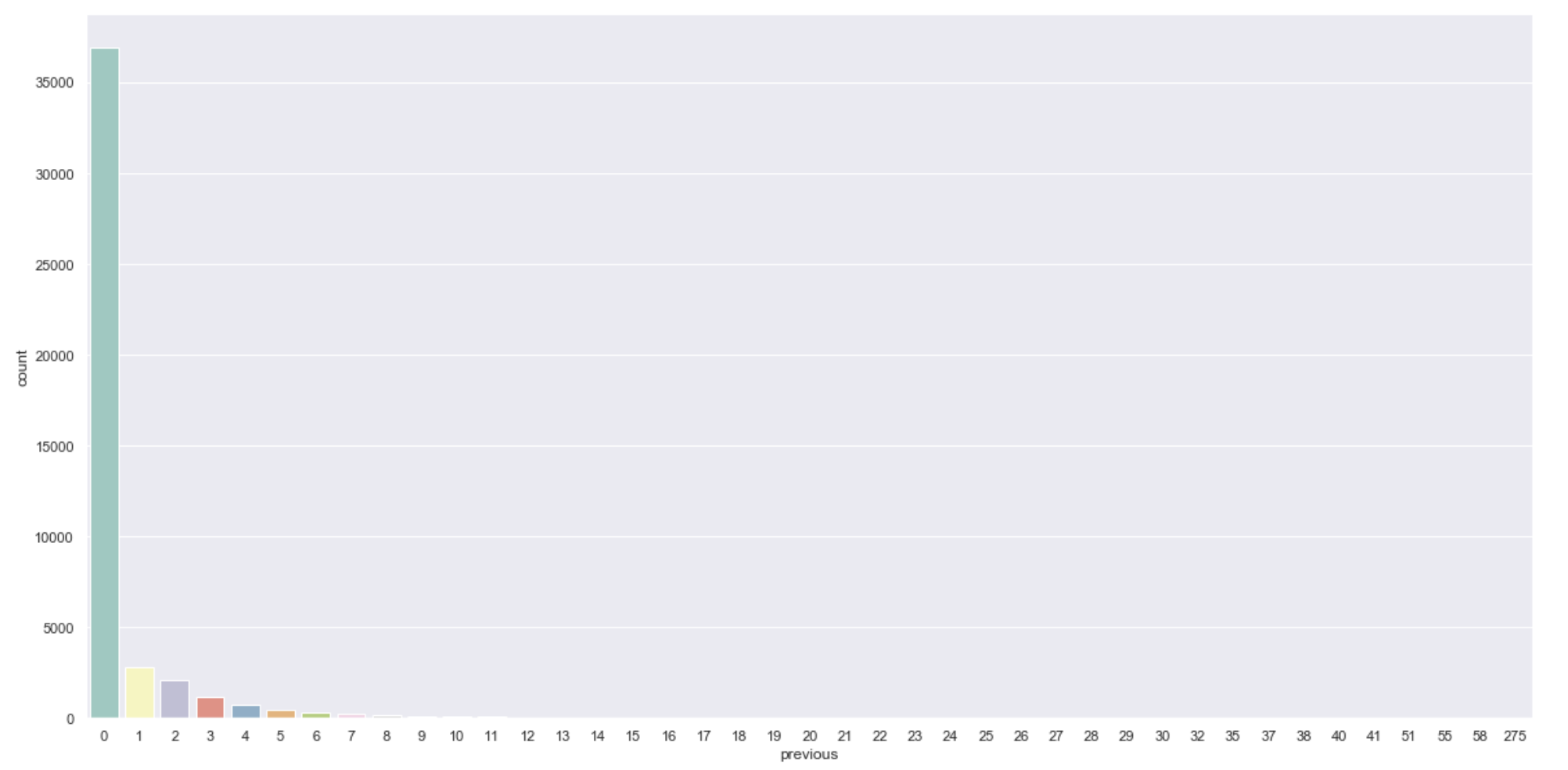


Figure. Count plot of previous

Variable ‘previous’ means the number of contacts performed before this campaign and for this client. Similar to variable ‘campaign’, ‘previous’ also has a heavy tail on small numbers, with number of value ‘0’ being the peak and much more than that of other values.

