

Figure 1: Fist 5 row of given data

According to Figure 1, we have 14 columns which 5 of them have numerical and 9 of them have categorical data.

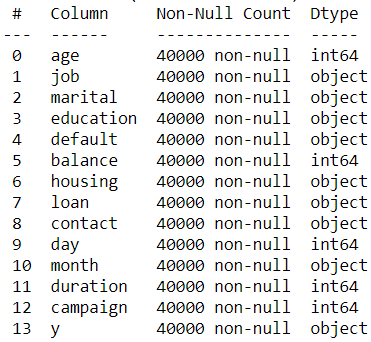


Figure 2: Information about data

When we examine the information given from pandas, data has no null object inside it. Therefore, there is no imputation for missing values.

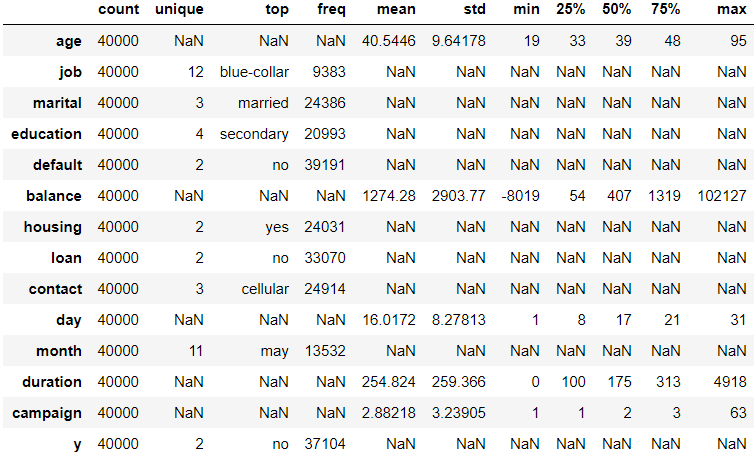


Figure 3: Description for data

From given figure at Figure 3, description for numerical and categorical columns can be seen. For categorical columns, unique values are more than or equals to 2. Hence, we can say that our categorical columns have variance to teach our model useful insights. Target column, which is y, has 2 different value. If we choose to use neural networks, we can select sigmoid function for activation. For numerical columns unique values are NaN, because of the continuous values. When we examine mean and standard deviation of these columns. We can see that; their values are different from each other. It can be problem for some algorithms. So, we need to scale these columns to prevent overfitting. However, this situation is not a problem for Tree Based Algorithms. We can see that standard deviations of numerical columns are not equal to 0. So, these columns have variance to teach our model useful insights.

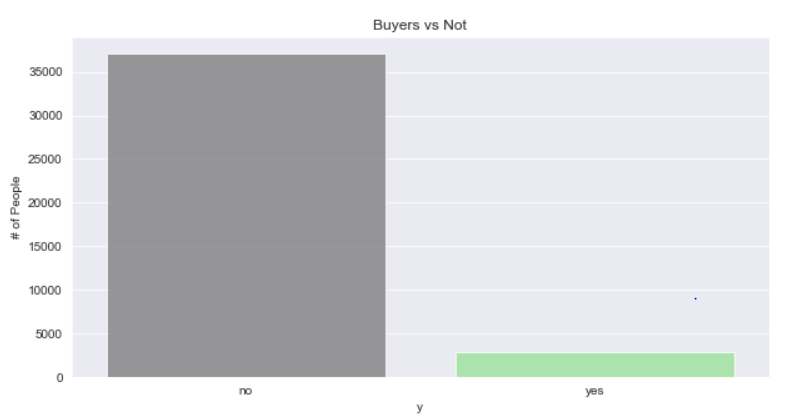


Figure 4: Value counts for target column

According to given figure, we have an imbalance dataset. Total data length is 40000. According to Figure 4, we have 2896 buyer for term deposit and 37104 non-buyers. We are going to use Over Sampling method to create synthetic data for buyer. Thereby, our algorithm can see more buyer data and classification can be easier.



