Predicting the Success of Bank Telemarketing- Project 1030, Brown University

Sagarika Ramesh

https://github.com/sagarika251/Project1030.git
December 2021

1 Introduction

In the banking industry, a lot of importance is given to telemarketing as high profits are obtained with reduced costs. An intelligent decision support system(DSS) is required to efficiently help in the prediction of the client's subscription to the deposits. It is valuable information for the banking industry as they can now analyze the effectiveness of their telemarketing campaign on the clients and the type of clients they need to select.

This project attempts to create a tool to support the client selection decision based on telemarketing campaigns conducted by the banking industry. The classification goal of this model is to predict if the client will subscribe to a term deposit or not. The dataset is obtained from the UCI Machine Learning repository. It has 45211 data points and 17 features. The features it consists are: age, job, marital, education, default, balance, housing, loan, contact, day, month, duration, campaign, previously contacted days, previous campaigns, previous outcome, subscription of a client to a term deposit.

The dataset has been used by several authors for different purposes, these include bank telemarketing, credit scoring and other decision support system (DSS). In one of the publications, the authors use the dataset to predict if a client successfully subscribed to a long-term deposit. Data models: logistic regression, decision trees, neural networks and support vector machines were compared using two metrics area of the receiver operating characteristic curve (AUC) and the area of the LIFT cumulative curve (ALIFT). The best results were obtained by the neural networks, and the results were AUC of 0.80 and ALIFT 0.67[1]. They also observed that an increase in other highly relevant attributes enhances the probability of a successful client subscription. In another publication, the authors used the dataset to validate the effectiveness of credit scoring models. They have presented a comparative analysis with nine ensemble learning approaches with various classification approaches. They found that MultiBoost and Dagging with a multilayer ensemble frame is the best approach for credit score classification. The accuracy of 92.86% was achieved[2].

2 Exploratory Data Analysis

This section contains several plots that we obtained during the exploratory data analysis.

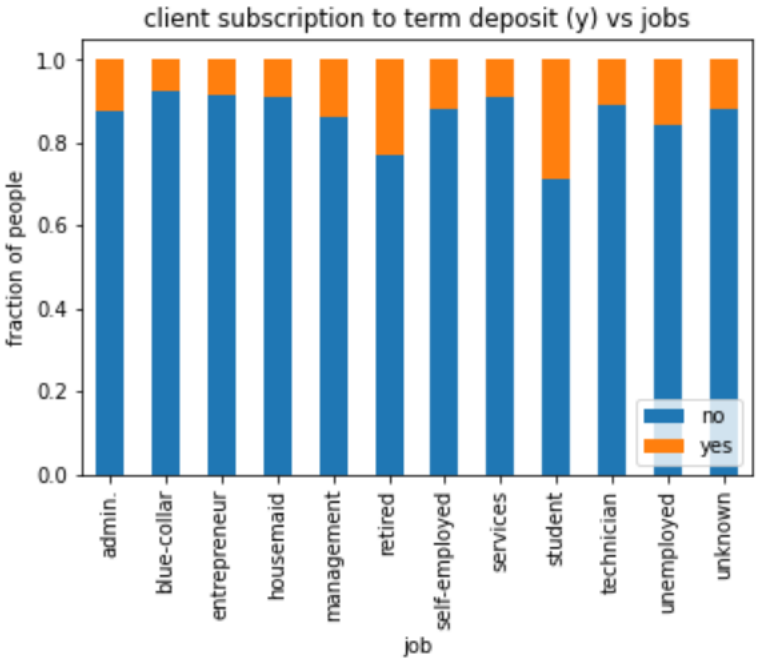

Figure 1 In the above plot of Client subscription to term deposit vs jobs we can find how the jobs are affecting the subscription to term deposit. We can find that students and retired people tend to subscribe to the term deposit. While majority of with blue-collar, entrepreneur, housemaid and services are least subscribed to the term deposit. This is valuable information to predict who will subscribe to the term deposit.

