CAPSTONE PROJECT – PREDICTING CUSTOMER TERM DEPOSIT

This is a modified version of the marketing dataset posted on UCI repository by a Portuguese bank. The competition is related to a direct marketing campaign where the goal is to predict if a client will subscribe a term deposit or not.

Part 1 – Data Wrangling

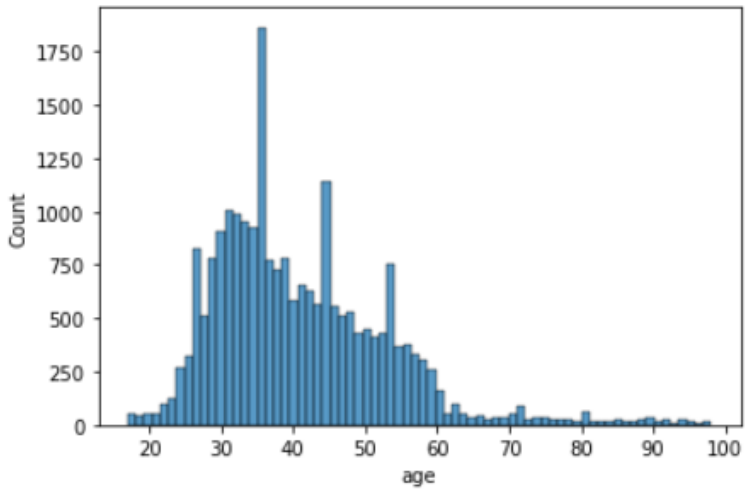
The initial data has 22500 rows X 22 columns. There were several categorical columns eg 'married, 'housing', 'loan' etc others were integer or float eg, age, employment variation rate etc. Initially there were no "null" values. There were three columns with "yes" and "no" values that were converted to "1" and "0". During conversion, some "unknown" values in those columns were converted to null and thus were dropped. The correlation matrix was first calculated. It seemed the features that are most relevant are employment variation rate, euribor3m and number of employees (in that order)

	subscribe
RecordID	0.268302
age	0.093892
default	0.160437
housing	0.004047
loan	0.023371
duration	0.037470
campaign	0.165172
pdays	-0.100082
previous	0.032854
emp.var.rate	-0.270444
cons.price.idx	-0.074710
cons.conf.idx	0.026914
euribor3m	-0.181128
nr.employed	-0.176219
subscribe	1.000000

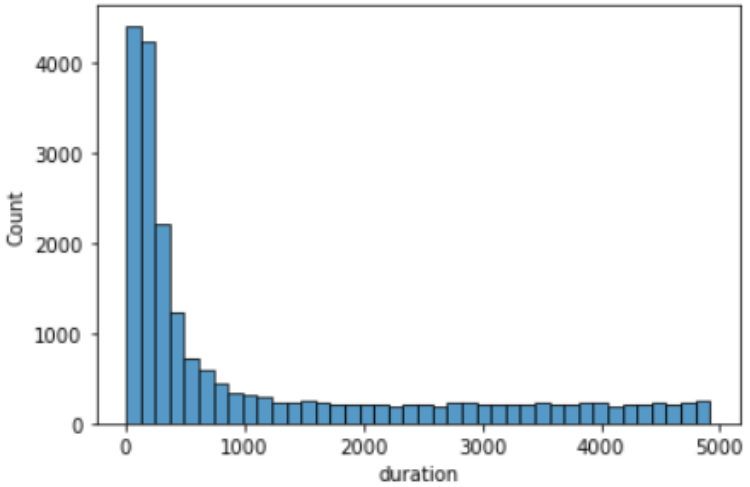
The describe function was called and it was found that columns 'duration' and 'pdays' had high maximum values that need to be normalized.

A check on duplicate values on "RecordID" revealed no duplicates.

Several features were plotted to check their distribution. Except for "age", other features were not gaussian. Hence, The outliers from "age" column were removed using quantiles.

