GEN 511 - MACHINE LEARNING
HACKATHON

Subscription to a Term Deposit

Mohammad Khalid – IMT2017028
Anurag Pendyala – IMT2017504
K. Sailesh – IMT2017524

CONTENTS

1 Abstract

Term Deposit: Term Deposit is a cash investment held at an any financial institution for a fixed period of time on which an interest is paid at a fixed rate agreed before hand by the institution and depositor on the amount invested.

Term deposits are a win-win for both the financial institutions and the investors. Unlike the usual customers who deposit and withdraw money at their wish, Term depositors cash in only after end of the Term period. This gives the institutions a good amount of cash reserves which they lend or invest elsewhere and generate profits.

Thus financial institutions spend considerable amount of resources in identifying a good term depositor. Then the institution use the fund for large investments in future (indirectly getting investment seed capital). Hence a Machine Learning model can be trained using a few features and predictions can be made whether the person would subscribe for a term deposit or not.

2 Problem statement

The goal is to predict if the client will subscribe (yes/no) a term deposit (variable y) using a data set compiled by a Portuguese bank during one of their digital campaigns.

3 Data Set

The data is related with direct marketing campaigns to attract potent investors for term deposits in a Portuguese banking institution.

The source of the data set is : [Moro et al., 2014] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014

Each bank client's data contains 20 attributes. A brief description of each attribute are given in the next section.

3.1 Attribute Information

1. age : Age of the person. Numeric in nature.
2. job : Type of job. Categorical in nature and includes the following:

    - admin
    - blue-collar
    - entrepreneur
    - housemaid
    - management
    - retired
    - self employed
    - services
    - student
    - technician
    - unemployed
    - unknown

3. marital : Marital Status of the person. Categorical in nature and includes the following:

    - divorced
    - married
    - single
    - unknown

    Note: "divorced" means divorced or widowed

4. education : Education standard of the person. Categorical in nature and includes the following categories:

    - basic.4y
    - basic.6y
    - high.school
    - illiterate
    - professional.course
    - university.degree
    - unknown

5. default : If the person has a credit in default. Categorical and binary in nature. Includes yes, no or unknown.
6. housing : If the person has a housing loan. Categorical and binary in nature. Includes yes, no or unknown.
7. loan : If the person has a personal loan. Categorical and binary in nature. Includes yes, no or unknown.
8. contact : Mode of communication with the person. Categorical and binary in nature. Includes cellular and telephone.
9. month : Last month of contact with the person. Categorical in nature and includes all the months of the year, jan, feb, ... , nov, dec.
10. day_of_week : Last contact day of the week. Categorical and includes:

    - mon
    - tue
    - wed
    - thu
    - fri

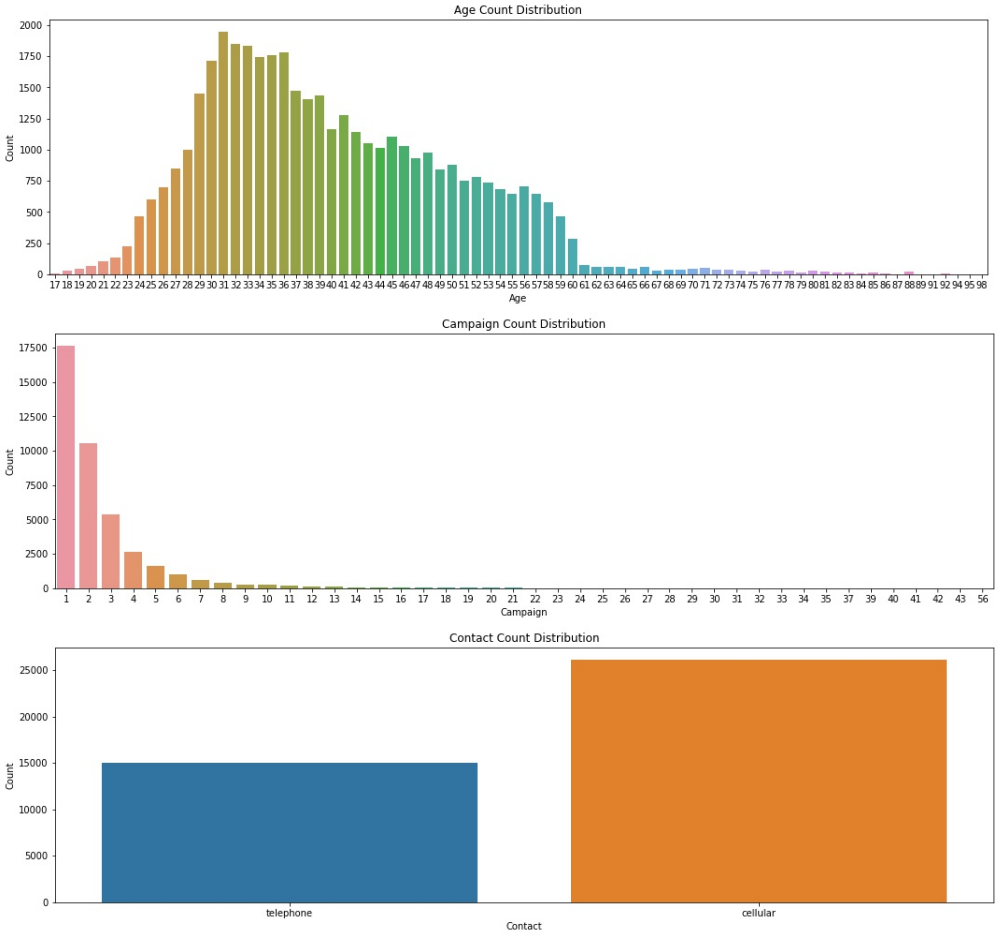
11. duration : Last contact duration in seconds. Numeric in nature.
12. campaign : Number of contacts performed during this campaign. Numeric in nature.
13. pdays : Number of days that passed after the client was last contacted from previous campaign. Numeric in nature. Note: 999 means client was not previously contacted
14. previous : Number of contacts performed before this campaign.
15. poutcome : Outcome of previous marketing campaign. Categorical in nature and includes the following categories:

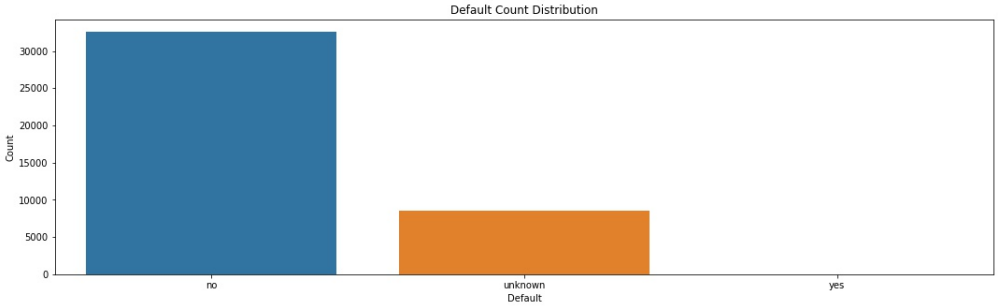
- failure
- nonexistent
- success

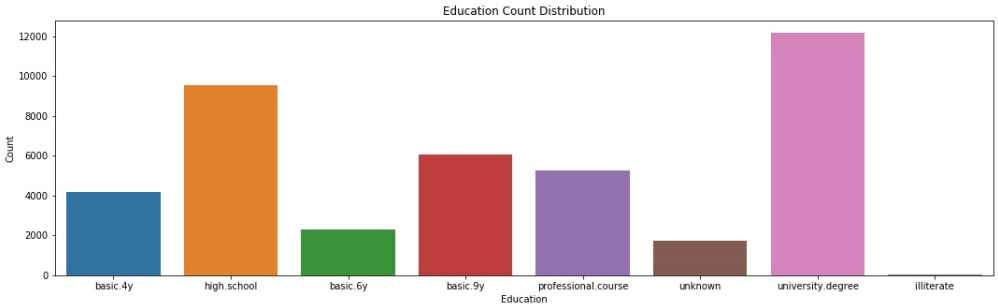
16. emp.var.rate : Employment Variation Rate. Quarterly Indicator and Numeric in nature.
17. cons.price.idx : Consumer Price Index Quarterly Indicator and Numeric in nature.
18. cons.conf.idx : Consumer Confidence Index Quarterly Indicator and Numeric in nature.
19. euribor3m : Euribor 3 Month Rate. Daily Indicator and Numeric in nature.
20. nr.employed : Number of Employees Quarterly Indicator and Numeric in nature.
21. y : Output variable. Has the client subscribed a term deposit or not.

