Case Study

By: Manoj Venkatachalaiah

Contents

EDA

1. Data Description

The dataset has 20 independent features and one dependent feature (target variable). The 20 dependent features and further classified into:

- **Numerical**: ['age', 'campaign', 'pdays','previous', 'emp.var.rate', 'cons.price.idx', 'cons.conf.idx', 'euribor3m', 'nr.employed']
- **Ordinal**: ['job', 'education']
- **Nominal**: ['marital', 'default', 'housing', 'loan', 'contact', 'month', 'day_of_week', 'poutcome']
- **Target**= 'y'

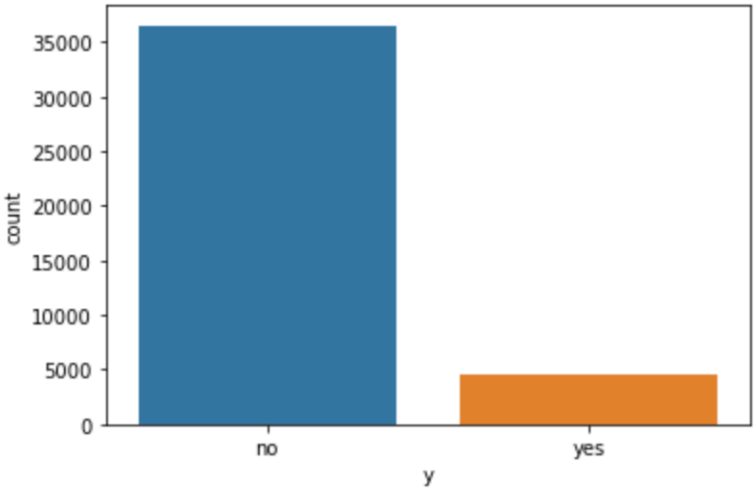
Note: The 'duration' feature was dropped before any of the future steps.

2. Preprocessing- filling missing values with mean imputation

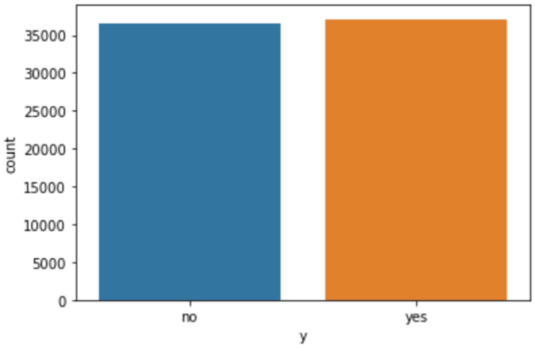
'age' and 'cons.price.idx' features had missing values in the dataset. The missing values have been filled with the **mean** of the respected columns.

3. Inspecting target variable

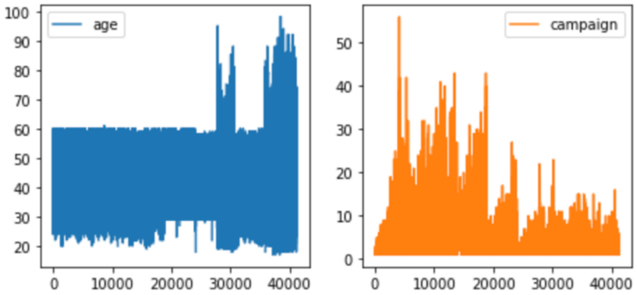
The dataset is heavily imbalanced. The ratio of 'no' to 'yes' in the target variable is 7:1.