**poutcome:**

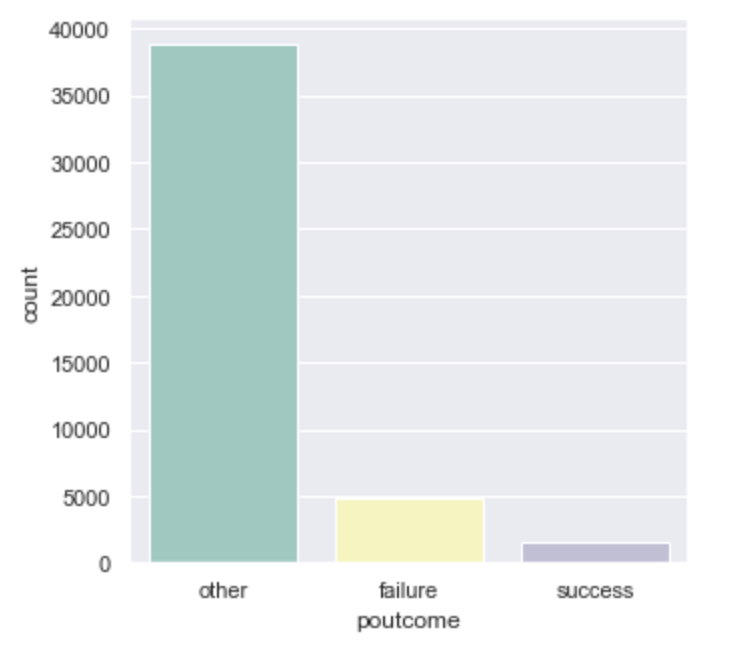


Figure. Count plot of poutcome

Variable ‘poutcome’ represents the outcome of the previous campaign. Here ‘other’ means the previous campaign does not exist and accounts for a majority of all the observations. The number of failure is multiple times of counterpart of success, which is reasonable in reality.

