CS7CS4 Research Assignment 2 - test

TEAM ID: 25 TASK ID: 104 - DATA AMOUNT

Bhaskar Rao
raob@tcd.ie
18300829

Bhavesh Mayekar
mayekarb@tcd.ie
18301407

Hamid Hassani
hassanih@tcd.ie
15338952

CONTRIBUTION

All the team members have contributed equally during the project. The team met every second day at the Lloyd building to brainstorm, plan project and discuss the results and also the team had several skype to do the project. Team finalized the three datasets to be used for the research question (available on the Github repo). Team then distributed datasets between themselves (1 for each member) and started their analysis. Individual work contribution has been detailed below:

Bhaskar Rao: He worked on the public "Bike Sharing Dataset" from UCI machine learning repository. The dependent variable (continuous) was the count of the number of bicycles rented per day and dependent variables had information about the weather conditions and specifics of the day (holiday, weekend or weekday). He did EDA, data cleaning, feature engineering on the dataset before implementing machine learning models to predict the count of bikes rented. He also performed cross-validation and employed ensemble techniques to improve the models. He the analyzed the effect of training data size on the accuracy of the model. He discussed his results and findings with the team and asked for help whenever required.

Bhavesh Mayekar: He worked on the publicly available "Census Income Dataset" from UCI machine learning repository. He performed various data cleaning and feature engineering operation on the dataset and implemented various machine learning models. The aim was to predict the whether an individual has a salary above 50k based on various input parameters. He tested various models with varying sizes of training datasets and analyzed the accuracy. Later the results and findings were discussed with the team. This dataset was not added in the final report, as there was a limitation of 1500 words. Also, he helped the team in data preprocessing and feature engineering of other two datasets and reviewed codes.

Hamid Hassani: He worked on the publicly available "Bank Marketing" dataset from UCI machine learning repository. He did detailed EDAs to visualized and understand the data and find out the relation between input features and target variable. Also, he shared insights with other teammates. He implemented multiple classification models to predict if a client will subscribe a term deposit based on marketing campaigns ran by the bank or not. He used this dataset to find out the effect of training data amount on the accuracy of models. He also presented his results and incorporated team's feedback to further improve his analysis.

Apart from the individual analysis, team was also reviewed each other's code and methodology. Team also collaborated to complete the research paper. Everyone prepared his own points for the paper and sat together to discuss and consolidate the key points using which the paper was finalized.

WORK COUNT

Word count excluding cover sheet, title, author names, tables and figures, references, acknowledgement, and appendix is **977**

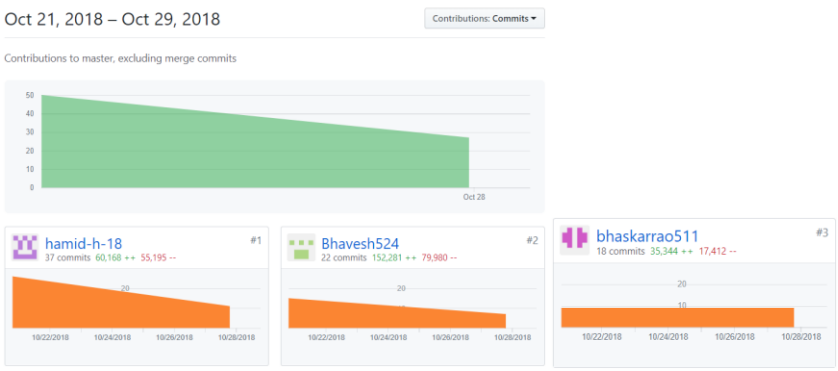
URL TO SOURCE CODE REPOSITORY

https://github.com/bhaskarrao511/CS7CS4--task-104--team-25/tree/master/sourceCode

URL TO SOURCE CODE REPOSITORY ACTIVITY

https://github.com/bhaskarrao511/CS7CS4--task-104--team-25/graphs/contributors

COMMIT ACTIVITY OF TEAM MEMBERS



We would like to thank Prof. Joeran Beel and Prof. Douglas Leith for their teachings without which this research would have been impossible