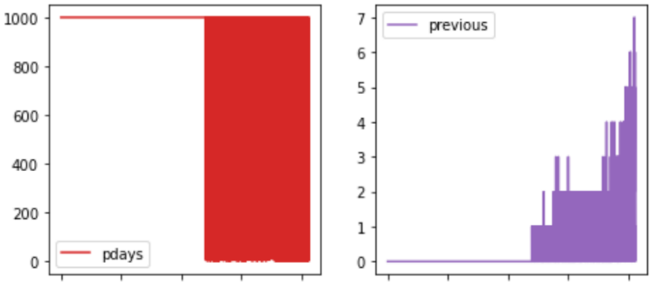


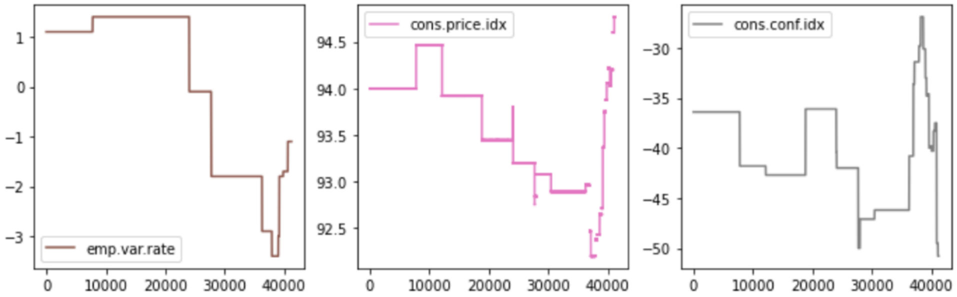
Hence, up sampling was done to balance the dataset. Below is the result after the dataset was balanced.