**duration:**

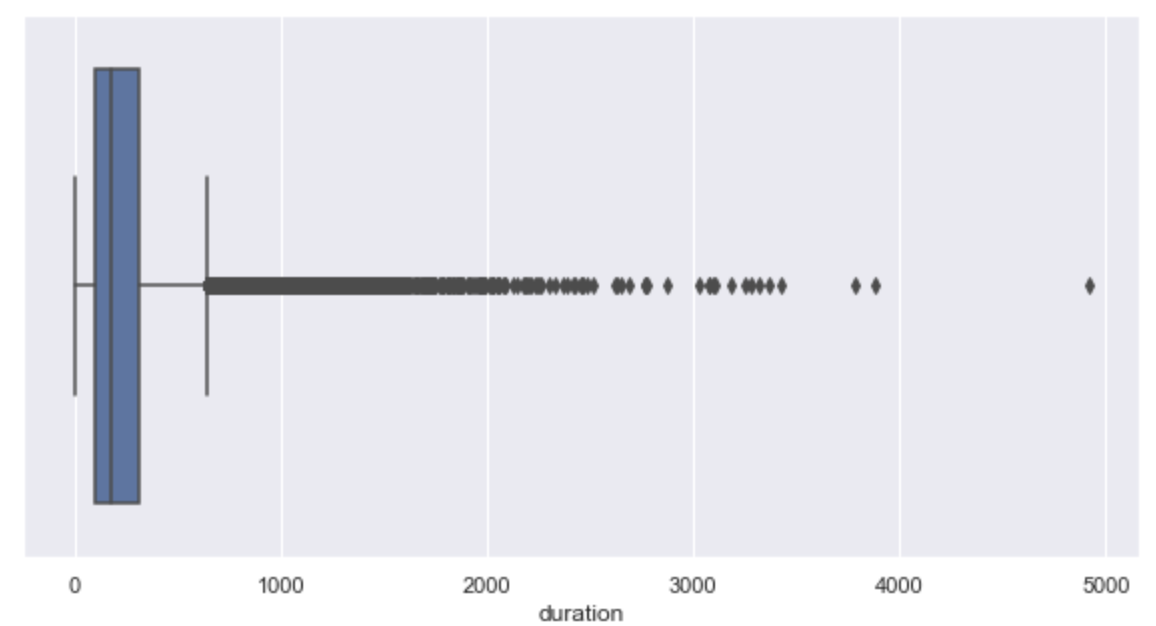


Figure. Boxplot of duration

The variable ‘duration’ represents last contact duration in seconds. It is a unique variable in that if the value of this variable equals to 0, then the output should be 0 correspondingly. It is easy to comprehend that if the duration is 0, which means the phone connection is a failure, then it is impossible for the client to subscribe the term deposit. As a result, the variable should only be included for benchmark purposes and should be eliminated for a predictive model, which is exactly what we do.

**3.3 Bagging and Boosting**

The random forest model would implement the bagging and boosting methods which can help address our issues with the data-set such as data imbalance and noise, bias and variance for some features in the data, and improve the stability and accuracy for the model.

Bagging helps us to generate additional training data by sampling with replacement. In our study, since each data point in the original data-set has the same probability to be selected in the training set. Thus we can have a better understanding of the mean and variance of the data and prevent the overfitting problem of the model. However, we are not sure if we have an excellent classifier for each tree, and bagging may further degrade the performance of the performance. Moreover, each customer is unique in this marketing data-set, improper modeling can cause even higher bias, and we would thus lose the interpretability of the model.

On the other hand, the idea of boosting is that it starts with a weak classifier and make small changes in the data that can make significant improvements in the performance. The nature of our data is that we should not treat every client. Equally, some clients are rather easy to predict; for instance, they frequently communicate with the bank callers, have a decent job and a successful history with the bank. Such clients are very probable to subscribe to a term deposit, and that's not the area we need to work hard on, we should emphasize on the most challenging cases by issue more weights on the misclassified cases. The improved trees would then focus more on these problematic clients.

Besides, each tree may focus on some of the aspects of making a perfect prediction. In our case, some trees focus on the client demographic, and some tree may focus on the quality of the call, and these aspects are not equally important. The second round weighting in boosting method is an excellent fit for our study. Trees with better prediction results would be assigned higher weights, and the weighted vote can give us a more accurate result and lower bias.

**4. Performance Metrics**

**4.1 Analysis goals**

There are many statistical metrics that are available to us when we do the prediction by category. For example, confusion matrix, F score, and AUC curve are all very important metrics in the statistical world. In our situation, however, not all the metrics have the same value to our analysis. For this project, we want to analyze the impact of adding macroeconomic features and use those macroeconomic features to boost the performance of the model we finally choose. Therefore, there are two main goals the chosen metrics should achieve eventually.

The first goal is to compare and evaluate the models that we have discussed in the previous section, we choose logistic regression as the baseline classification model, and choose tree structured models (Random Forest, XGboost, LightGBM) as potential candidates. We discussed the theoretical pros and cons for each model in the model selection section, but it is better to assess them with the data to gain a practical insight of the models.

The second goal is to assess the macroeconomic features’ impact on the performance of the models. There are two processed datasets for this project, one does not have the macroeconomic features, and the other one has the macroeconomic features. We not only look at the model performance on the datasets individually, but also want to gain information on the difference of the performance. This is very important because some models are more flexible than the other ones. A model can potentially have an outstanding improvements on the performance with some key variables added, but may do not perform very well in the first place.

**4.2 Confusion Matrix - False Negative Rate**

Recall that our business goal is to accurately reach high-potential customers. Therefore, the worst thing we can think of is to lose potential customers. Our primary metric would be the one that can somehow evaluate the chance of missing calling potential customers. One natural metric is the false negative rate of the model’s prediction, since the metric directly maps to the rate of losing potential customers.