Effects of varying training data amount on Machine Learning algorithm

Bhaskar Rao
Trinity College Dublin
raob@tcd.ie

Bhavesh Mayekar
Trinity College Dublin
mayekarb@tcd.ie

Hamid Hassani
Trinity College Dublin
hassanih@tcd.ie

1 INTRODUCTION

Machine Learning is a process by which computer can make prediction through analysing the input data and it is either curve fitting or classification tasks. [1] In last few years, the use of machine learning has been increased tremendously due to increase of computational power. A report published by McKinsey Global Institute claims that ML will revolutionize the future innovation [2].

In ML, data plays an important role and in order to train the algorithm, the data is divided into training and testing datasets. Therefore, the first question that arise is that how much data is required to train the model effectively. As author's knowledge, there is no definite answer to this, but in most scenario, it depends on various factors like complexity of the algorithm, input features, correlation between data etc.

Keywords – Machine Learning, Training Data, Accuracy

2 RELATED WORK

In [3], it was proposed that training size should be defined by specifying confidence interval widths for classification algorithm in bio spectroscopy field. As mentioned in [4], increasing the training dataset will overfit the model. It was found in [5] that how the performance of models vary with the training dataset size in biomedical applications. The investigation in [6] describe about how much training data is required to have an accurate model in medical image deep learning systems.

3 METHODOLOGY

In this research, following steps were followed: Identifying relevant datasets and their target features, data pre-processing, breaking the dataset into test and train, splitting training dataset into chunks of different length, build models upon these chunks, evaluate these models with the test data and compare the accuracy vs train data size.

3.1 Datasets

Two multivariate datasets from the UCI machine learning library were used for the research, "Bike Sharing" and "Bank Marketing" datasets. Former dataset has 17,379 observations and 16 features recorded for two years at day-hour level. The target variable is the number of bikes rented and features are environmental conditions at the hour. The later dataset is a bank customer level data which has more than 40K instances and 10 features. Target variable is if the contacted customer subscribed to the bank term deposit or not and the features are bank client data and some information related to the last contact of current customer.

3.2 Data Pre-processing

In both the datasets, null values were treated, and EDA (exploratory data analysis) was performed to understand the datasets. Label encoding was applied on categorical variables.

4.2.1 Pre-processing for "Bike Sharing" Dataset

Correlation analysis was performed to understand the relationship between the features and the dependent variable. From the heat plot in Fig. 1, there is no correlation between 'windspeed' and 'cnt' target variables, therefore it could be dropped. A high correlation exists between 'casual' and 'registered' features ('registered' being the subset of 'cnt', won't be used), 'casual' will be dropped, as well. New features like Sunday flag (day is Sunday or not) and day period ('noon', 'evening' etc.) were created to extract more useful information from data. Categorical variables like 'season' and 'weather' were label encoded.

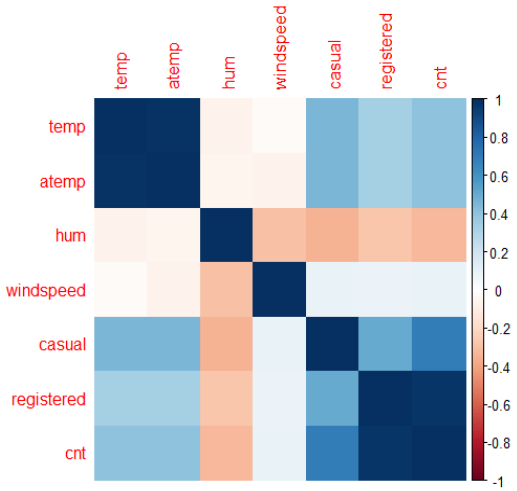


Figure 1: Correlation analysis for bike-sharing dataset

4.2.2 Pre-processing for "Bank Marketing" Dataset

In bank marketing dataset, most of the features were categorical, so they were encoded and transformed to quantitative data. Then, the outliers were detected with respect to quartile range and filled with closest values in range. Also, there was no any missing data in this dataset. In order to do feature selection in this dataset, some features will be visualized and analyzed in this section and reminded features in Appendix A.

