أكاديمية سدايا
SDAIA Academy

Classification Project
Term Deposit Subscription

Elyas Maghrabi

Abstract

The data is related with direct marketing campaigns (phone calls) of a Portuguese banking institution.

Classification goal

The classification goal is to predict if the client will subscribe a term deposit (variable y).

Objective

The objective is to analyze dataset based on several variables and create a classification algorithm.

Dataset

The dataset was collected from UCI Machine Learning Repository:

http://archive.ics.uci.edu/ml/datasets/Bank+Marketing#

EDA Phase:

Pandas

Matplotlib

NumPy

Machine Learning Phase:

Sklearn

UCI - Bank Marketing Data Set

Dataset contains:

- We have 41188 instances and 21 features..
- y - has the client subscribed a term deposit (yes, no) target.

Bank Client Data:

Age, Job, Marital, Education, Default, Housing, Loan

Last contact / Campaign:

Contact, Month, Day_of_week, Duration, Campaign, Pdays, Previous, Poutcome

Social and economics:

Emp.Var.Rate, Cons.Price.Idx, Cons.Conf.Idx, Euribor3m, Nr.Employed

Target outcome:

Y - Subscribe term deposit or not

The problem

Preprocessing

- How to deal with missing values
- How to encode categorical variables
- Imbalance Target

What features?

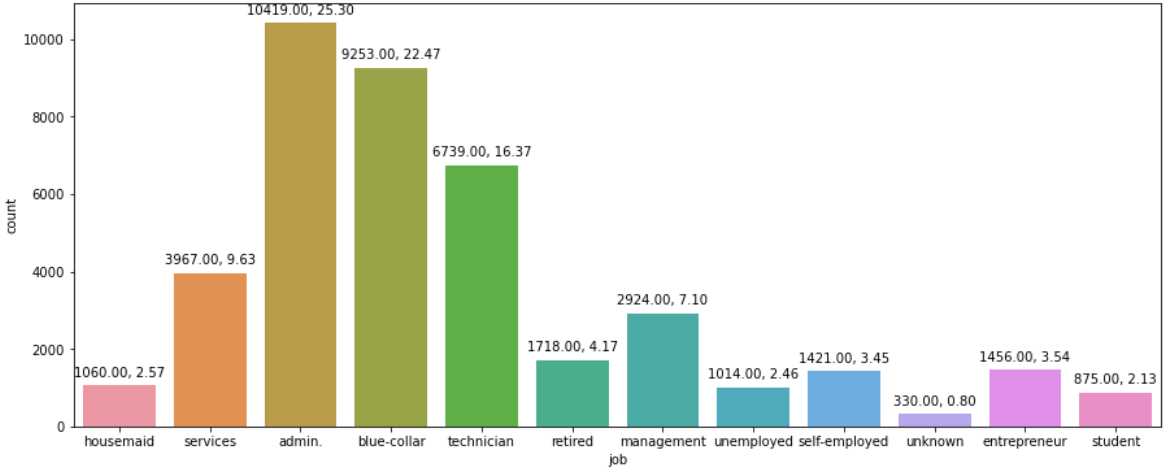
- What features are important to get customers subscribe in the term deposit.