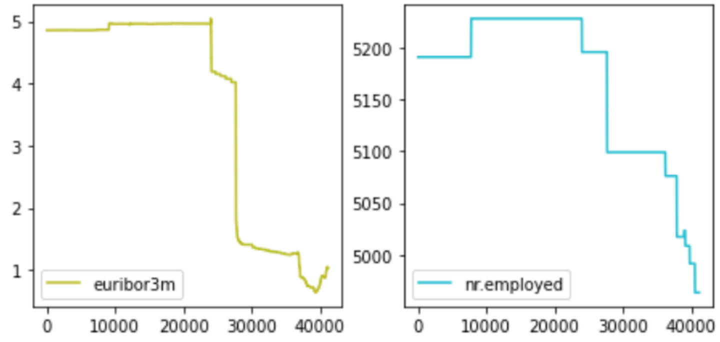


4. Inspecting numerical features







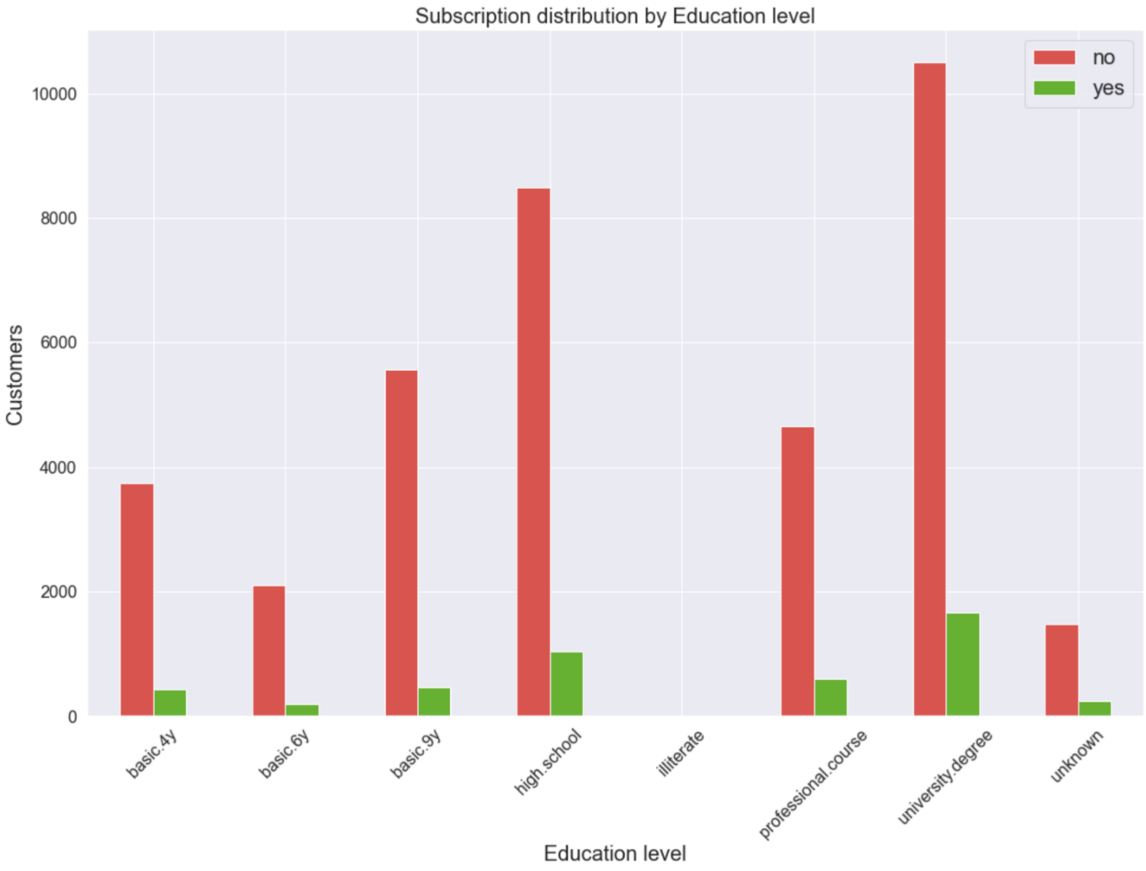


It’s clear that none of the numerical features have a normal distribution. Hence, the dataset will be standardized **after** train/test split.