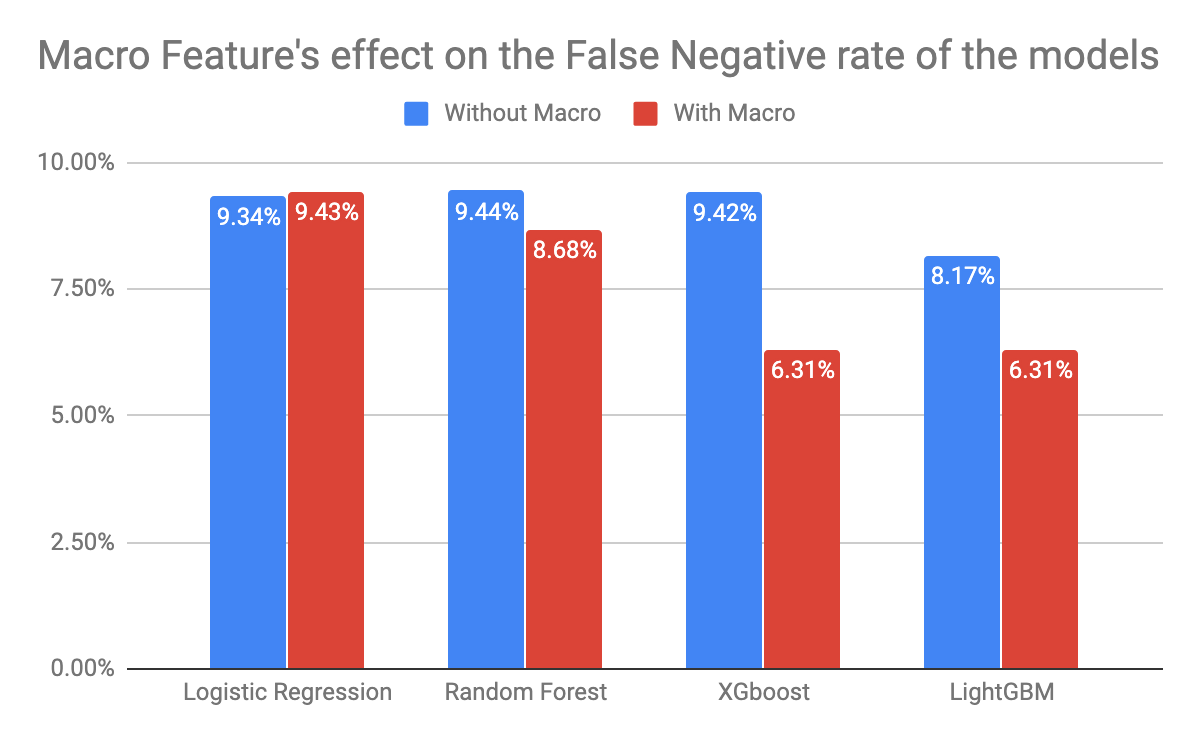


Figure. Macroeconomic Features’ effect on the False Negative rate of the Models

In our situation, since we do not want to lose chances to reach our potential customer, the false negative rate for the models are better with lower values. From the graph, we see that by adding macroeconomic features, the false negative rates decreased in all models except our logistic regression baseline. For more complicated models --- XGboost, and LightGBM, they are more capable of taking advantages of the added macroeconomic variables. Thus we can conclude that by adding effective macroeconomic models, the XGboost and LightGBM models is able to effectively reduce the false negative rate. We can conclude that if we use XGboost and LightGBM for the prediction, it is very possible to reduce the chance of losing the potential customers by adding effective macroeconomic features.

**4.3 Model Accuracy and F1 score**

The false negative rate contributes to the half of an accuracy measurement, but our previous discussion does not indicate that the other half is insignificant. Only assessing the false-negative rate does not make sense to our objective because we can build a blind model always to predict “1” and has zero false negative rates. As a result, the model would ask the company to reach as many people as possible without questioning. This will bring a considerable cost and trouble to the company if the company is trying to telemarket everyone. A big value-added part of our prediction is to effectively reduce the number of calls and therefore reduces the cost of this advertising process. By combining with false negative rate, the accuracy measurements serve as a metric to assess false positive rate as well.