Figure 5: Value counts for categorical features

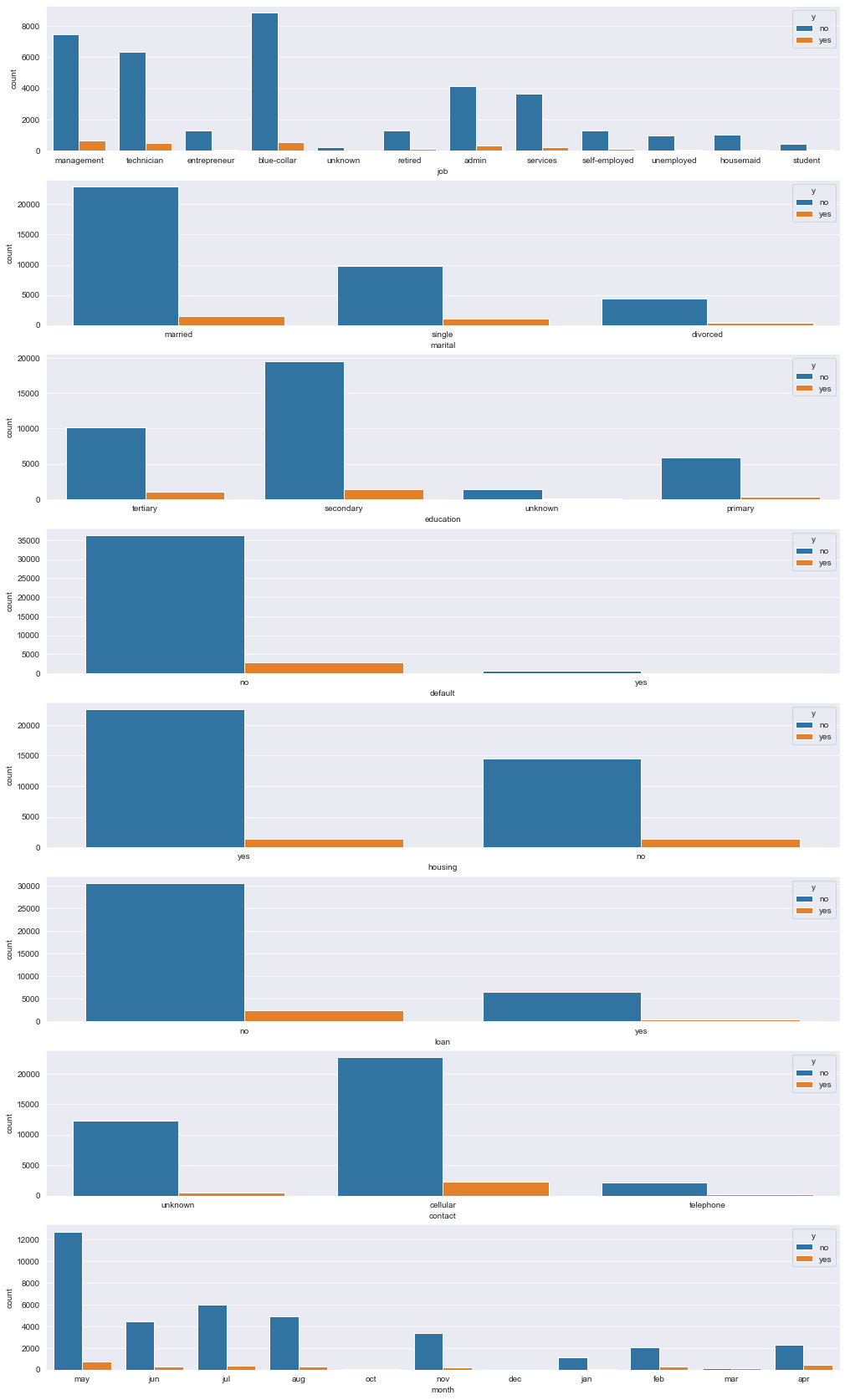


Figure 6: Value counts for categorical features with target value hue

When we examine categorical feature separately, first of them is jobs of customers. We see most people are blue collar. When we examine the numbers, most of buyers work in management. However, If we calculate the ratio of buyers and non-buyers according to work groups, we can see that students have the highest buyer rate with %15.64. If we draw the age chart, we can see that the highest rate of buyers is between the ages of 20-30. Thus, this situation makes sense. Second one is marital status. When we examine the numbers, most of buyers are married people according to numbers. However, If we calculate the ratio of buyers and non-buyers according to marital status, we can see that single people have the highest buyer rate with %9.43. . Third one is education. When we examine the numbers, most of buyers has secondary education according to numbers. However, If we calculate the ratio of buyers and non-buyers according to education, we can see that people, who have tertiary education, have the highest buyer rate with %9.18. When we examine the default, housing and loan features. We can see that when people have no debt, their buyer rates are high.