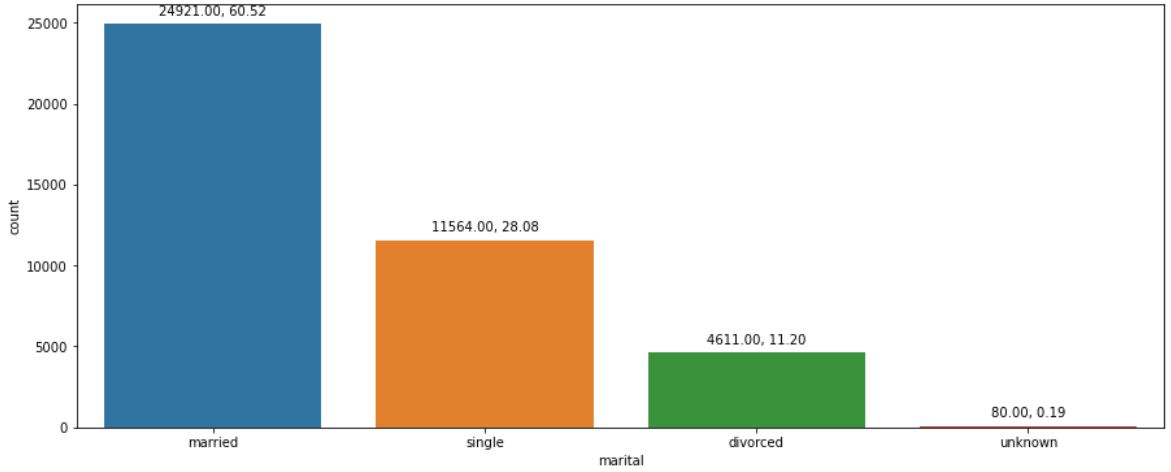
Prediction Model

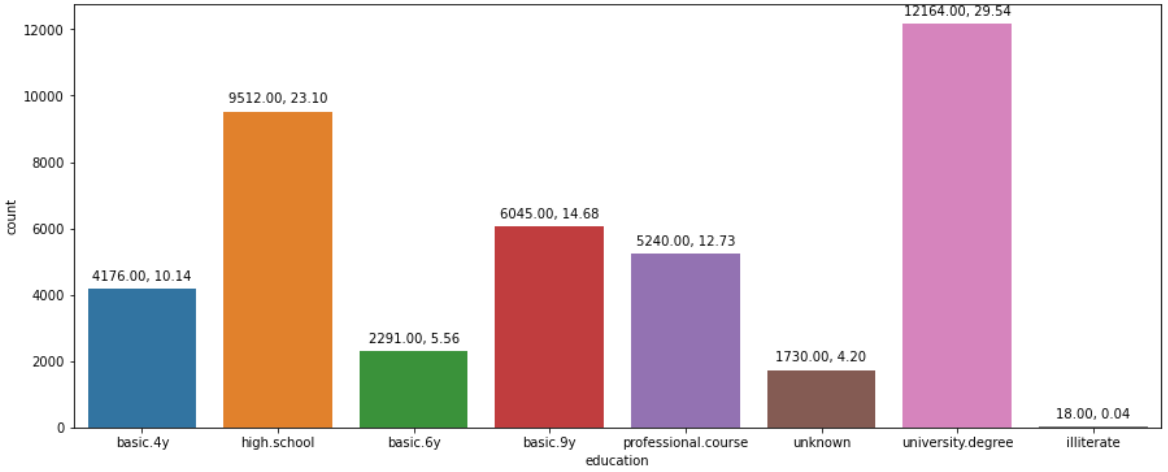
- Try to build a model to predict whether customers will subscribe for term deposit.