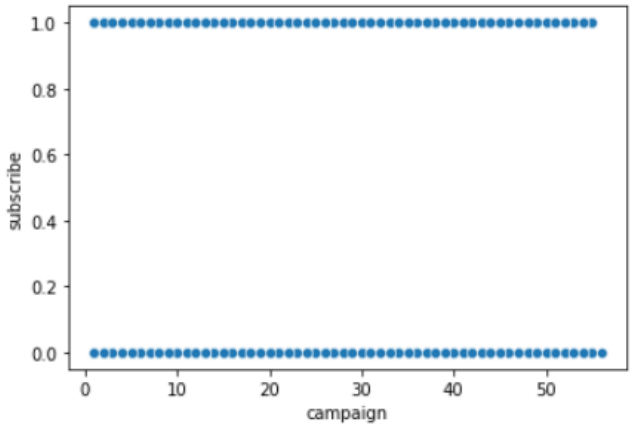


Age followed a gaussian distribution; hence outlier detection was carried out.



Duration showed 1-tail gaussian distribution; hence outlier detection was not carried out.

A check was done to see if features had subscribers/non-subscribers (both 1/0) consistently along low and high values.



Consistent 1/0 subscribers through all values of "campaign"

Part 2 – Pre-processing

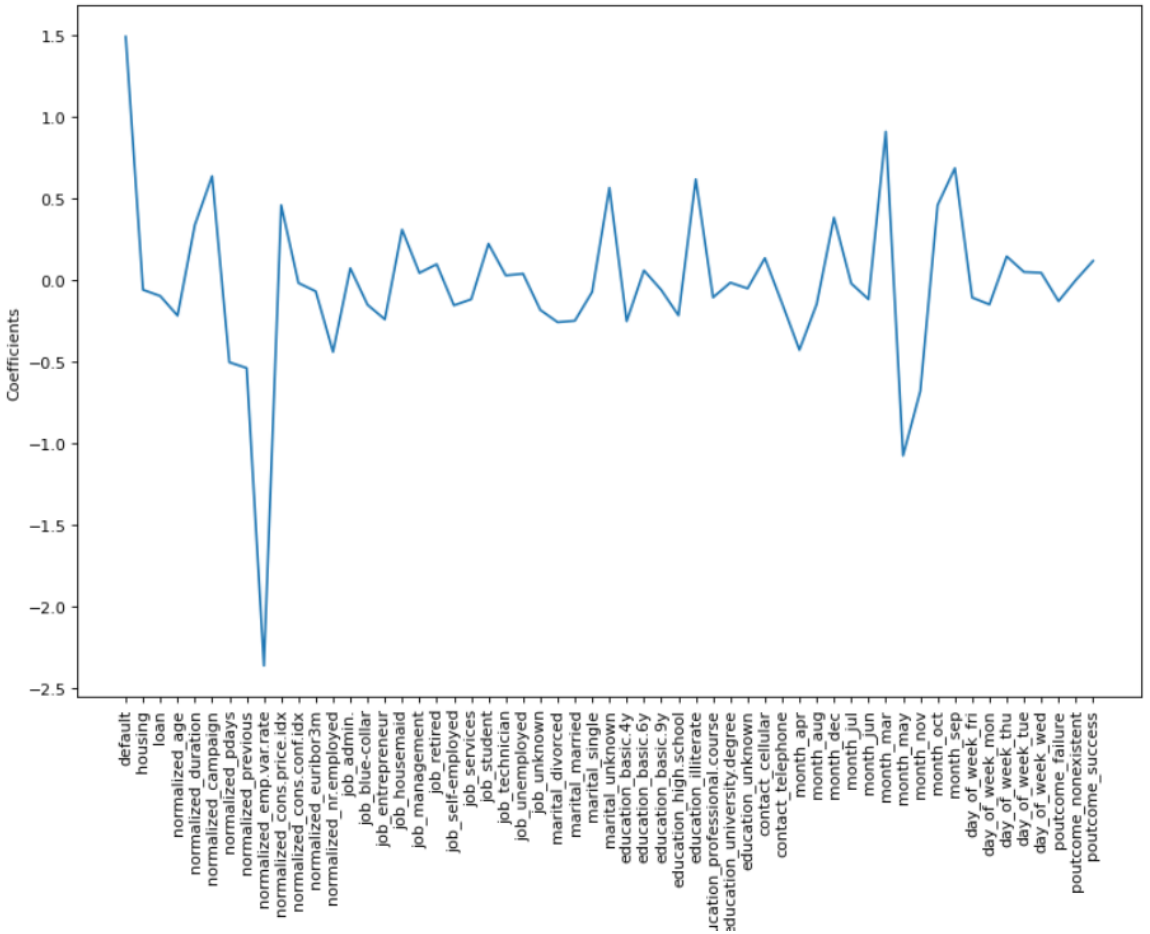
First the RecordID was dropped from the columns and the target label "subscribers" was stored as a separate data frame. The y target label "subscribe" was dropped from the main dataframe and stored in a separate dataframe. A standardization was applied to all numeric values that have max values greater than 1. Since the distribution on most features are non-gaussian, a MinMaxScaler was applied to standardize the features.
The module get_dummies was called to OneHotEncoding all target variables in order to model the categorical features. The number of columns increased to 57 from 20 after OneHotEncoding. Before running several models, a train-test split was carried out on training data with 70/30 split.