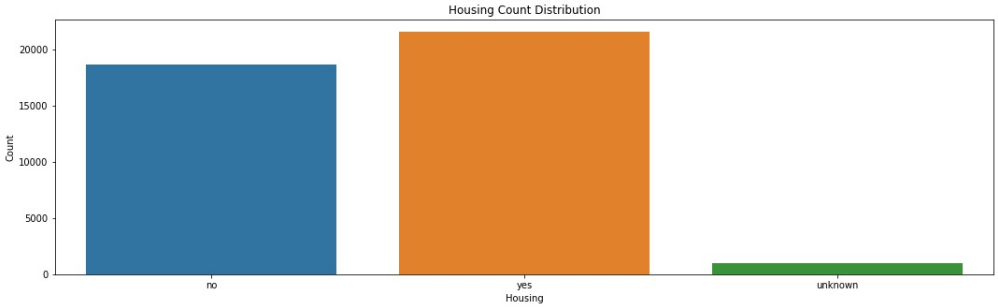
4 Exploratory Data Analysis

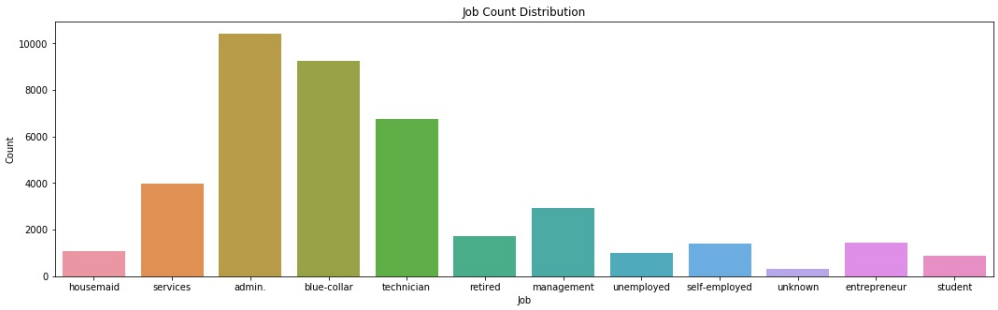
The count plot for few of the features are as follows:

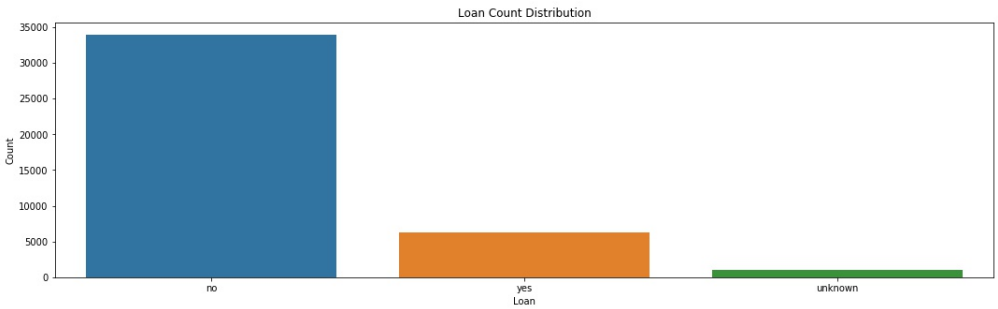


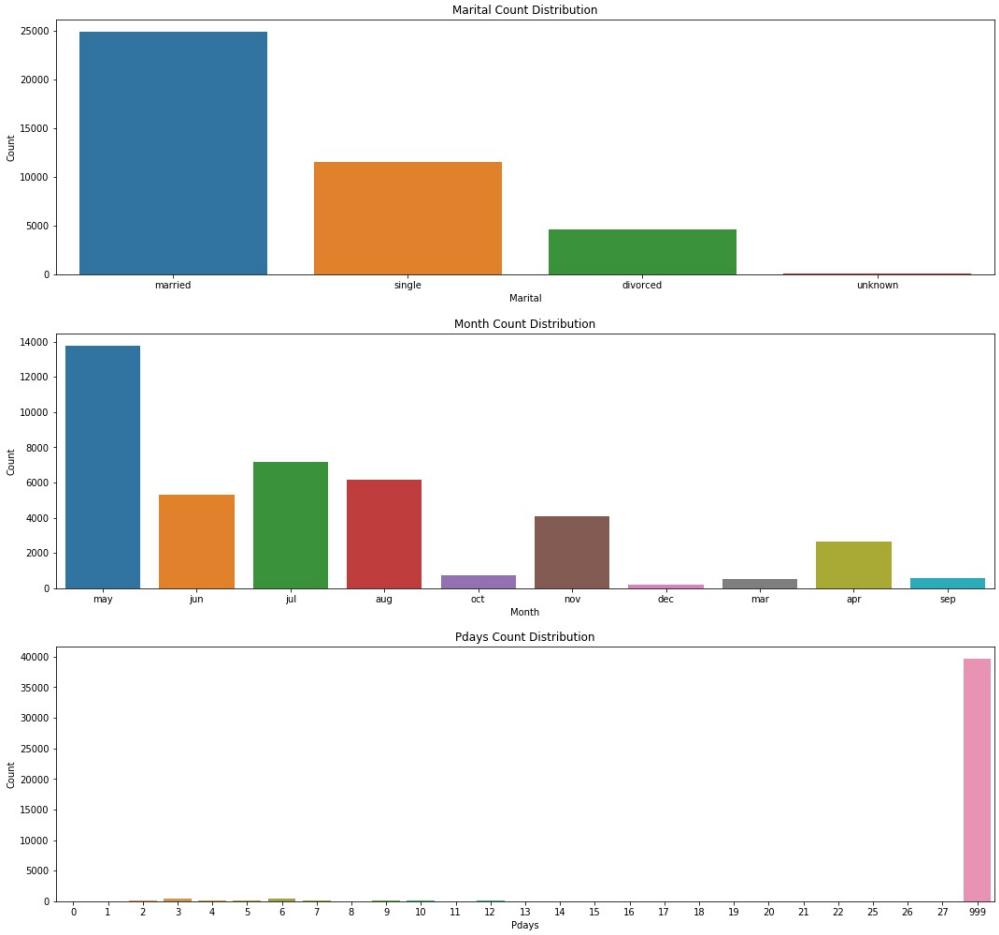

Default Count Distribution
Count
Default
no
unknown
yes


Education Count Distribution
Count
Education
basic.4y
high.school
basic.6y
basic.9y
professional.course
unknown
university.degree
illiterate


Housing Count Distribution
Count
Housing
no
yes
unknown


Job Count Distribution
Count
Job
housemaid
services
admin.
blue-collar
technician
retired
management
unemployed
self-employed
unknown
entrepreneur
student


Loan Count Distribution
Count
Loan
no
yes
unknown



5 Data pre-processing

The data contains around 41,000 data points with 20 features plus one output variable. After investigating the data set, it can be observed that there are around 330 unknown values in the job feature vector, 80 in marital, 1,731 in education, 8,597 in default, 990 in housing and 990 in loan. Unknowns here are nothing but missing data. So unknowns were replaced by NaNs.