Exploratory Data Analysis

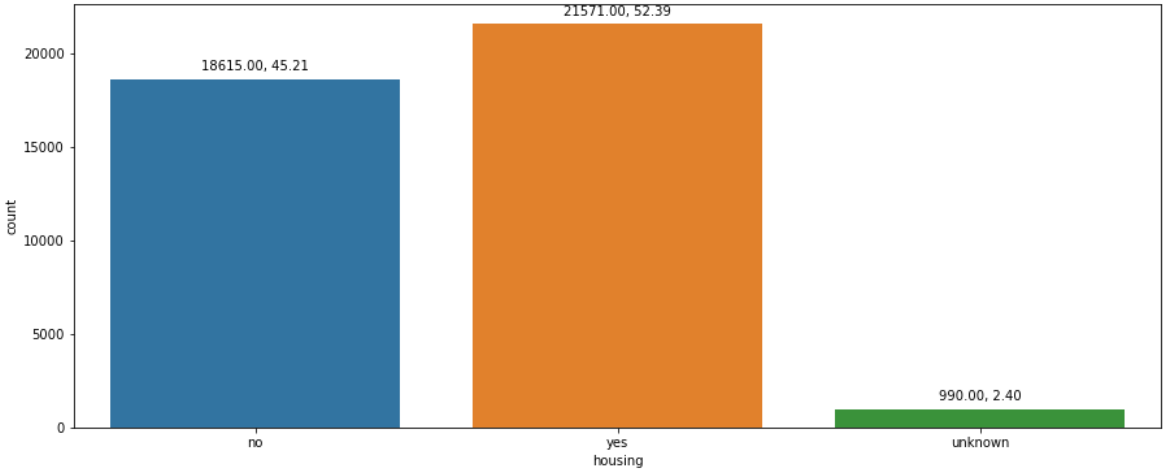
Data Analysis: Job, Marital

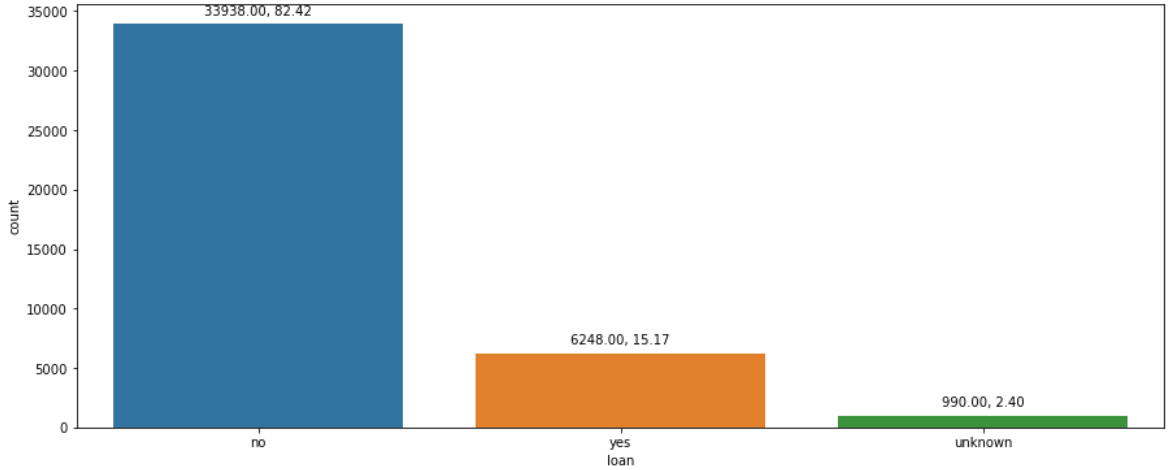


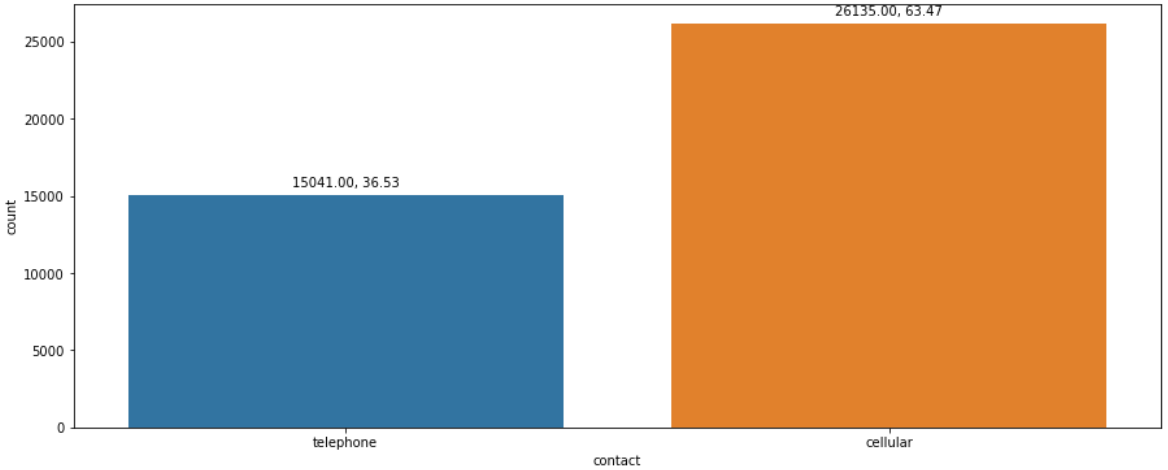




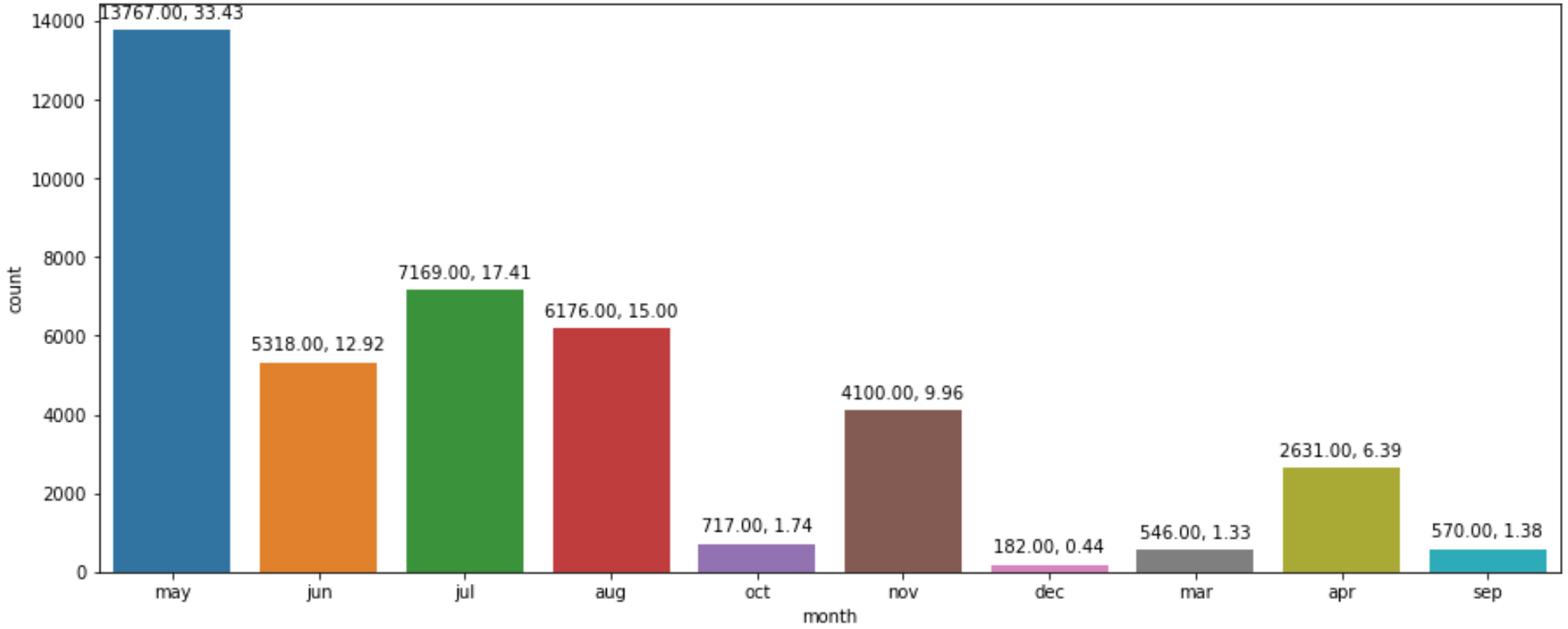
Data Analysis: Housing, Loan, Contact







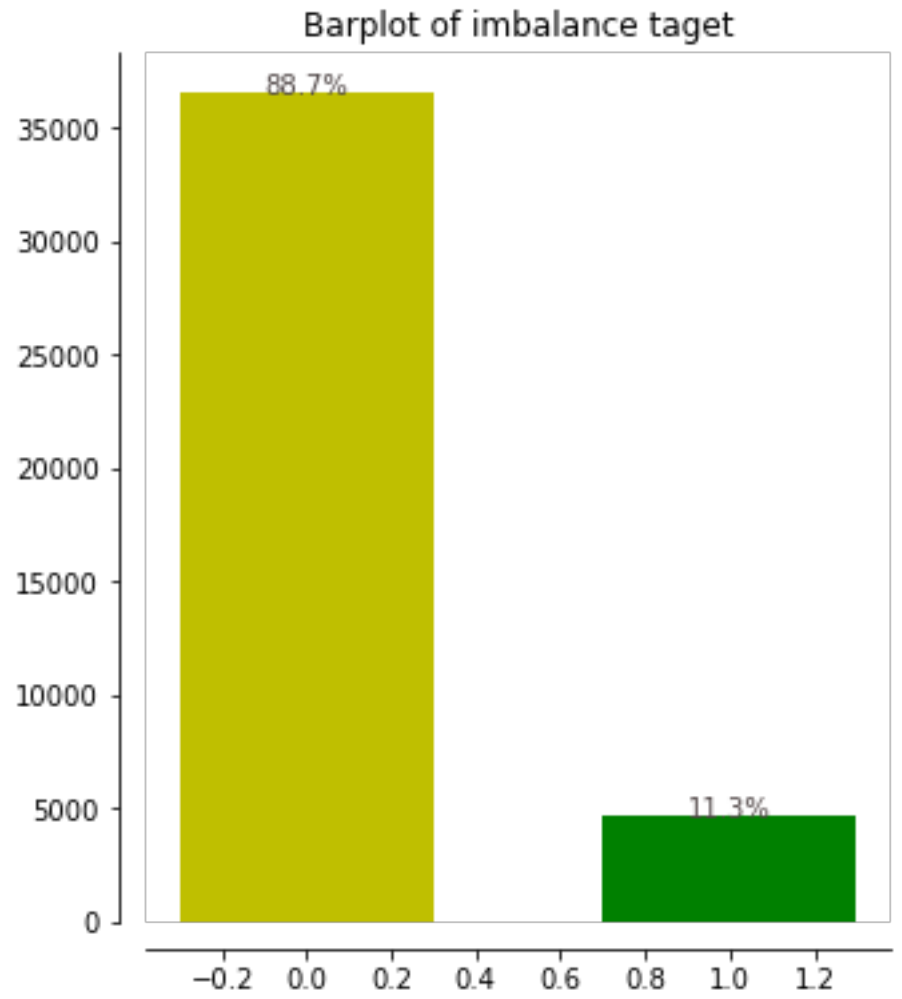
Data Analysis
Month



Imbalance Target

11% to 88%

(User class_weight = “balanced”)



Preprocessing

Categorical

- By description, all the unknown or missing values are represented as “unknown” in all the categorical values. And since most of them are not ordinal(nominal), therefore we would like to oneHotEncode them.

Numerical

- There are almost no missing value in the numerical columns, but one weird columns “pdays”, which is the number of days since last reach out, indicate no call before as 999. Therefore, we need to replace 999 with 0.

Modelling

Three Models

Logistic Regression

Baseline

The logistic regression performance is not very well, but its simplicity make it work very well as the baseline.

Random Forest

RandomizedSearchCV

Using random grid search , we can search for the best hyperparameters that goes into the model.

Ensemble model

Voting Classifier

With ensemble model of knn, decision tree, and svc, find the best combination of the model with their vote.