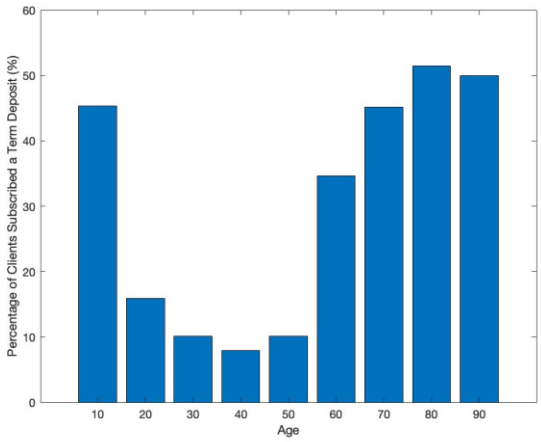


Figure 2: Distribution of subscribed clients vs. age

In Fig. [2, 3], it has been shown that older people are more probable to subscribe a term deposit. However, the large portion of dataset contains young people. Hence, it would be better to change the campaign with more promotions for young people in the future.

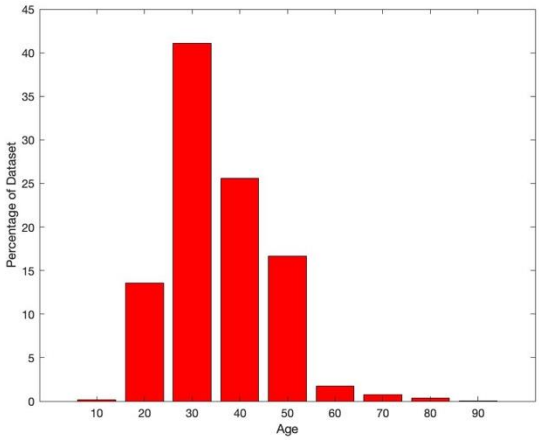


Figure 3: Distribution of clients vs. age in dataset

In Fig. [4,5], it has been shown that retired and students are more interested in subscribing a term deposit but they are less than 10% of whole dataset.

After did feature selection, house and personal loans features were removed due to low correlation with target variables. Then, the continues "age" values were discretized to a multiple of 10 to improve the accuracy. Formerly, features were rescaled with min-max normalization method before building the model.