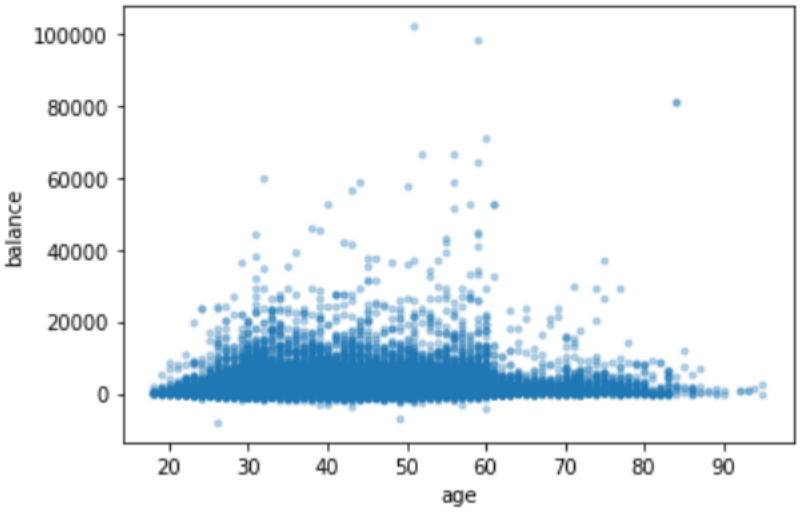

Figure 2 In the above plot we can see how the age and average yearly balance are related to each other. We can observe that the balance deceases with the increase in age and the balance is highest during middle age. The above plot provides valuable insight into how age and balance feature are related.

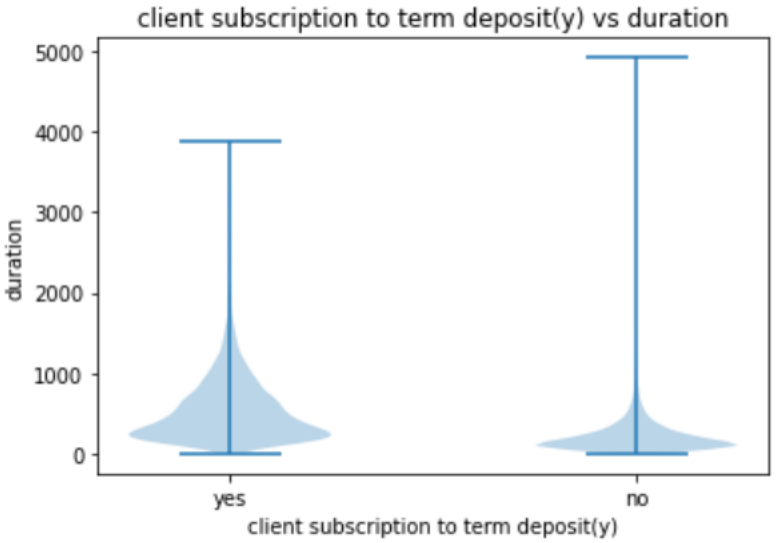

Figure 3 In the violin plot we can see how last contact duration is related to client subscription to term deposit.We can see that when the duration of contact is slightly higher than 0 seconds and less than 2000 seconds the clients tend to get subscribed to the term deposit.

3 Methods

3.1 Data Preprocessing

During the data splitting 20% of the data is allocated to testing and other 80% is sent to stratified kfold splitting as the target variable is found to be imbalanced. We do this to handle the imbalanced data set. 5 fold splitting is chosen. The Other(Used for validation) and test sets are split into 80% and 20% respectively.Random states are set in order to check reproducibility. The features are identified as categorical or continuous. Based on that we apply respective transformers. Ordinal transformers are applied to education,month, poutcome as they can be numbered in order. One hot Encoder is applied to job,marital,default,housing,loan,contact as the are categorical features. Minmax scaler is applied to age feature as it is continuous and bounded. Standard Scalar is applied to balance,day,duration,campaign,pdays,previous as they are continuous features.I have used GridSearchCV which helps in hyperparameter tuning. Then we fit it on the validation set and predict the target variable. Number of features in the preprocessed data is 34. The target variable is encoded since it has 2 categories.