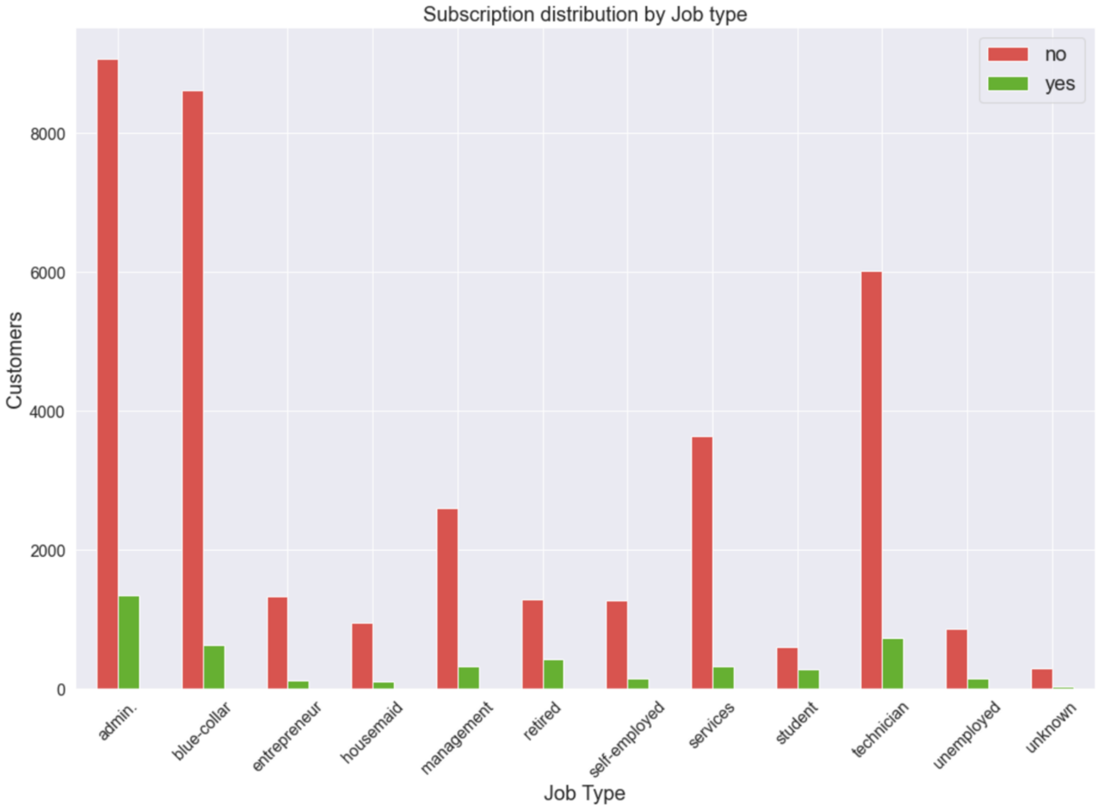
5. Inspecting categorical features

The categorical features that had the most impact on the target variable were ‘education’, ‘job’ and ‘marital.

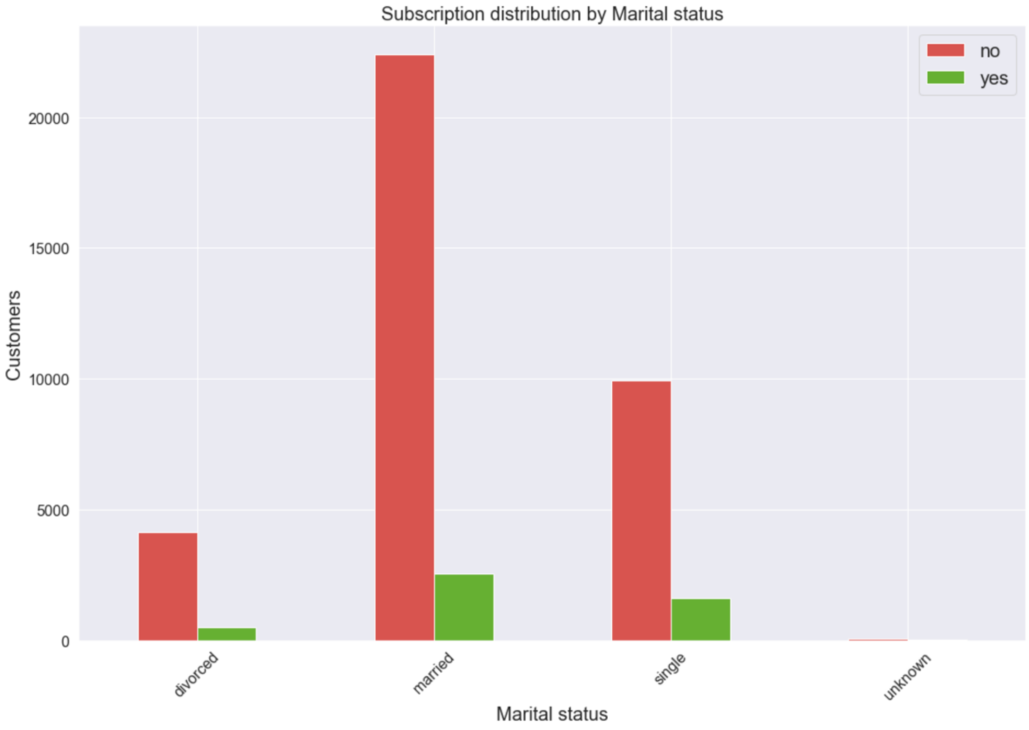
Education level:



Job type:



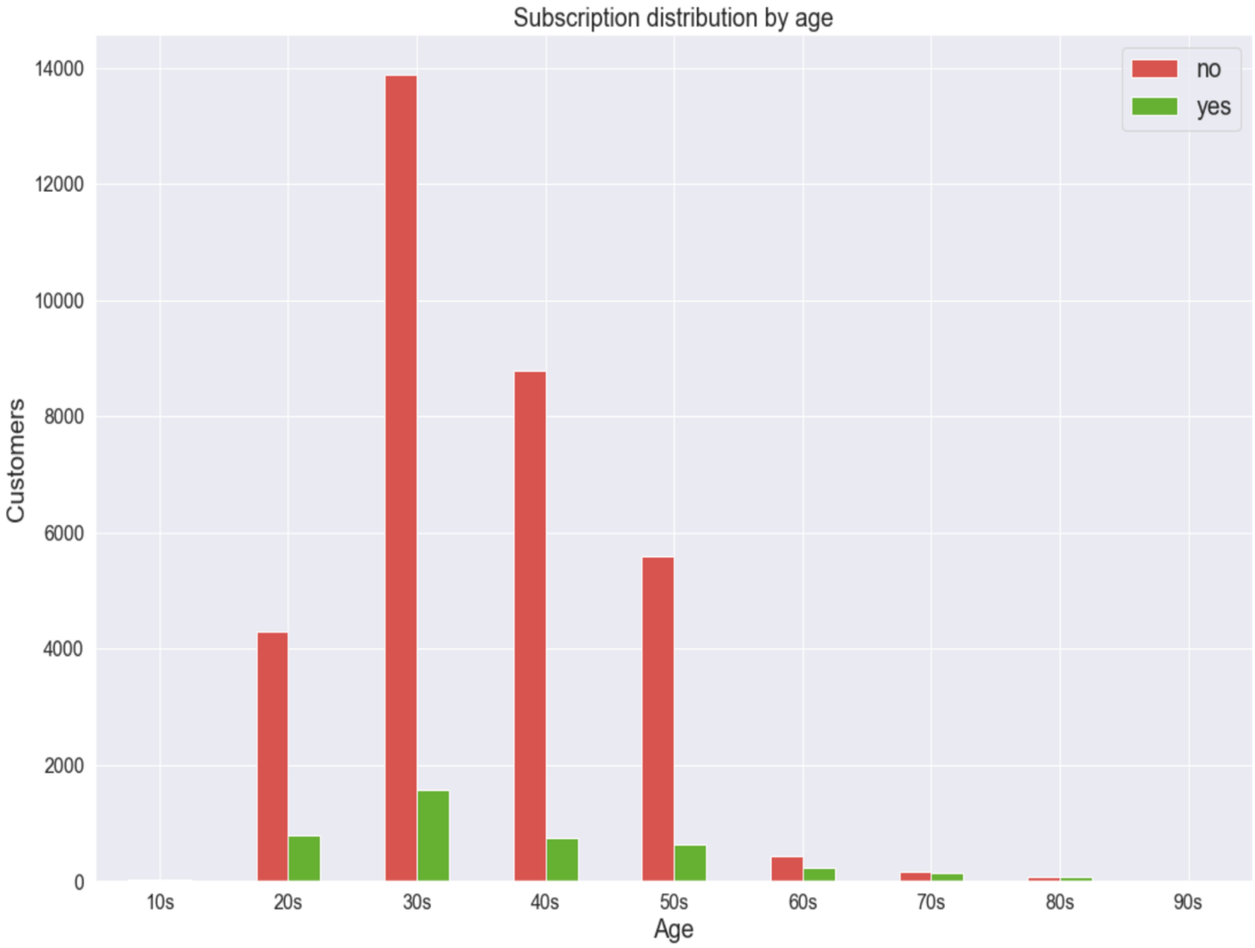
Marital status:



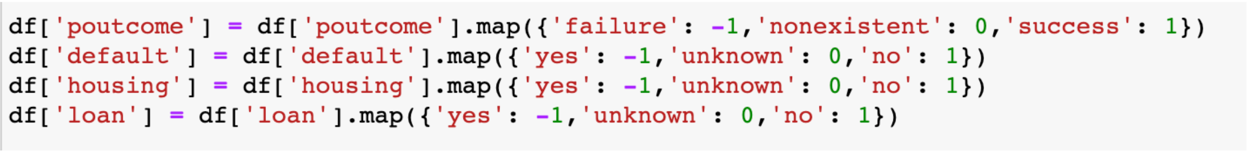
Distribution of other categorical features can be found on the notebook in the github repository.