5.1 Dropping Columns

The columns pdays, default and loan have been dropped as they don't contribute much towards the classification of data. In the original data set *default* had around 90% of nos. Therefore it doesn't have much variance and is a redundant feature. A similar argument holds for *pdays* and *loan*. There are around 39,000 people(around 95%) who weren't contacted and around 90$ nos for pdays and loan features respectively. These columns hardly have any useful data and hence don't help in classification of the data.

5.2 Predicting Missing Values

Most of the missing data is part of a categorical non-numerical columns. Before imputing the missing data categorical data has to be converted to numerical form, since the library implementation of imputers require the data in each of the features to be a real number.

Models like decision trees can handle input data which has both numerical and categorical data. But a large number of models cannot handle both numerical and categorical data together when passed as input for training. Most of the models are optimized to handle data of single type efficiently and hence give up on handling different types of data. Logistic Regression has been used to predict the missing values.

5.3 Encoding Categorical Data

There are 2 popular methods to handle categorical data :

- One Hot Encoding
- Label Encoding

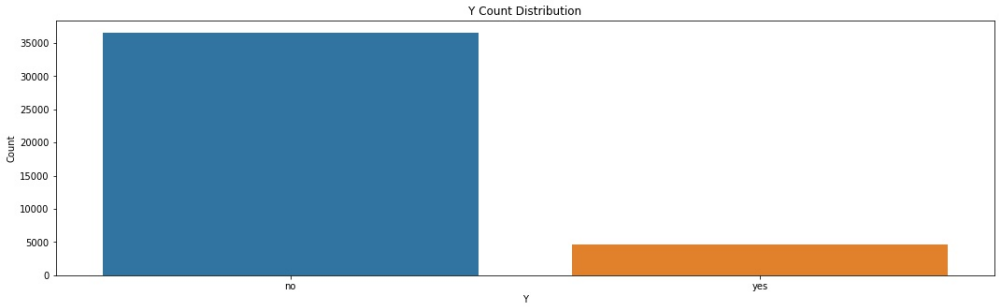
A Label Encoder encodes labels with a value between 0 and $n-1$ where n is the number of distinct labels is the feature. In some cases Label Encoding is not preferred. Sometimes a label encoding of a class gives some sense of hierarchy when it is clearly not needed.

To avoid this, the column can be One Hot Encoded. One hot encoding takes a column with categorical data and splits it into n columns, where n is the number of distinct categories in the column. The class of the data point is preserved by setting that particular column with one and remaining columns with zeros.

In case of binary features, label encoding is preferred as it just replaces the 2 labels present with 0 or 1. All the categorical features with only 2 classes have been label encoded and the remaining ones have been One Hot Encoded.

The categorical features which have more than two labels multi-labeled data can be converted to numerical form using One Hot encoding and the same has been done. The columns job; marital; education; month; day_of_week have been One hot Encoded. Now the data is ready in a single format and can be fed to models for training.

5.4 Balancing the Data set



After imputing the missing data in the data set, the frequency plots showed that the data set was heavily skewed towards No class with 36548/40,000 data points in it, where as Yes class had only 4640/40,000 data points. Since the data set was heavily skewed it was re-sampled such that number of data points in the data set in both the classes were equally distributed. This is to prevent the model getting skewed towards the dominant class

The re-samplings done with the help of Synthetic Minority Oversampling Technique (SMOTE) from imblearn library . After re-sampling the data there were 73096 data points in the new data set with 36548 data points in No class and 36548 data points in Yes class.

Additionally data has been normalized to avoid lopsiding effect by a particular feature.

6 Model Selection

The re-sampled data set has been split into train and test subsets in the ratio of 80:20 Logistic Regression and kNN have been used to predict the output label(Classify as 0 or 1).