3.2 Model Selection

After preprocessing and splitting 4 different models are trained and compared. The 4 models are: a Logistic Regression with L1 regularization, a Logistic Regression with L2 regularization, a Random Forest Classifier, an XGBoost Classifier. All the models are subjected to hyperparameter tuning by using GridSearchCV to find the best parameter combination for the models. For XGBoost Classifier brute-force grid search method is used to find the best parameters for the model.The models are run on 5 different random states instead of 10 because increase in random states increases the time required for the model to run. Below are the parameter that are tuned and values applied to each model.

Models	Parameters tuned
L1	C: [100.0, 35.93, 12.914, 4.64, 1.66, 0.59, 0.21, 0.07, 0.027, 0.01]
L2	C: [100.0, 35.93, 12.914, 4.64, 1.66, 0.59, 0.21, 0.07, 0.027, 0.01]
RF	max_depth: 1,3,5,7,10,25; max_features:1,3,5,7,10,25;min_samples_split:2,5,7
XGBoost	max_depth: 1,2,3,6; subsample: 0.4,0.5,0.66

Figure 4 Parameters that are tuned in each model

Best parameters are extracted by comparing the model's best f_beta scores. I have taken beta value as 1.5 because it is an imbalanced data set and it is cheap to act problem, which means it does not harm when a client

is falsely classified as subscribed to the loan. Below are the average f_beta scores for best of each model across 5 random states.

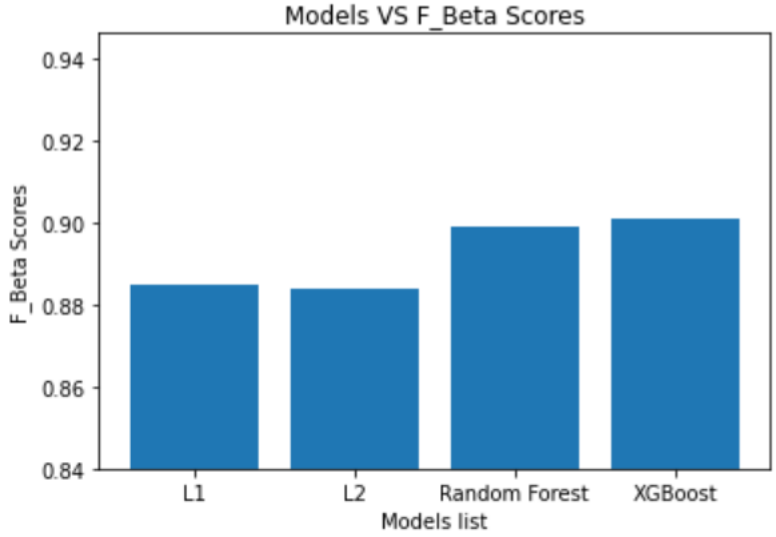

Figure 5 Average f_beta score for the best model over 5 random states

As we can see from the plot that the average performance of all the models are quite similar. XGBoost has the highest f_beta score of 0.901. The f_beta score of Random Forest is 0.89 which is very close to the highest score. I have chosen XGBoost as the best model and the parameters chosen for it are max_depth of 6 and subsample of 0.5 because the model has achieved the highest score of 0.911 on this parameter combination when run on 5 different random states.

3.3 Final Model Formulation

Now, the best Model and the hyperparameter choice are trained on new 10 random states after split. During each split 20% of data was allocated to test dataset and 80% is sent to stratified kfold splitting which splits into 60% and 20% of training and testing data respectively. For each random state the f_beta score calculated are stored are stored.

4 Results

4.1 Evaluation of Models

The baseline accuracy is found to be 0.847. The XGBoost model which is run over 10 random states has an average f_beta score of 0.904 and standard deviation of 0.002. The baseline accuracy is 28.5 standard deviations below the mean f_beta score achieved by our model.