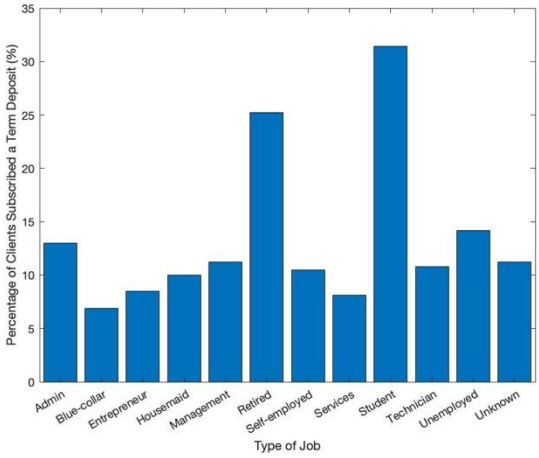


Figure 4: Distribution of subscribed clients vs. type of job

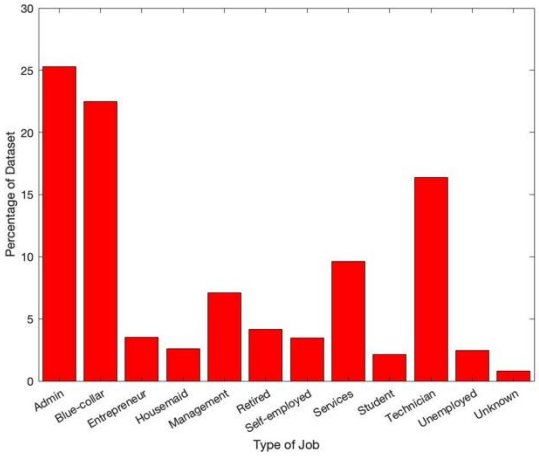


Figure 5: Distribution of clients vs. type of job in dataset

3.3 Train data splitting

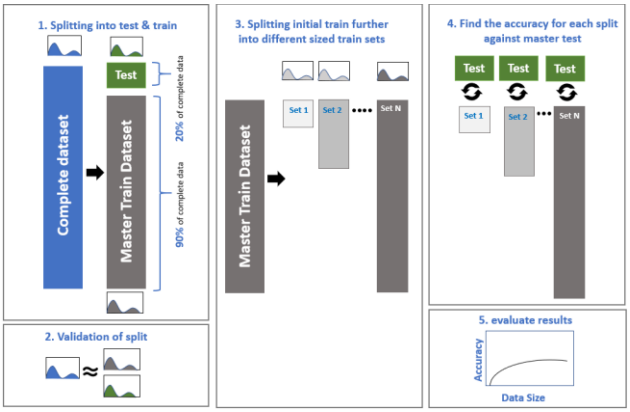


Figure 6: Data splitting approach

To analyze the impact of data amount on the performance of ML algorithms, each dataset was split into validation and train parts in ratio of 20:80 as shown in Fig. 6. Then, some separate training datasets were further generated from the initial train set. The difference between size of consecutive training sub-sets was constant, i.e. 10% of initial train data size. Furthermore, distributions of all validation and train datasets were compared with that of the complete dataset to verify sample selection.

3.4 Model Building

All the models were implemented on the individual training sub-sets generated from the initial train dataset.

4.4.1 "Bike sharing" Dataset

As dependent variable in "Bike Sharing" dataset is continuous, linear regression (using R "Stats" package), ridge regression (using R "glmnet" package), support vector regression ("e1071" package) and random forest algorithms ("randomForest" package) were implemented and compared. The weighted ensemble of all the algorithms was also calculated to understand the effects of ensembling on data amount. Hyperparameters were tuned based on cross-validation. The optimum value for number of trees in random forest algorithm, penalty parameter (λ) in ridge regression and ensemble model weights were determined based on least error method. The hyperparameter used can be found in Table 1.

Table 1: Hyperparameters tuning for Bike Sharing data models

Model	Hyperparameter	Range	Optimum Value
Ridge regression	λ : Penalty parameter	-2 to +3	λ : 0.316
Random forest	Number of trees	1 to 300	#trees: 155
Ensamble of 4 models	"a" : weight of linear model "b" : Weight of ridge model "c" : weight of random forest model "d" : weight of SVM model	All individual weights can have either 1.0 , 0.5 , 0.25 or 0.75 values	a = 0.25, b= 1, c = 0.25, d = 0.25

4.4.2 "Bank Marketing" Dataset

In In "Bank Marketing" dataset, the target variable is binary, and the problem is a binary classification. Fig. 7 shows the distribution of the target variable in this dataset. It is obvious that it is an unbalanced dataset due to comparatively a smaller number of clients have subscribed a term deposit. Therefore, it is necessary to apply some metrics which are appropriate to this dataset such as F1 accuracy.