6. Feature Engineering

Upon inspection, it was found that age had a big impact on the target variable. Hence, ‘age’ was converted into a categorical feature by placing the ages in bins of range 10.

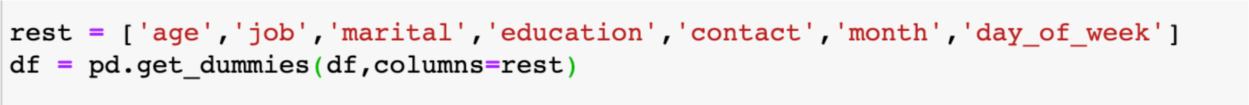


‘poutcome’, ‘default’, ‘housing’ and ‘loan’ were label encoded,



```
df['poutcome'] = df['poutcome'].map({'failure': -1,'nonexistent': 0,'success': 1})
df['default'] = df['default'].map({'yes': -1,'unknown': 0,'no': 1})
df['housing'] = df['housing'].map({'yes': -1,'unknown': 0,'no': 1})
df['loan'] = df['loan'].map({'yes': -1,'unknown': 0,'no': 1})
```

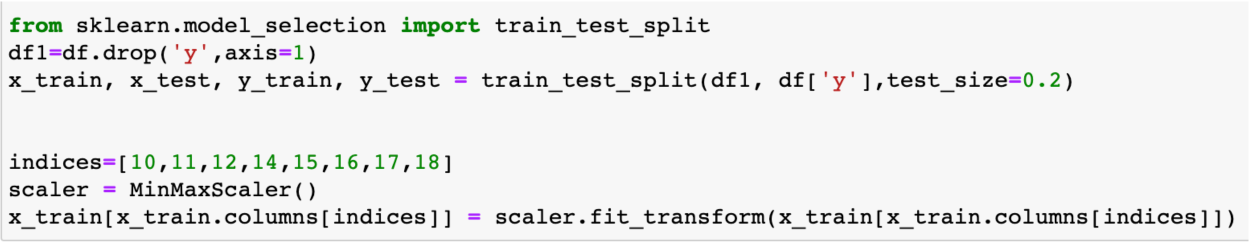
Whereas ‘age’ and the rest of the categorical features were one-hot-encoded,



```
rest = ['age','job','marital','education','contact','month','day_of_week']
df = pd.get_dummies(df,columns=rest)
```

7. Splitting into train and test, and Feature Scaling of train

The dataset was first split into training and testing sets. After which the training set was standardized using the MinMaxScalar()



```
from sklearn.model_selection import train_test_split
df1=df.drop('y',axis=1)
x_train, x_test, y_train, y_test = train_test_split(df1, df['y'],test_size=0.2)


indices=[10,11,12,14,15,16,17,18]
scaler = MinMaxScaler()
x_train[x_train.columns[indices]] = scaler.fit_transform(x_train[x_train.columns[indices]])
```

8. Final dataset description

The dataset used for modeling has 62 dependent features and 1 dependent features. Below are the dependent features:

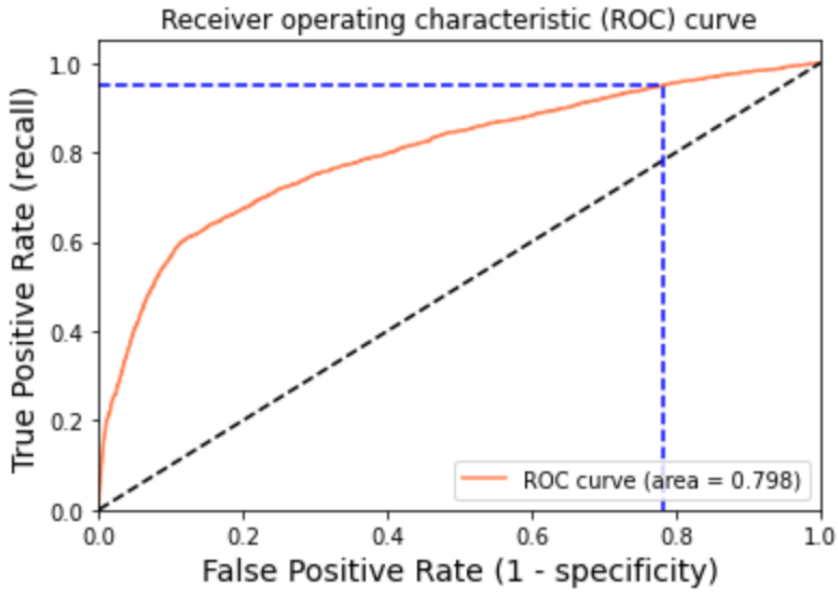
['default', 'housing', 'loan', 'campaign', 'pdays', 'previous',
'poutcome', 'emp.var.rate', 'cons.price.idx', 'cons.conf.idx',
'euribor3m', 'nr.employed', 'age_10s', 'age_20s', 'age_30s', 'age_40s',
'age_50s', 'age_60s', 'age_70s', 'age_80s', 'age_90s', 'job_admin.',
'job_blue-collar', 'job_entrepreneur', 'job_housemaid',
'job_management', 'job_retired', 'job_self-employed', 'job_services',
'job_student', 'job_technician', 'job_unemployed', 'job_unknown',
'marital_divorced', 'marital_married', 'marital_single',
'marital_unknown', 'education_basic.4y', 'education_basic.6y',
'education_basic.9y', 'education_high.school', 'education_illiterate',
'education_professional.course', 'education_university.degree',
'education_unknown', 'contact_cellular', 'contact_telephone',
'month_apr', 'month_aug', 'month_dec', 'month_jul', 'month_jun',
'month_mar', 'month_may', 'month_nov', 'month_oct', 'month_sep',
'day_of_week_fri', 'day_of_week_mon', 'day_of_week_thu',
'day_of_week_tue', 'day_of_week_wed']

Modeling

Note: Accuracy is not the best metric in this particular case study, because what matters more is the Recall the Roc_Auc score which tells us how good the model is at predicting if the customer is going to subscribe to the bank term deposit correctly.

GridsearchCV was used for hyperparameter tuning in each of the 3 cases, the best results are provided below.

