Data Analysis and ML Model building for bank term deposit

**Abstract**

My project's goal is to extract insights from the dataset and build a machine learning model on top of it. The data I gathered pertains to the bank's marketing campaigns. In order to carry out these efforts, phone calls were taken into consideration. It is sometimes necessary to contact a customer multiple times in order to make a campaign successful. We will investigate a dataset that will aid in making successful business decisions to reduce costs and enhance revenue, and we will use Machine Learning models to forecast whether or not a certain consumer will subscribe to a term deposit.

**Design**

The dataset, is publicly available on the Kaggle.com website, is about term deposit marketing efforts conducted by Portuguese banking institutions. Our goal is to discover relevant information such as if age has an effect on subscription or not, and whether bank balance has an effect on subscription. By uncovering these valuable insights, marketing budgets may be used more effectively and effectively target interested audiences, and Machine Learning models can be used to predict whether or not a certain person will subscribe to a term deposit.

**Data**

There are 9500 records in the dataset, with 17 columns/features. Some of the dataset's columns are categorical, while others are numerical. In comparison to those who subscribed, there were more people who did not sign up for a term loan.

**Balancing Dataset:**

As the dataset contains more number of samples for class no. I used smote an oversampling technique to balance the dataset. After sampling both the classes have equal number of samples.

**Feature Selection:**

To find the most effective features which are helpful in order to predict whether customer will subscribe or not term deposit. I used ExtraTressClassifer.

**Model Building and selection:**

For my classification task I used five classification models. K nearest neighbors, Decision Tree, Logistic Regression, Naïve Bayes and SVM. I used k-fold cross validation technique in order to select optimal model providing highest accuracy. I further improved model accuracy using hyper parameter tuning.

1. **Tools:**

Pandas: I used pandas to load the set

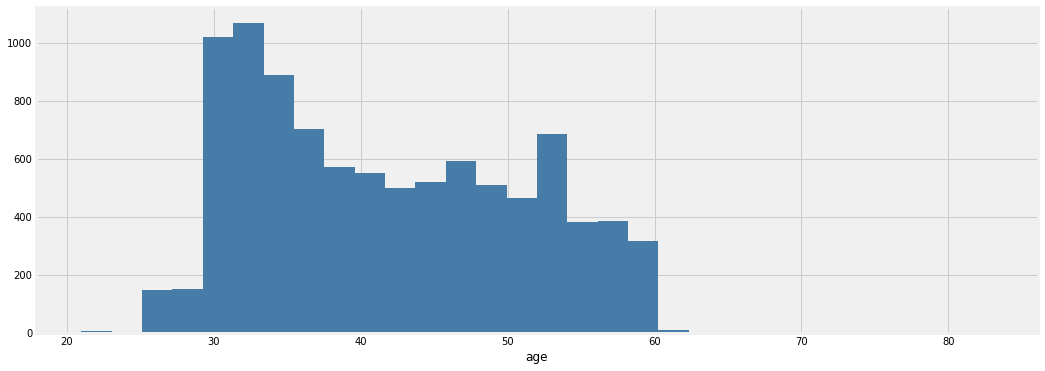
Seaborn: I used seaborn in order to visualize the dataset

Imblearn for balancing unbalanced data

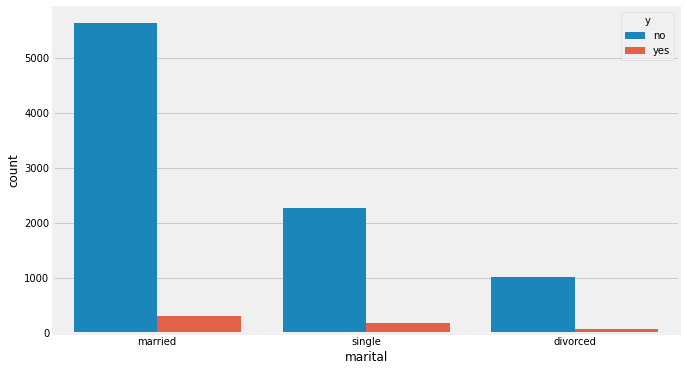
Sklearn: sklearn is used in model building and validation

1. **Communication:**

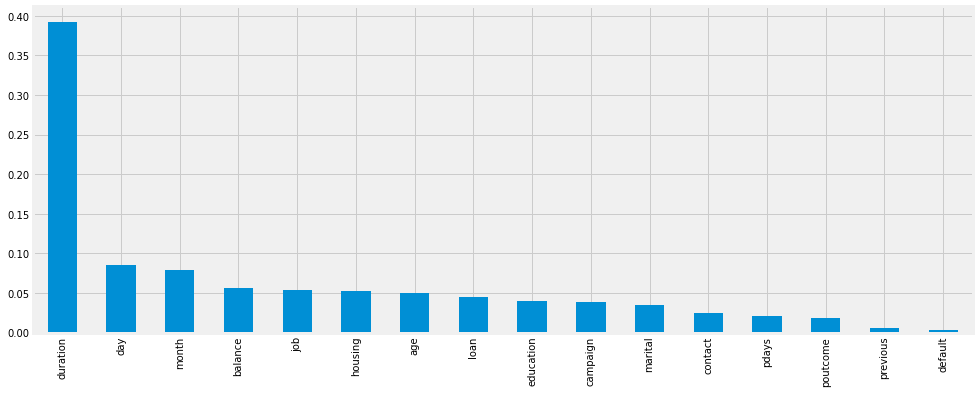
Age distribution of customer



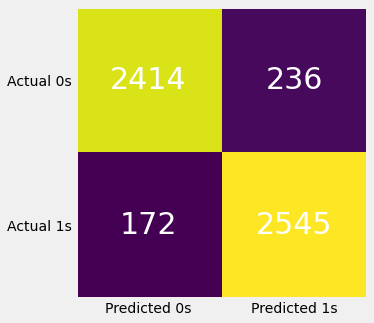
Number of people by subscription and marital status

****

Features by importance

****

Confusion matrix of model



Pair plot of numerical features