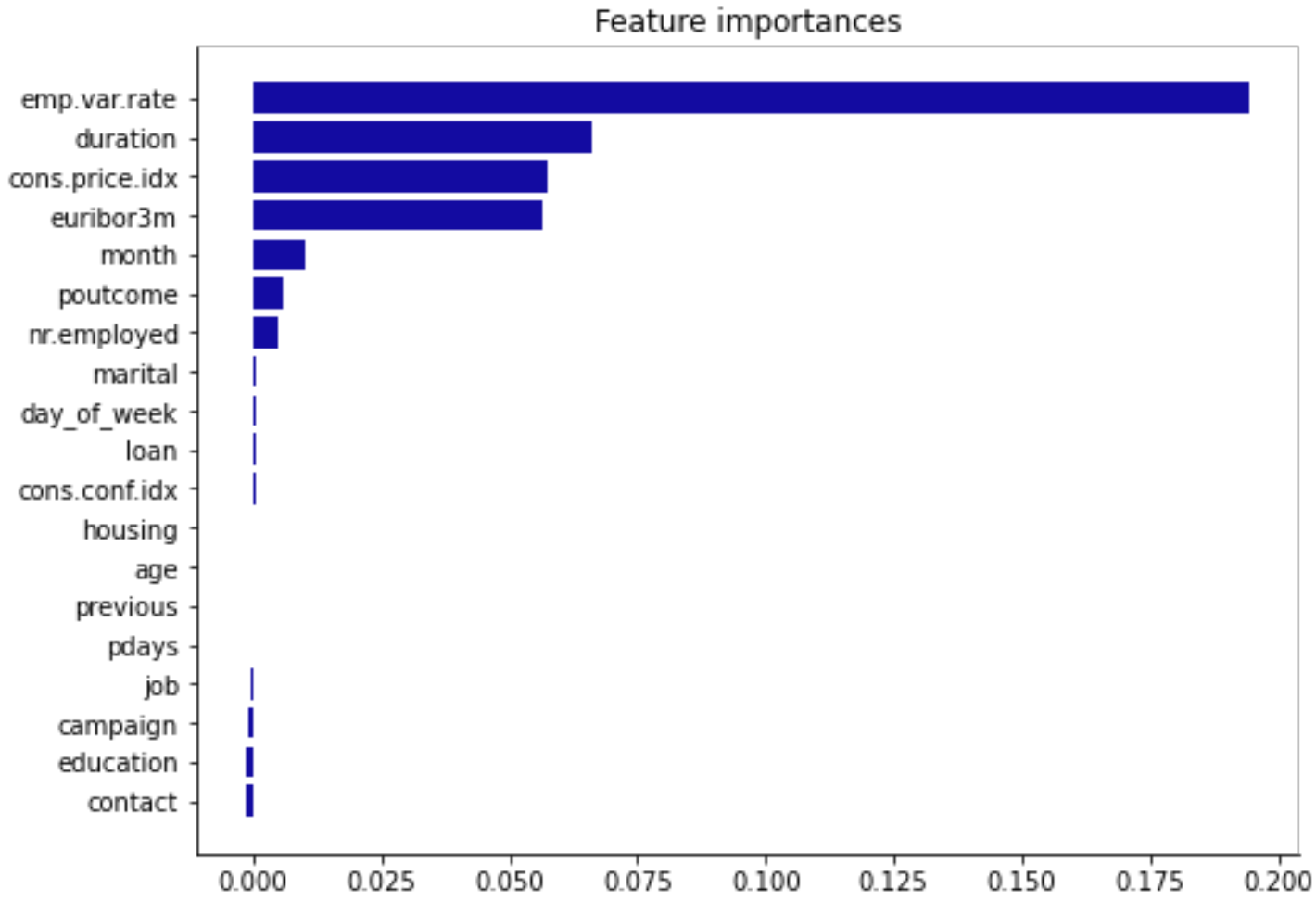
Metrics

Models		Validation score	F1 Score	Confusion Matrix
1	Logistic Regression	0.85	0.88	[[5838 1006] [100 777]]
2	Random Forest	0.90	0.88	[[6773, 71], [666, 211]]
3	Ensemble model	0.91	0.90	[[6642, 202], [482, 395]]

Most important features:

- Employment variation rate
- Duration
- Consumer price index
- Euribor 3 month rate

They are all social and economics and campaign related features.



Thank You