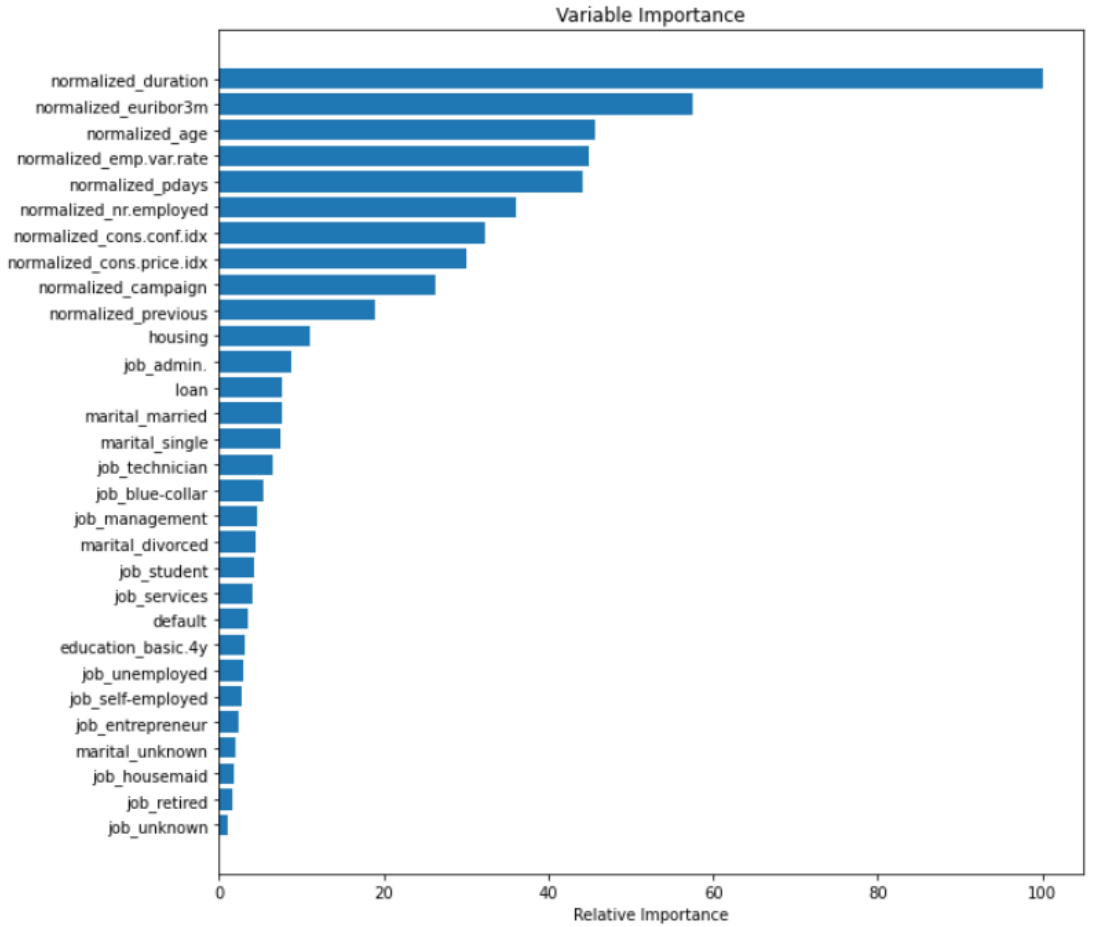
Part 3 - Modeling

Since there is a mix of categorical and numerical features, we tried using logistic regression, decision tree and random forest models to get the best Accuracy. A Gradient Boost model was also included in the mix.
Hyperparameter tuning was carried out for the models and the best parameters were selected which were used to predict the test values.