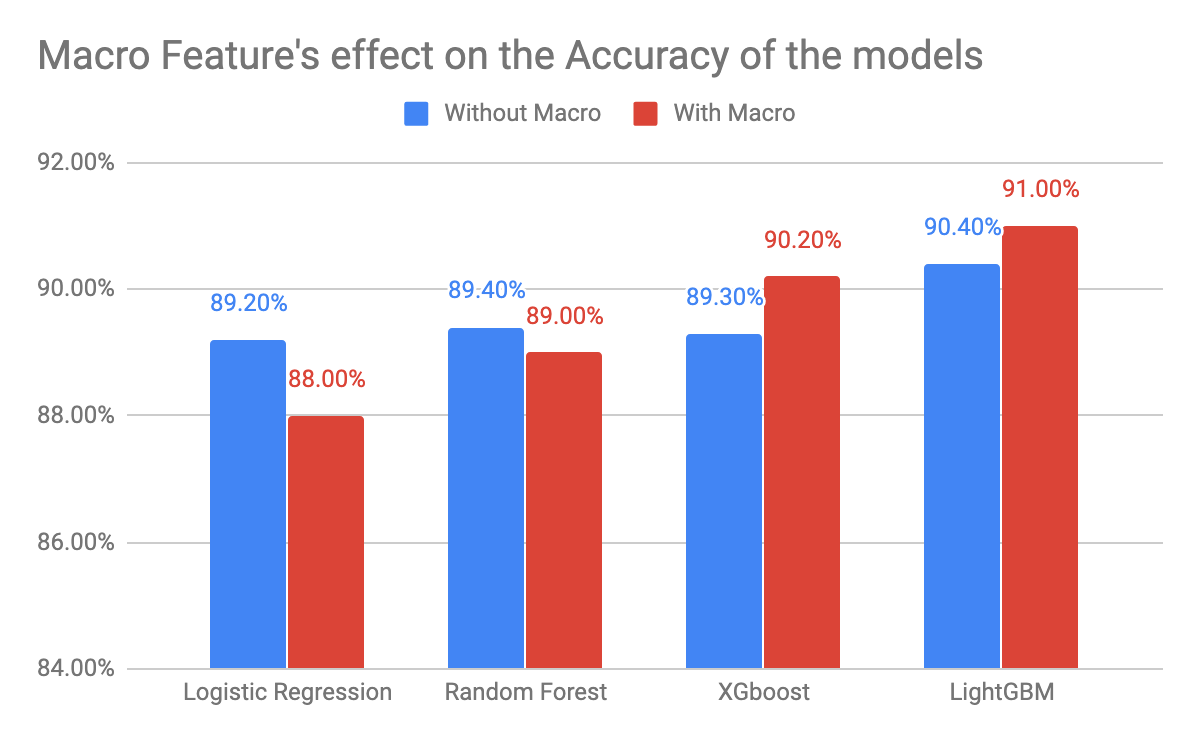


Figure. Macroeconomic Features’ effect on the overall Accuracy of the Models

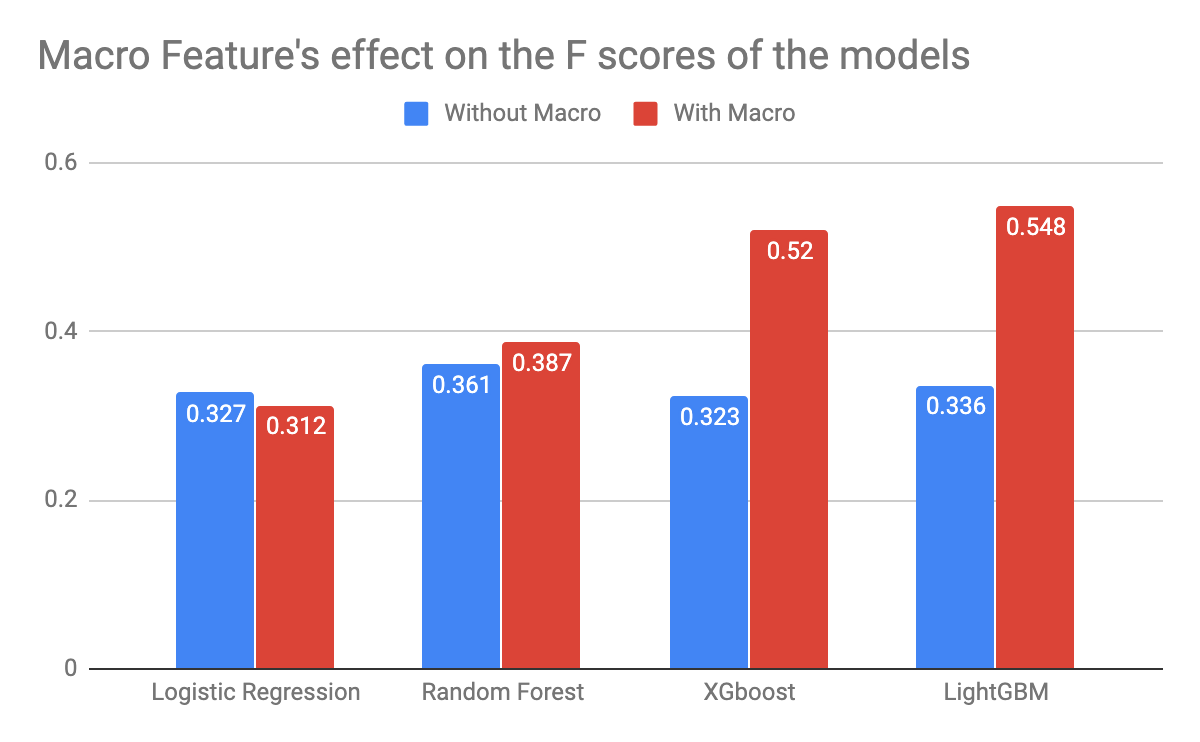


Figure. Macroeconomic Features’ effect on the F scores of the Models

As shown by the figure above, all the models that we choose outperform the Logistic Regression baseline in accuracy measurements. This is what we expected with the complexity of our data. Also, the logistic regression accuracy decreases as more macroeconomic variables are added. This is potential because the model is weak against outliers and overfitting situations.

From the graph, we also see that the added macroeconomic variables give XGboost and LightGBM models a boost in their accuracy measurements. This fact agrees with what we have discovered in the previous section, that XGboost and LightGBM are more capable of taking advantages of the added macroeconomic variables to improve their overall accuracy.

The LightGBM has the best performance in accuracy, as it's both accuracy and F scores stand out among other algorithms. A side note is that XGboost and LightGBM have lower true positive but higher accuracy. It is because, in the term deposit data, there are much more clients who subscribe to the service. LightGBM captures this imbalance, so it performs well according to the accuracy metric.

In conclusion, we find out that LightGBM performs the best in accuracy measurements with/without macroeconomic features, and the added macroeconomic features can boost its accuracy. With the outstanding performance of LightGBM, we can target our potential customers more accurately.

**4.4 ALIFT**

Lift is a way to evaluate our model as well since the success rate for our data set is low. For instance, if the truth is 1 out of 100 customers would subscribe for the term deposit, and our model predicts “not subscribe” for all the clients, the accuracy rate is 99%. However, we can not interpret the model as a near perfect model.

The idea of lift is that our model should make sure that the predicted success rate should directly proportional to the actual rate of success, and also if we predict a customer would not subscribe for the term deposit, we should be sure that this client is not likely to do so.

The life score is calculated as the predicted rate divided by the average rate of the data set. That is, for the same data-set with 13% success rate, two classifiers that give 10% and 26% successful rate would have 0.77 and 2 lift score respectively, and that indicates we would have 2.6 more times of success cases if we use the second one.