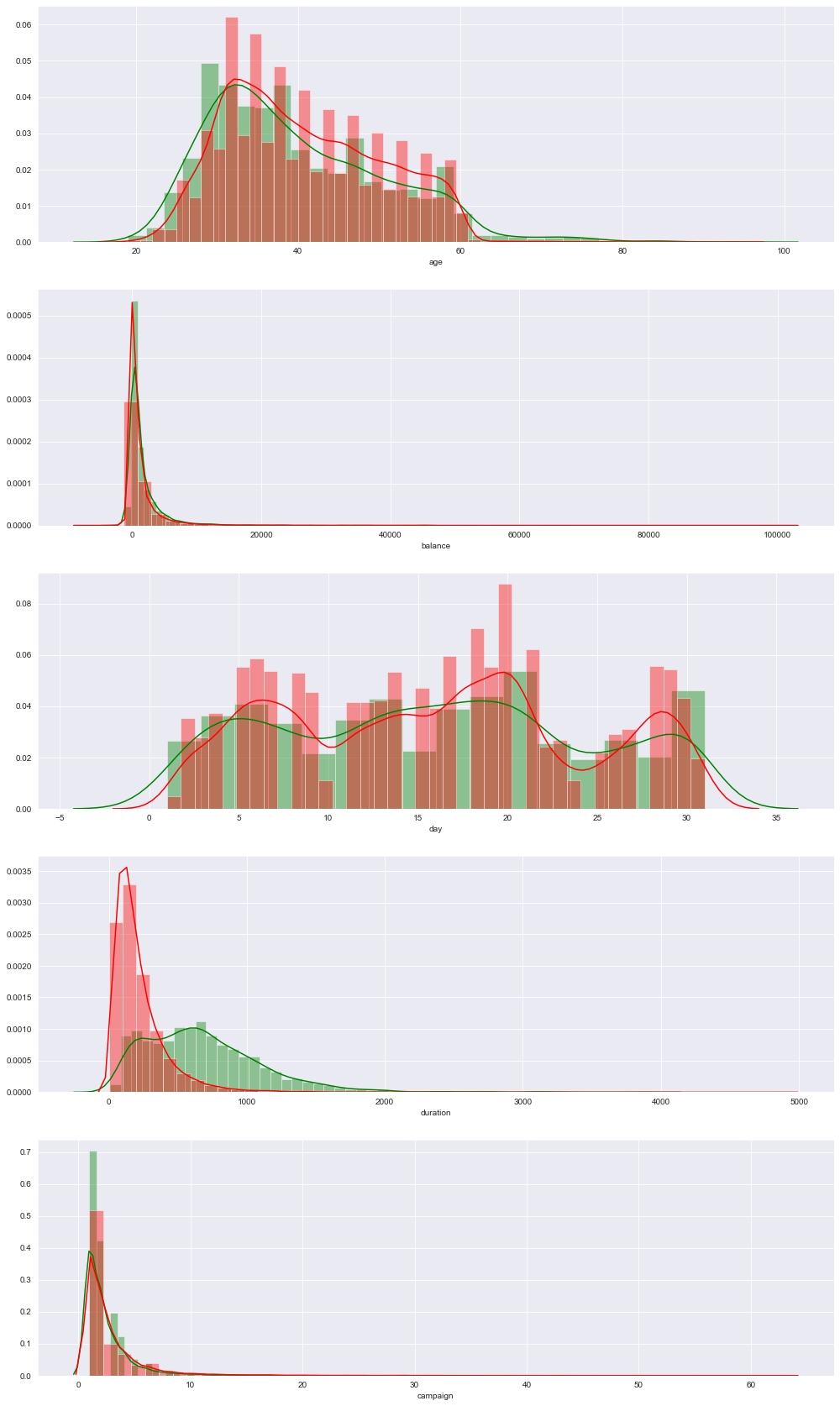


Figure 7:Value counts for numerical features with target value hue

When we examine Figure 7, under 30 years of age most of customers bought the term deposit. The probability distribution of age feature is close to normal distribution and it makes our model‘s learning easier. Most of the customers who have balance under 0 didn’t buy the deposit. When we examine loan, default and housing features, we found similar insight. Most customers purchase the product when the contact time exceeds 400 seconds. This significantly changes the customers' decision. In addition to these explanations, campaign number between 0 and 6-7 affect the buy rate positively.

Before the model was established, 5 different algorithms were tested and the best was preferred and hyper parameter editing was made. These algorithms are Random Forest, AdaBoost, Gradient Boost, Bagging Classifier and XGBoost. Preferred algorithm is XGBoost.