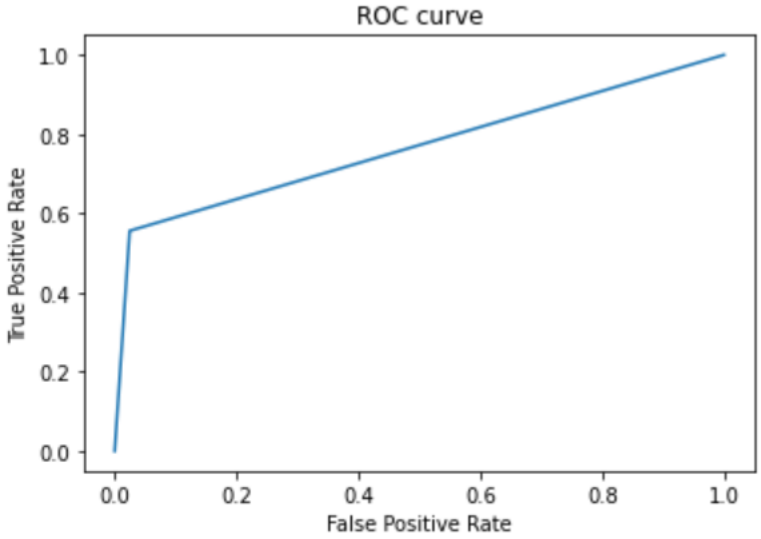

Figure 6 ROC curve of our XGBoost model

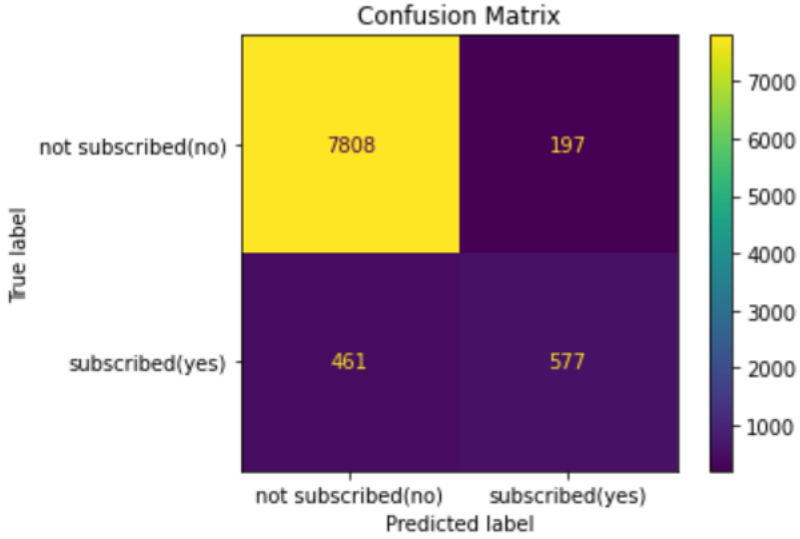

Figure 7 confusion matrix of our XGBoost model

We can see that the model has a fair AUC value and it could be improved by tuning more parameters. I was not able to tune many parameters due to long run time. From confusion matrix we can see that some clients are falsely detected as not subscribed.

4.2 Interpretation of Findings

Global feature importance is calculated on 3 different metrics: gain, cover, weight using importance_type. Below are the results of 3 different metrics.