In our study, the model with Gradient boost has the life score closest to 1, such that XGBoost model gives the best performance to predict for the actual success of subscribing term deposits. Using Gradient Boosting, the bank can target potential clients with a high response rate and classify clients according to predicted probabilities with high accuracy.

Utilizing ALIFT to evaluate the model can obtain more information about the actual performance of the model due to the imbalanced nature of the data set, it can present more than just accuracy in our case.

**4.5 AUC-ROC**

Considering our goal of predicting selling of long-term deposit, evaluating the performance of a model is not trivial. Therefore, we used a variety of metrics to evaluate our models. AUC, the area under the ROC curve (receiver operating characteristic), is often preferred over accuracy for binary classification. AUC deals with situations where you have a very skewed sample distribution and aims to not overfit to a single class. It measures how true positive rate and false positive rate trade-off.

In the business setting, we do not want either high false positive or high false negative, which may be overlooked by the accuracy metric but will be captured by AUC. For example, the marketing department would not like to have potential customers to be misclassified as “0”, i.e., will not purchase. Similarly, we do not want to spend resources targeting customers that are less likely to use the service. Therefore, alongside accuracy, we include AUC as one of our performance metrics.

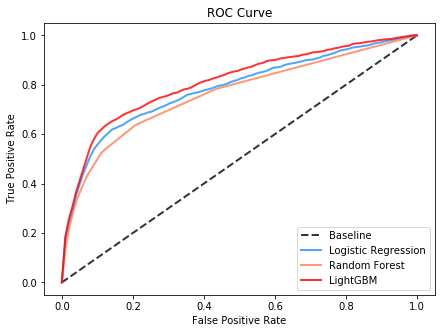


Figure. AUC-ROC Plot for Models without Macroeconomic Features

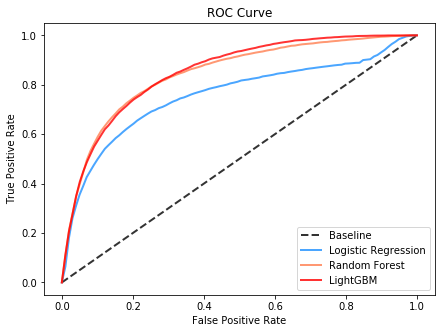


Figure. AUC-ROC Plot for Models with Macroeconomic Features

As shown above, in general, tree-based algorithms outperform Logistic Regression in terms of both AUC-ROC curves. LightGBM model achieves the best performance with/without macroeconomic variables. The added macroeconomic features give a more substantial and smoother area under the AUC-ROC curve for LightGBM. Therefore, LightGBM has achieved the best balance in dealing with the skewness in our data, and the macroeconomic features are going to make it even better. As a result of the analysis, LightGBM is the most capable model of handling our data and would be the most accurate model that leads the telemarketing team to the highest potential customers.

**5. Conclusion and Future Work**

LightGBM outperforms all the other models in all the performance metrics we choose and maintains good computing speed. LightGBM is very sensitive to the useful macroeconomic features we add and can turn the added features into its significant performance improvement. Those excellent features make LightGBM an ideal model in our predictions.