According to given data and hyperparameters, accuracy of 5-fold cross validation is 0.93. However, when modelling with imbalanced dataset, it is not correct to consider only accuracy score. Therefore, we use precision, recall, f1 score and auc score. With the recall score, it is observed how many of the actually positive classes were correctly predicted. With this score, the authenticity of the predictions made according to the given features is determined. It is observed how many of the classes predicted are actually positive positively with the precision score. Thus, the cost of false estimates is also considered. The f1 score was used to consider two scores at the same time.

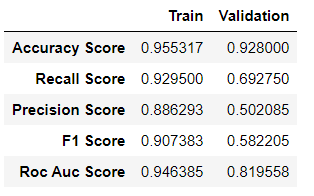


Figure 8: Metrics for XGBoost model

Bonuses:

1. According to exploratory data analysis, we can determine the segment of customers like:

AGE: 20 – 30

JOB: Student

MARITAL: Single

EDUCATION: Tertiary

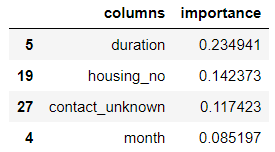
BALANCE: Higher than 0 or Equal to 0

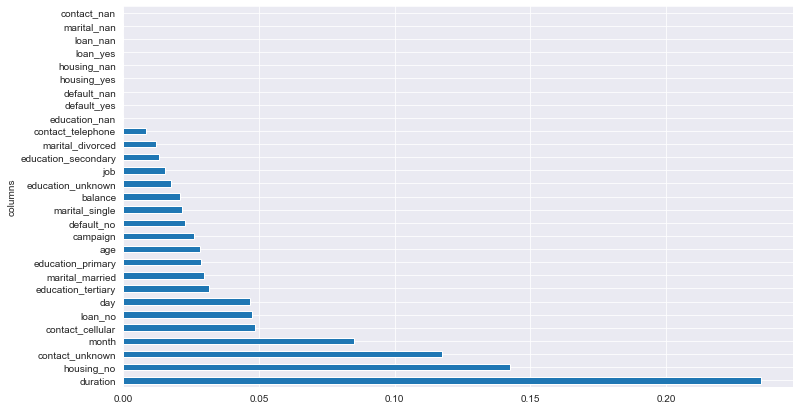
DEFAULT: No

LOAN: No

HOUSING: No

1. Feature importances according to model:





Last contact duration is highly effective over decision. When contact duration is higher than 400, most of the customers buy the product.