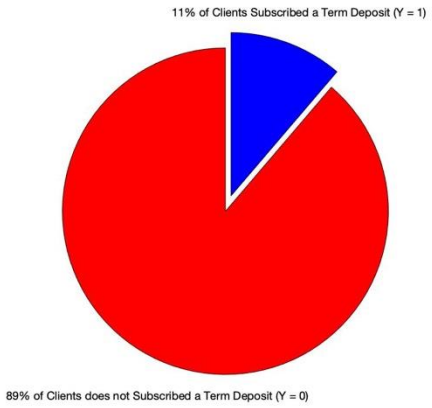


Figure 7: Distribution of target variable Y

In order to find a proper model, logistic regression, K nearest neighbor (KNN), decision tree, random forest and Gaussian Naive Bayes (GNB) algorithms were implemented in Python using scikit-learn library [10].

4 RESULTS

4.1 "Bike sharing" Dataset

For the Bike Sharing dataset in Fig. 8, all the algorithms apart from linear regression and ridge regression behave differently when size of training dataset is increased. For linear regression model, mean absolute percentage error (MAPE) increased till 20% of train data size and then eventually becoming nearly constant after 40%. Ridge regression performs very similar to linear regression (almost perfect overlap in graph X), this is because of low multicollinearity in our data (reduced via correlation analysis). In case of support vector regression model, error decreases till the train size is 30% of overall training data but beyond this, error becomes almost constant. For more complex model like random forest, model's error decrease constantly when size of training data increase. The results of these models were ensembled to get the weighted average prediction, this ensemble performed the best and have the lowest MAPE of ~ 88% amongst all the models. The error of the ensemble remained constant with increase in data size, this is because both bias and variance are reduced by the weighted average ensembling of different models. To sum up, size of dataset has a significant impact of machine learning models up to a certain level. Complex models will have better accuracy compared to linear/simple models and ensembling will give the best results consistently.

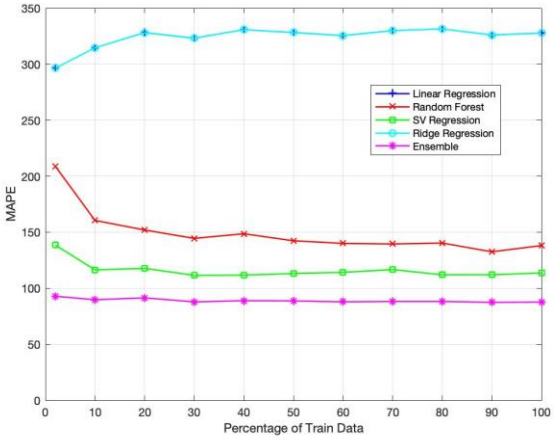


Figure 8: MAPE vs. train data size for Bike Sharing dataset

4.2 "Bank Marketing" Dataset

Fig. 9 shows the impact of varying training dataset size is related to selected classifier algorithm, in Bank Marketing dataset. In this figure, the first point on x-axis is 0.01% (>200 instances) and F1 accuracy calculated using bootstrapping with 20 times resampling. As shown in Fig. 9, logistic regression, linear SVM and Gaussian NB show a lower accuracy because of the underfitting phenomenon. However, KNN and Random Forest have higher performance. Then after increasing the size of dataset beyond 10%, there is no significant enhancement in the performance of the models are visible. It can be seen that more complex models like Decision Tree and Random Forest reach higher accuracy compared to the simple models like Logistic Regression, GNB and KNN. In addition, it is clear that after increasing the size of training dataset to 10%, all the applied methods in this setup behave in a same fashion without any obvious change in the overall accuracy. Therefore, the size of training data changes the accuracy in some algorithm, but complex ones are more robust to decrease the number of input instances for training.