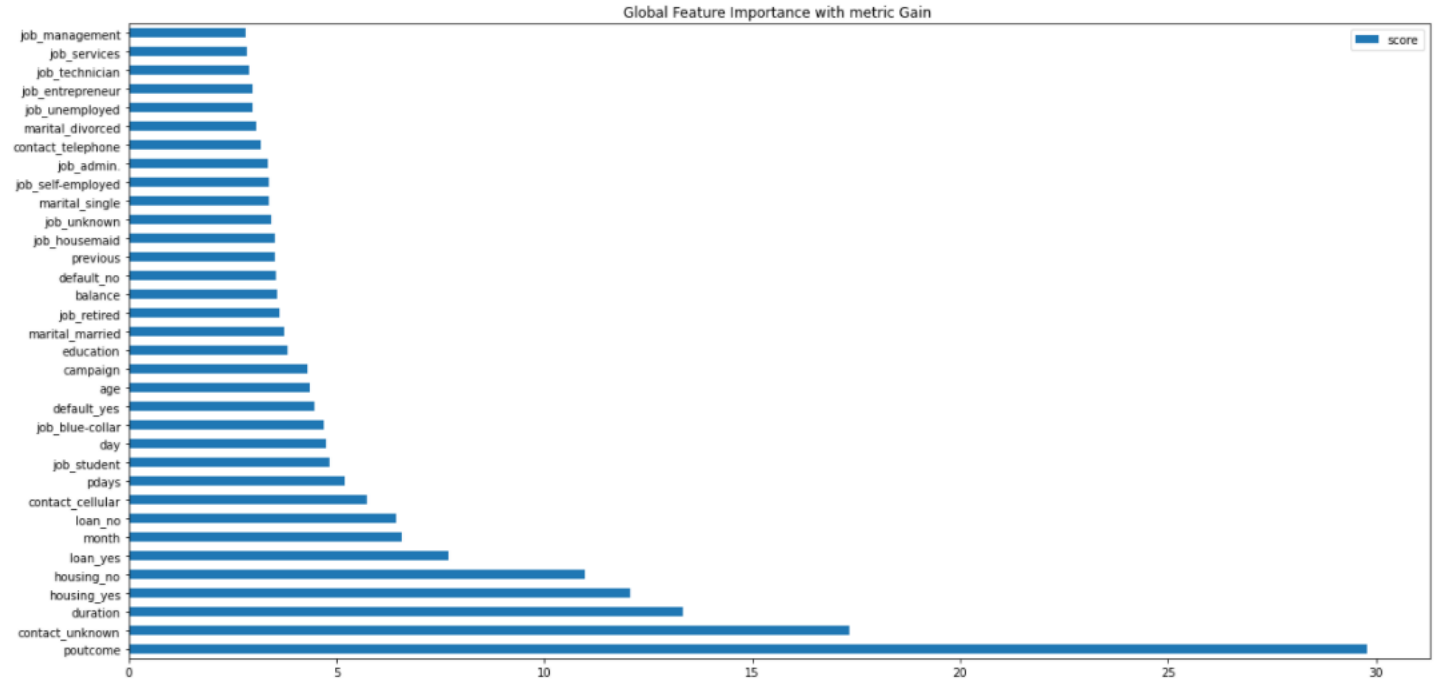

Figure 8 Global feature importance with metric gain

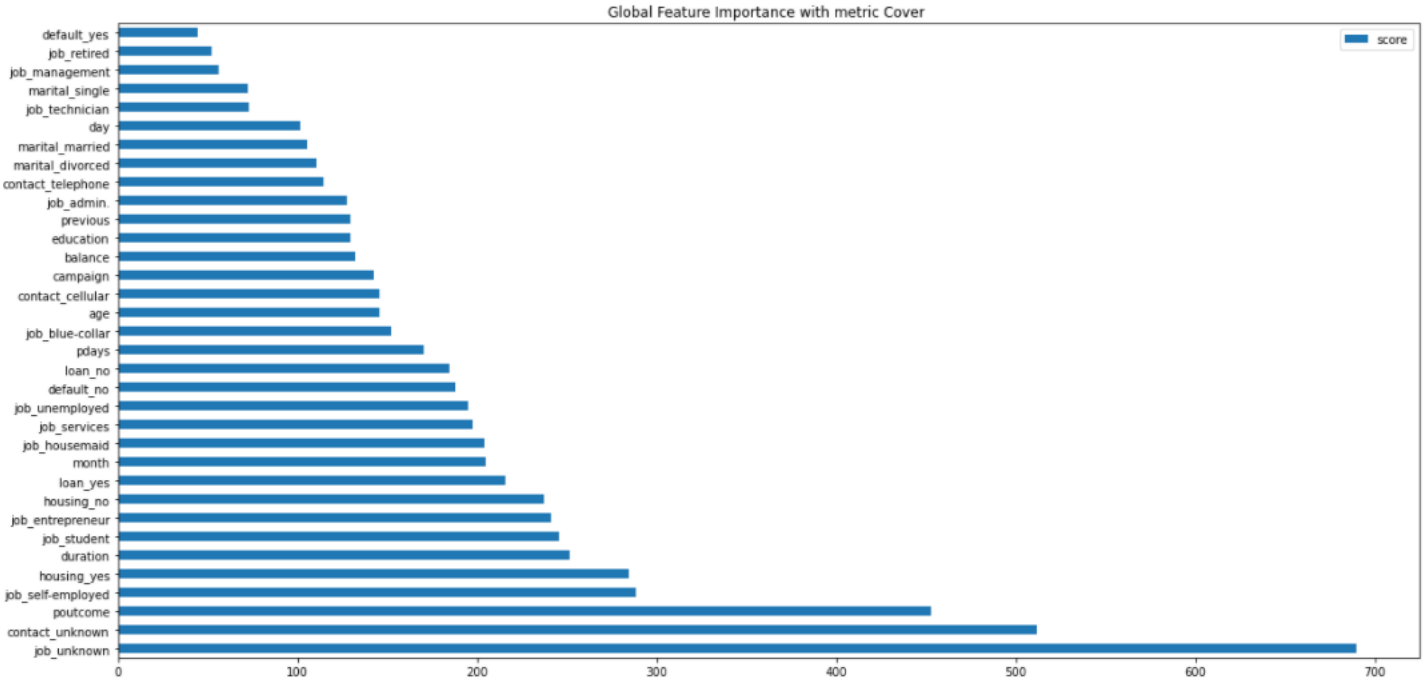

Figure 9 Global feature importance with metric cover

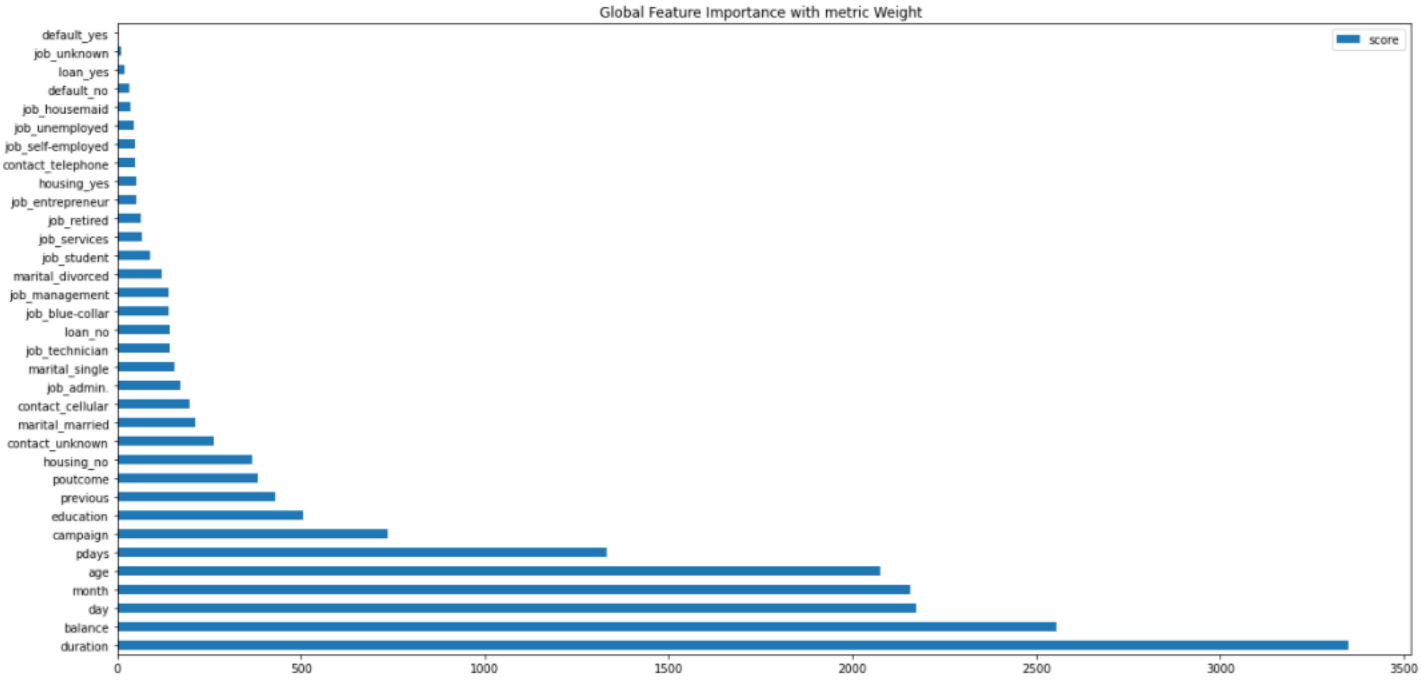

Figure 10 Global feature importance with metric weight

From the plots we get to know that the outcome of the previous marketing campaign is the most important feature for the prediction and has brought most improvement in accuracy. mode of contact which is not disclosed has the most number of observations related to this feature. Duration of last contact is used the most number of times to split the data across all trees.

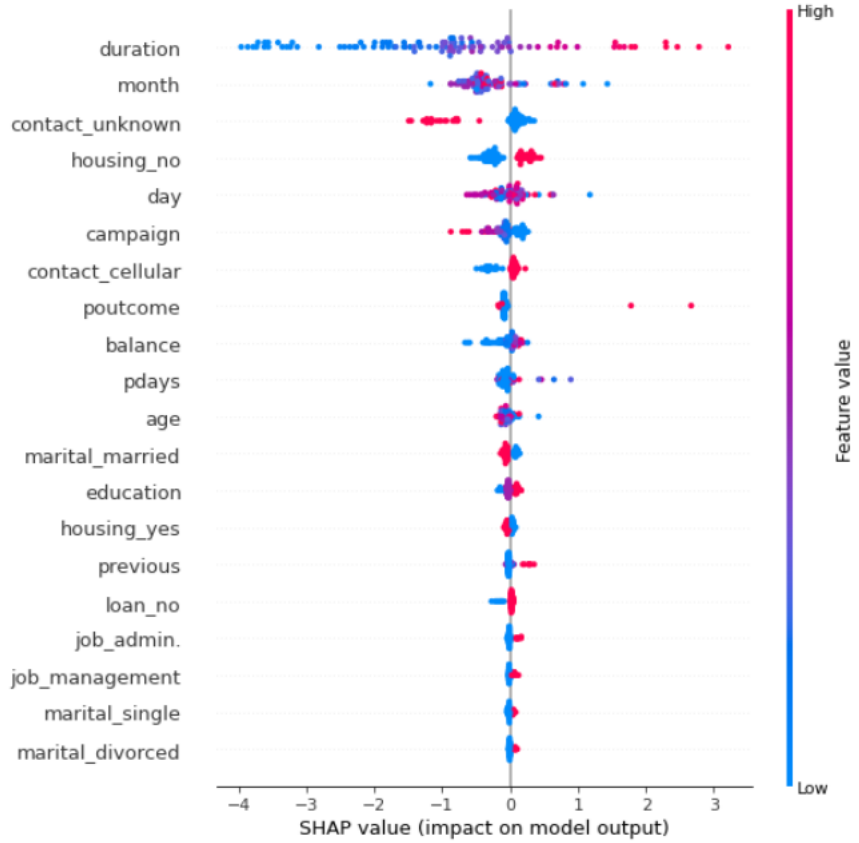

Figure 11 Local feature importance with SHAP

Local feature importance is calculated using SHAP. From the plot we get that duration of last contact has the most

impact on the model's prediction. Marital status of the individual is the least important.

5 Outlook

Our XGBoost model can be improved by tuning more parameters with more values in it on many random states. In order to perform this we need a high performance computer. We can also try training SVM model to check it's performance. I have not implemented it in the project as the run time was very high. By falsely detecting the clients as not subscribed can harm the bank as they might loose valuable clients. We can try to reduce it by collecting more data on clients who are actually subscribed in order to prevent imbalance of the data. More importance should be given to outcome of the previous marketing campaign and duration of last call as they are most important for prediction. similar features related to these important features can be collected to improve the prediction of the model.

6 References

[1]: 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, June 2014

[2]: Thripathi. D, Shukla, A K, Reddy. B R and Bopche. G S. Credit Scoring Models Using Ensemble Learning and Classifcation Approaches: A Comprehensive Survey, Springer,16 September 2021

7 Github repository

https://github.com/sagarika251/Project1030.git