Logistic Regression: best results



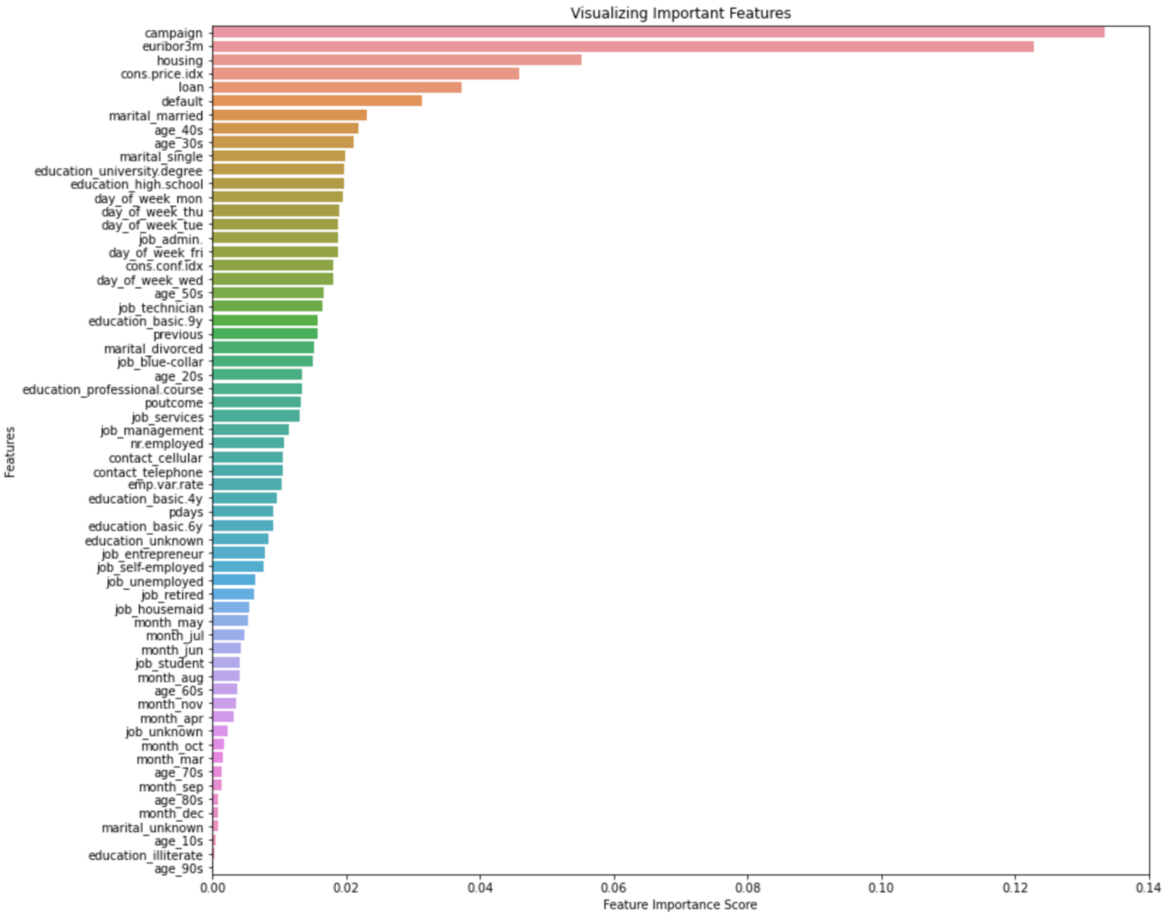
Accuracy: 73%
ROC_AUC score: 79%
Recall: 79%

Recursive Feature Elimination (RFE) was used to select the features for Logistic Regression Modeling. Below are the selected features:

['default', 'previous', 'poutcome', 'emp.var.rate', 'euribor3m', 'nr.employed', 'age_10s', 'age_60s', 'age_70s', 'age_80s', 'job_student', 'marital_unknown', 'contact_cellular', 'month_apr', 'month_aug', 'month_dec', 'month_jul', 'month_jun', 'month_mar', 'month_oct']

Random Forest Classifier: best results

Feature importance was used to select the set of features for RFC. The 6 strongest features were used for modeling.



Results:
Accuracy: 80%
Recall: 78%
ROC_AUC score: 79%

Features used:
['campaign', 'euribor3m', 'housing', 'cons.price.idx', 'loan', 'default']

Support Vector Machine: best results:

kernel: ‘Linear’

Results:
Accuracy: 71%
Recall: 70%
ROC_AUC score: 70%

The model I would prefer using:
Although Logistic regression has its perks when it comes to binary classification, I think ensemble methods such as the Random Forest Classifiers are best suited to this particular case study. The decision trees in the RFC take into consideration the most important features in making predictions. It can also be attributed to the number of dependent features in. this case, which is fairly large. And decision tree classifiers along with Logistic Regression are best suited in such cases.

Business Value of the analysis

The case study gives insights on what kinds of customers the bank should target while looking increase the number of subscriptions. In the EDA section we saw that Employment related features and demographical features such as Age, Job type, Education Level and Marital status had a big impact on the subscription outcome.

It makes sense that employment related features had the most importance in the RFC classification model. Subscribing to a term deposit isn’t financially viable to people not doing well with employment.

Other important features were the loan variables. It makes sense that people who are already associated with the bank via a personal loan or housing loan would be willing to make a term deposit.

For instance, ages between 30 to 50 had the most subscriptions. The bank can use the analysis to narrow down their list of target customers. Similarly, customers with a university degree had the most subscriptions, customers who were married and/or were an admin or a technician had the most subscriptions.

These insights can help the bank target certain customers in a certain way (the analysis also showed that cellular contact had more subscriptions than telephone.