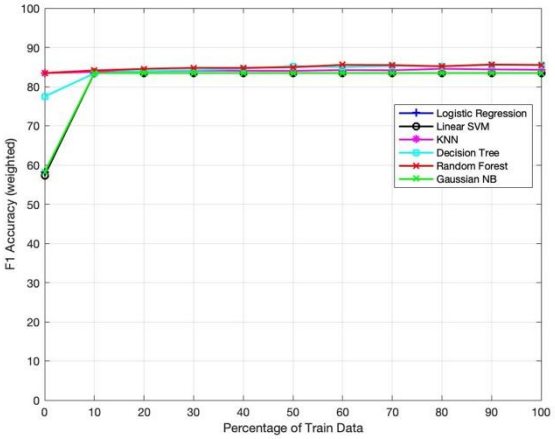


Figure 9: Accuracy vs. train data size for Bank Marketing dataset

5 LIMITATIONS AND OUTLOOK

To further investigate about the impact of training dataset size on accuracy, the current models can be improved by using optimized feature engineering. Feature selection and missing value imputation have an impact on the accuracy.

For bank marketing, we can do feature engineering and combine some features and generate promising new features. In addition, we can use some wrapper methods to remove the highly correlated input features and reduce the complexity of model.

ACKNOWLEDGMENTS

This analysis was conducted as part of the 2018/19 Machine Learning module CS7CS4/CS4404 at Trinity College Dublin).

REFERENCES

[1] P. Domingos. 2012. A few useful things to know about machine learning. *Communications of ACM*, vol. 55, no. 10, 78–87.

[2] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. Byers. 2011. Big data: The next frontier for innovation, competition, and productivity. Technical report. McKinsey Global Institute.

[3] C. Beleites, U. Neugebauer, T. Bocklitz, C. Krafft, J. Popp. 2013. Sample size planning for classification models. *Analytica Chimica Acta*, Volume 760, 25-33.

[4] Friedman, J., Hastie, T., and Tibshirani, R. 2001. *The elements of statistical learning*, (2nd. Ed.). Springer Series in Statistics, New York.

[5] Hajian-Tilaki, K. 2014. Sample size estimation in diagnostic test studies of biomedical informatics. *Journal of biomedical informatics*, Vol 48, 193–204.
[6] Cho, J., Lee, K., Shin, E., Choy, G., and Do, S. 2015. How much data is needed to train a medical image deep learning system to achieve necessary high accuracy? arXiv: 1511.06348 . https://arxiv.org/abs/1511.06348.
[7] Sun, C., Shrivastava, A., Singh, S., and Gupta, A. 2017. Revisiting unreasonable effectiveness of data in deep learning era. arXiv:1707.02968. Retrieved from https://arxiv.org/abs/1707.02968.
[8] Fanaee-T, Hadi, and Gama, Joao. 2013. Event labeling combining ensemble detectors and background knowledge', Progress in Artificial Intelligence, pp. 1-15, Springer Berlin Heidelberg.
[9] Moro, P. Cortez and P. Rita. 2014. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31.
[10] Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

APPENDIX A

In this section, the other input features of Bank Marketing dataset will be discussed in more detail as follows:

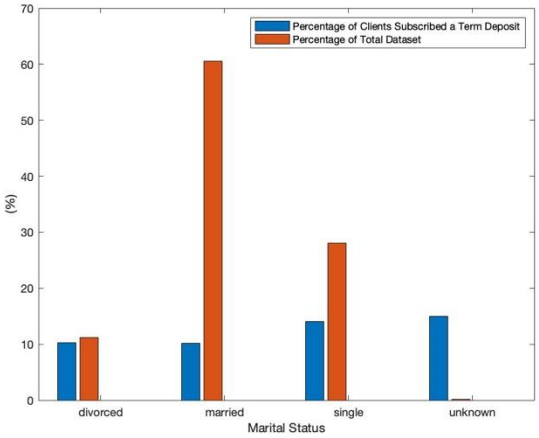


Figure 10: Analyze "Marital" feature

Fig. 10 shows the relation between marital statues and subscribing to a term deposit. It could be seen that single clients are a bit more probable to subscribe a term deposit.