Model	Accuracy (before tuning)	Accuracy (after tuning)
Logistic Regression	*0.876*	*0.877*
Decision Tree	*0.851*	*0.884*
Random Forest	*0.892*	*0.892*
Gradient Boost Model	*0.860*	*0.897*

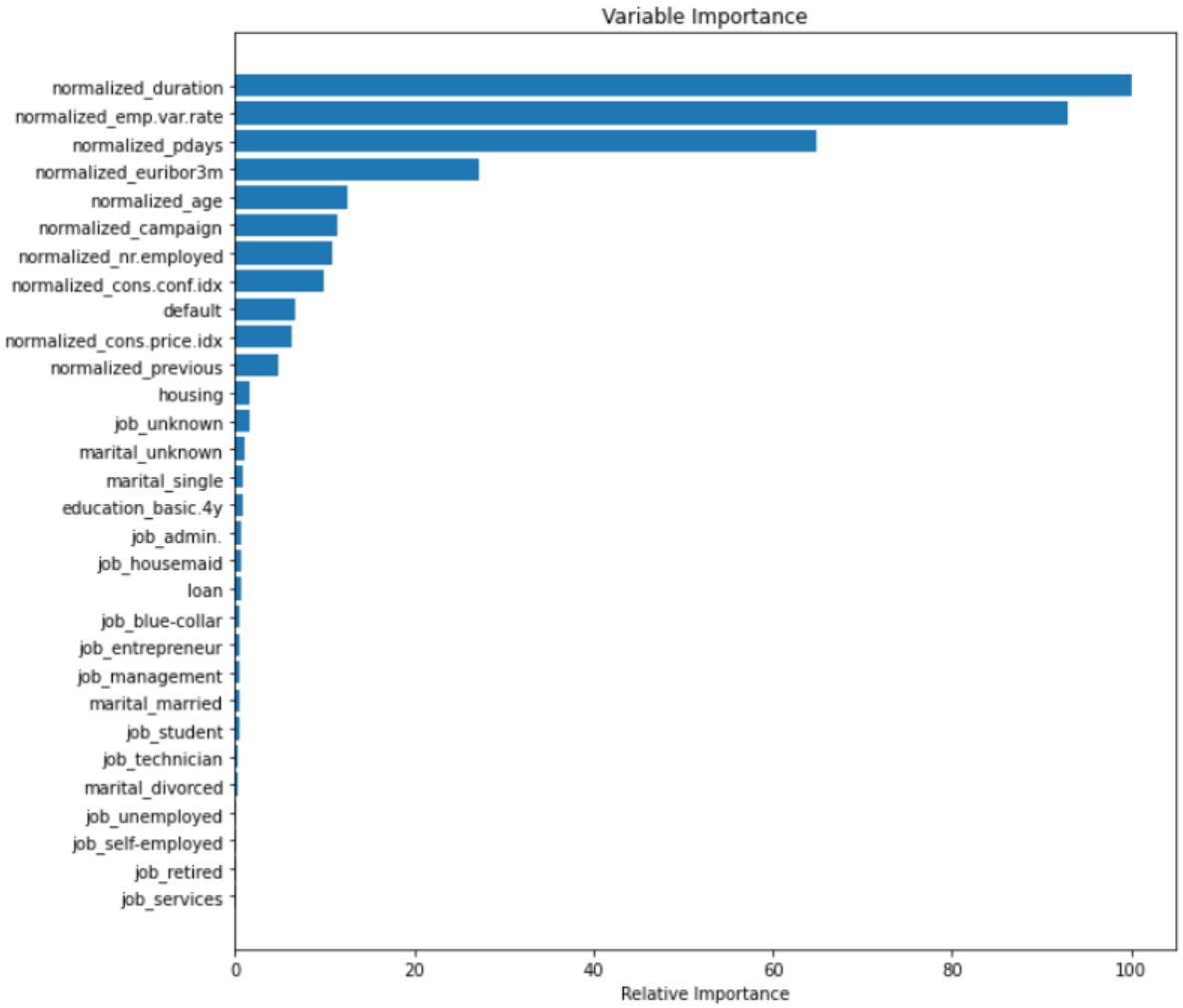
The feature importances were displayed for both logistic regression and random forest classifier:

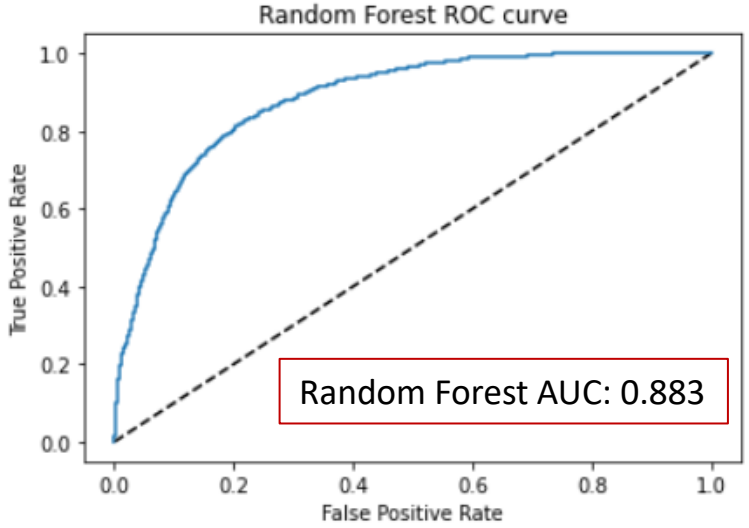


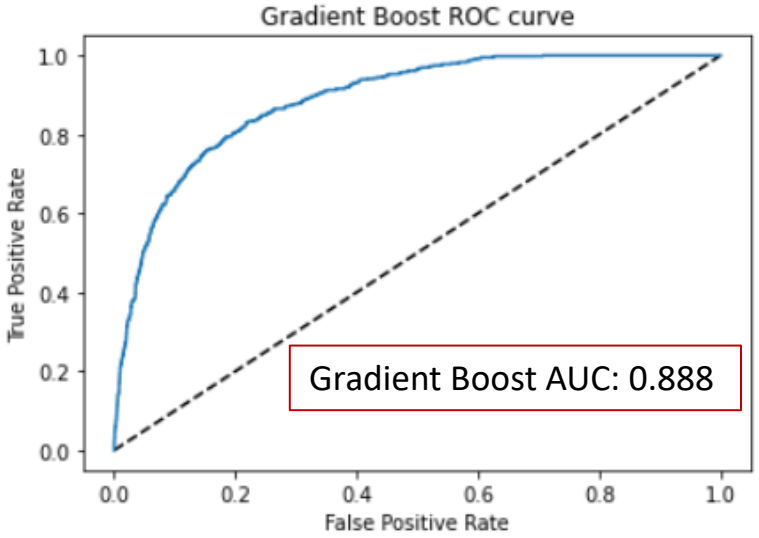


It seems like the feature importances from the random classifier model to an extent match with the logistic regression coefficients. Both models seem to agree that features like "duration", "employment var. rate", "consumer price index", "married" are important. However, logistic regression Gives high importance to "default" whereas random forest classifier does not.

The feature importances were displayed for Gradient Boost Classifier. The ROC and AUC was plotted and calculated for Random forest Classifier and Gradient Boost Model.