**5.1 Added Macroeconomic Feature Importance**

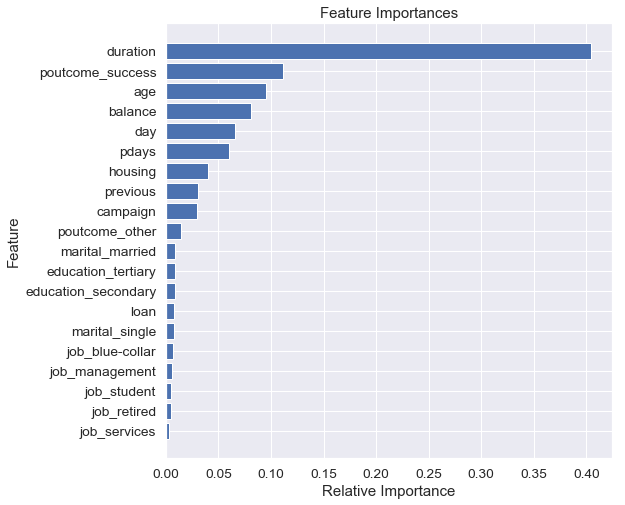


Figure. Feature Importance by LightGBM without macroeconomic variables

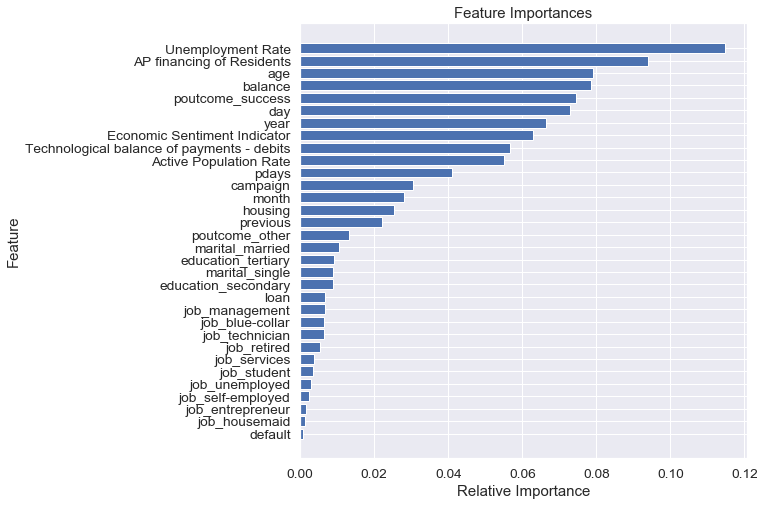


Figure. Feature Importance by LightGBM with macroeconomic variables

From the two graphs above, we can see how added macroeconomic variables are affecting the model’s decision-making process.

As shown in the figure of feature importance by LightGBM without macroeconomic variables, the variable ‘duration’ has the most substantial impact on the final classification result, which is consistent with our analysis before. Other significant features include if the previous campaign is successful, age, balance, day, and the number of days passed by after the client was last contacted from a previous campaign. These features dominate the classification process.

After we add macroeconomic variables into the LightGBM model, with variable ‘duration’ eliminated, the five added features take over the original dominant features. We can see from the figure of feature importance by LightGBM with macroeconomic variables that variable ‘Unemployment Rate’ and ‘AP financing of Residents’ rank 1st and 2nd respectively in terms of feature importance. The rest three added macroeconomic variables ‘Economic Sentiment Indicator’, ‘Technology balance of payments - debits’ and ‘Active Population Rate’ also rank in the forefront of feature importance order. Except for the added five macroeconomic variables, the relative ranking of feature importance in terms of the features in the original dataset remains approximately the same.

**5.2 Conclusion**

In this study, we propose a decision support system using machine learning approach for the selection of bank telemarketing clients. We analyze a sizeable Portuguese bank dataset, collected from 2008 to 2010, with a total of 45,211 records. The goal is to model the success of making a long-term deposit using attributes that were known before making the call. We have placed a particular emphasis on feature engineering and have confirmed that macroeconomic indicators are of great significance in optimizing targeting in telemarketing. We merge macroeconomic dataset to the client dataset such that we expand the study with a new level.

In conclusion, our hypothesis is proved, and the added macroeconomic variables do make a substantial impact on machine learning models. The improvement of the performance testified the positive effect of macroeconomic data when applied with some machine learning models. Given the model outperformed our baseline, the other analysis we conducted in this paper also provide value added into each part. The paper could also help research and business intelligence team to make better strategies in telemarketing campaigns.

**5.3 Future Work**

The goal of this model is to accurately predict the client behavior and discover the most critical impacts on successful term deposit. We improve the object further with macroeconomic data, and the final result proves that it works well. The future work is to find more hidden decisive relationships to making client subscript term deposit. To accomplish the goal, we will need to find more dimensions of information, for instance, clients transaction records, shopping behaviors, and languages used during the call. With more dimensions of information in the future, we expect a more robust and interpretable model.