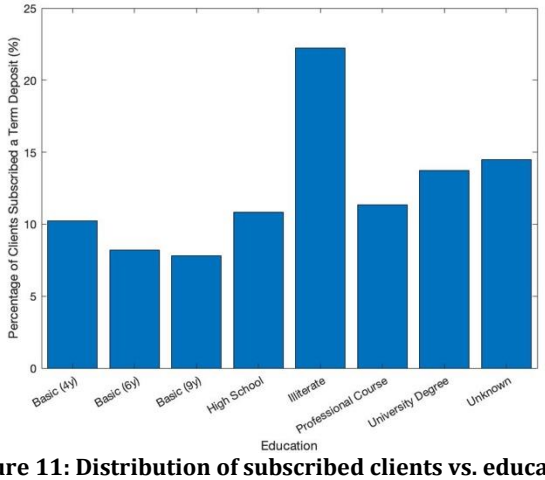


Figure 11: Distribution of subscribed clients vs. education

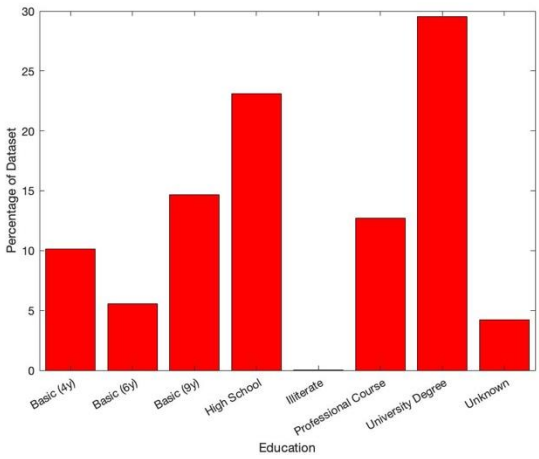


Figure 12: Distribution of clients vs. education in dataset

In Fig. [11,12], the impact of education level in target variable has been shown. Here, the illiterate people are more interested in subscribing a term deposit, but they are less than 1% of dataset. In addition, the probability of subscribing for other education levels are roughly close to each other. So, we should not expect effective information from this feature.

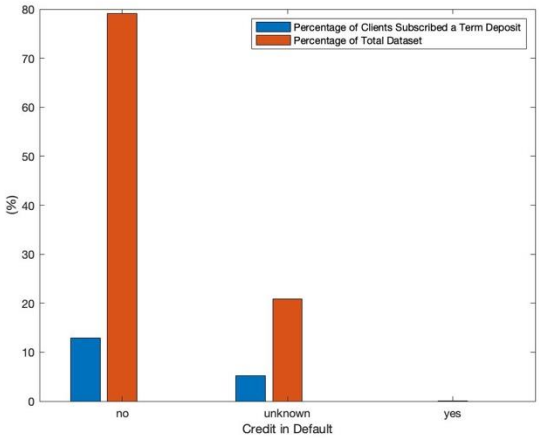


Figure 13: Analyze "Credit" feature

In Fig. 13, it is clear that clients with no credit are more probable to subscribe a term deposit.

As shown in Fig. 14, it is not possible to extract useful information from this feature as the probability of subscribing are the same for all groups. Hence, it could be removed from independent variables.

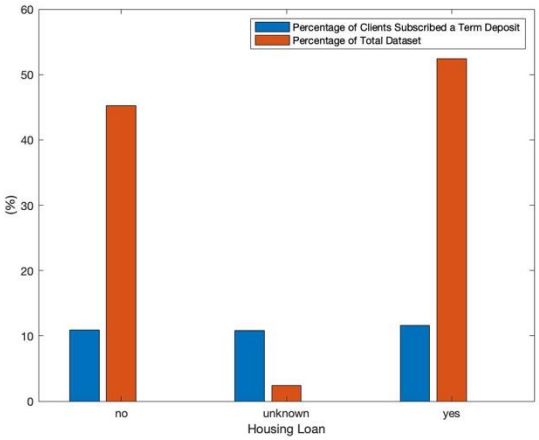


Figure 14: Analyze "Housing Loan" feature

In Fig. 15, the dataset is balanced regards number of clients have housing loan or not. Therefore, useful information is less probable.