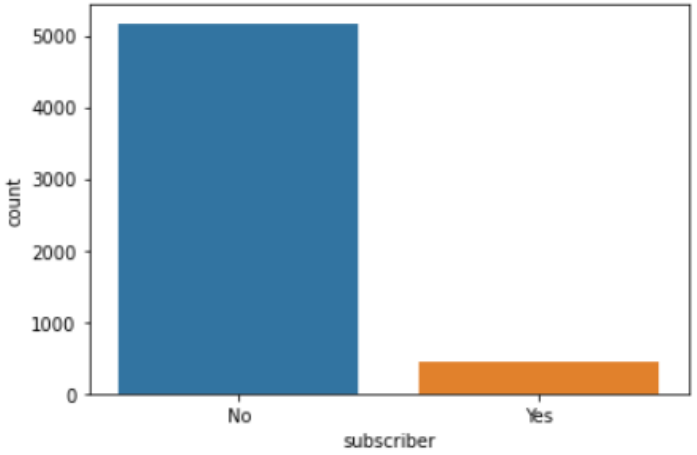


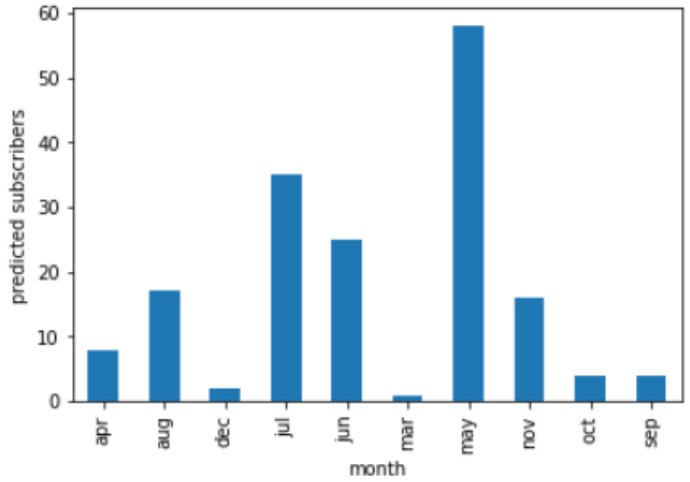


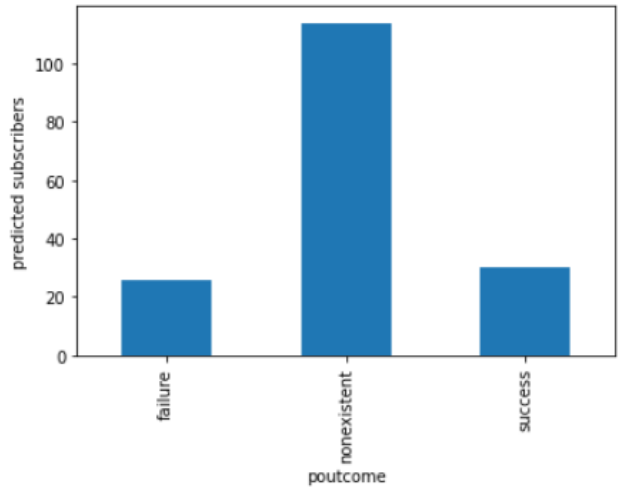


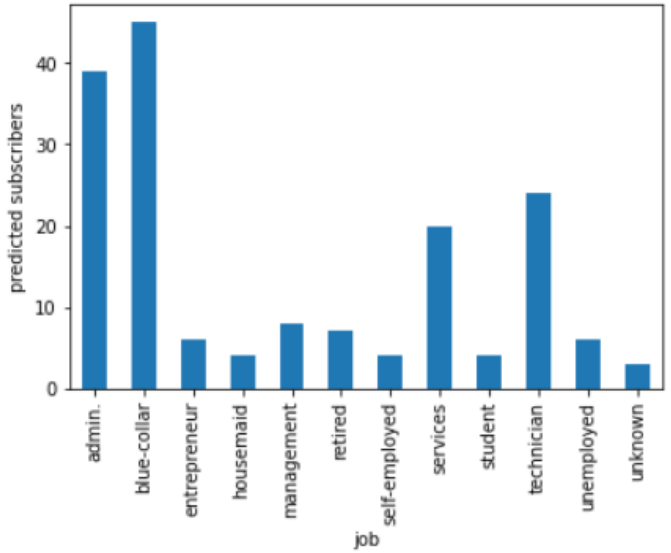
Part 4- Model testing and Visualization

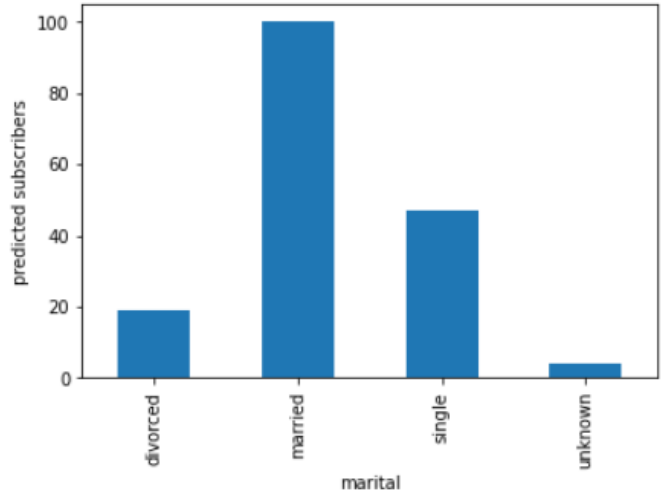
The Gradient Boost Classifier Model gave the highest accuracy of 0.897 on training data was used to model the test data. The test data was first wrangled and pre-processed before running the model. The model predicted 456 candidates out of 5626 people to subscribe to term deposits. This is slightly more than 8% of the total candidates. It seems customers that were contacted in 'May" had the largest subscribers. Other characteristics that contribute most to subscription in each group are "married", "blue-color" job holder among all job holders and "university degree" in education.











Predicted subscribers among various categories.