The parameters used for logistic regression are :

- Penalty : *L2*
- Tolerance Factor : 0.0001
- Class Weight : 1 for each class.
- Maximum Iterations : 100
- Solver : warn
- max_iter : 100

The parameters used for kNN are :

- k_Neighbours : variable
- Weights : 1 for each class.
- Metric : minkoswki
- Algorithm : Ball Tree, KD Tree or Brute

7 Analysis

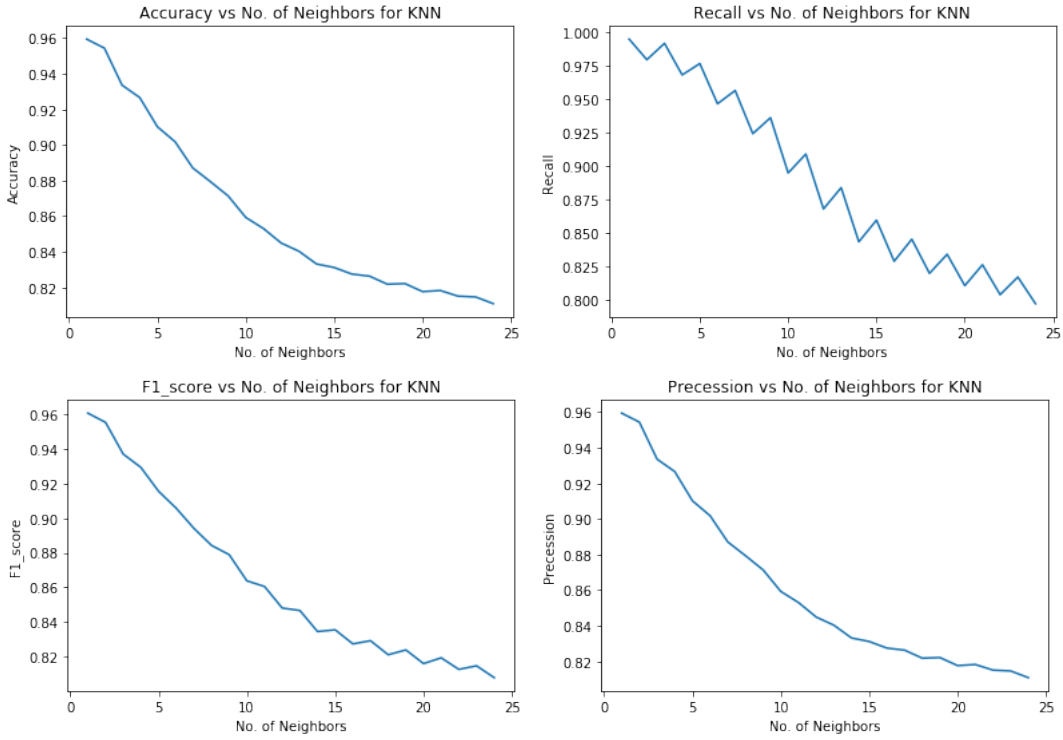
7.1 Logistic Regression

On training a Logistic Regression model, with the above mentioned parameters, to classify the data, the following metrics have been achieved :

- Accuracy : 87.03%
- F1 Score : 87.04%
- Recall : 88.47%
- Precision Score : 85.64%

7.2 k Nearest Neighbors

A kNN model has also been trained across different number of neighbors. It is evident from the results that the performance of the model decreases with increase in the number of neighbors. After the number of neighbors reach a sufficiently large value, the rate of change of performance is almost negligible. All of this has been represented with the graphs below:



The best results were achieved when the number of neighbors is equal to 1.

- Accuracy : 95.60%
- F1 Score : 95.75%
- Recall : 99.36%
- Precision Score : 92.39%

8 Conclusion

Credit Scoring and identification of potential investors is an area gaining a lot of importance off-late. This project is just a starting point towards building such applications. One other way this problem can be modified to help an individual identify a potent term deposit.

9 References and Citations

- '[Moro et al., 2014]' S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, In press http://dx.doi.org/10.1016/j.dss.2014.03.001
- Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.
- https://www.investopedia.com/terms/t/termdeposit.asp
- https://towardsdatascience.com/understanding-feature-engineering-part-2-categorical-data-f54324193e63
- https://machinelearningmastery.com/tactics-to-combat-imbalanced-classes-in-your-machine-learning-dataset/