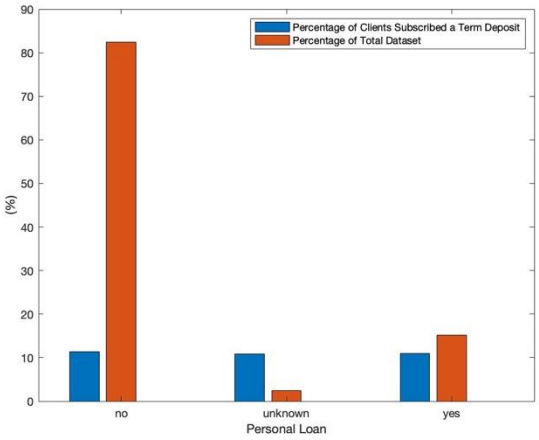


Figure 15: Analyze "Personal Loan" feature

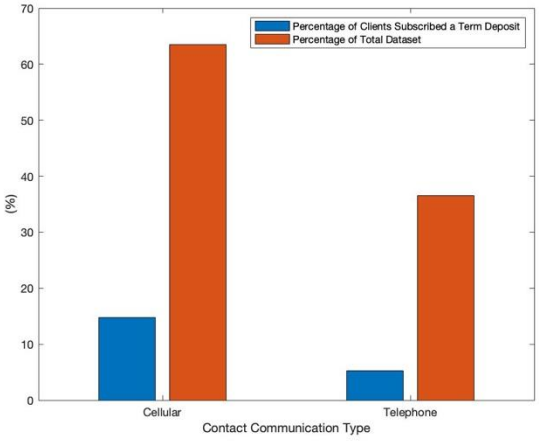


Figure 16: Analyze "Contact" feature

Fig. 16 shows that using fixed-line network to talk with clients increases the risk of refusing to subscribe a term deposit by a client.

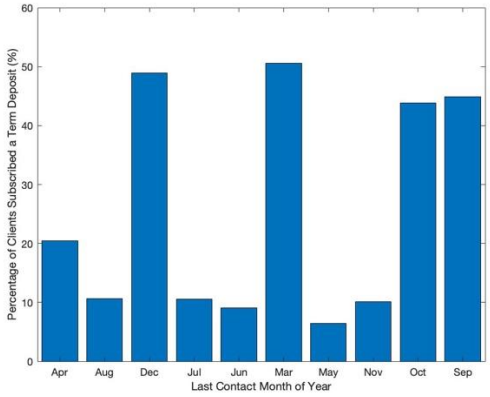


Figure 17: Distribution of subscribed clients vs. Last Contact Month

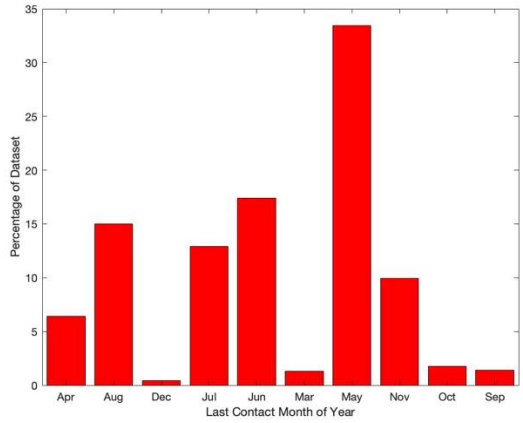


Figure 18: Distribution of clients vs. Last Contact Month in dataset

In Fig. [17,18], the impact of contact month on subscribing a term deposit is visualized. Here, the probability of subscribing a term deposit is highly related to some specific months such as December and March.

It is worth mentioning that the feature "duration", the duration of last time the bank has called to a client in seconds, is highly correlated with target variable because the more the bank talks with a client the more expected the client will subscribe to a term deposit because he/she shows higher interest. Therefore, to have a realistic model, it was removed in this study.