**CONTENTS**

**1. Abstract**

1.1 Problem Statement  
 1.2 Client  
 1.3 Dataset

**2. Data Inspection**

**3. Data Pre-Processing**

**4. Exploratory Data Analysis**

4.1 EDA with numeric variables  
 4.2 EDA with categorical variables   
 4.3 Box plot of all the numeric variables  
 4.4 Effect of classes of response variable with respect to other variables

**5. Feature Engineering**

5.1 Consolidate category classes  
 5.2 Binning ‘age’   
 5.3 Categorize ‘day’   
 5.4 Merging ‘marital’ and ‘age’  
 5.5 Inclusion and Exclusion of ‘duration’  
 5.6 Treating outliers

**6. Ready for Machine Learning**

6.1 Standardization and Normalization  
 6.2 Upsampling and Downsampling  
 6.3 Dummy Variables

**7. Machine Learning**

7.1 Logistic Regression  
 7.2 K-Nearest Neighbors  
 7.3 Support Vector Machines  
 7.4 Random Forest

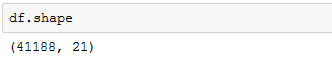
**8. Choosing the best model**

**1. Abstract**

Given some information of a marketing campaign, the goal of this project is to *predict whether or not they end up subscribing for a term deposit*.

**2. Data Inspection**

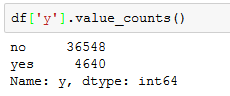
***df.shape***



**Interpretation:** There are 41188 rows and 21 features.

***df.response\_variable.value\_counts():***

The number of positive responses (yes) is largely fewer than the negative responses (no) implying that the dataset is significantly imbalanced.

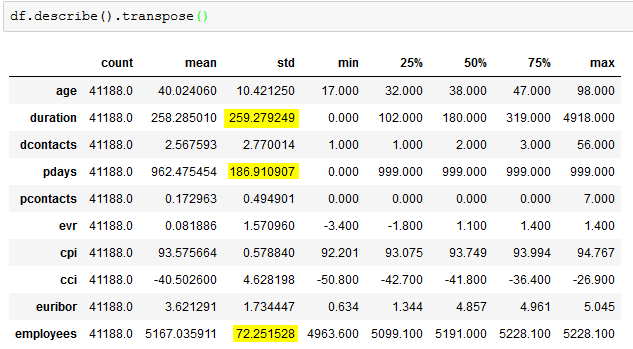


**Interpretation:** Business problems in financial and banking industries often have to deal with datasets that are massively imbalanced. Considering the reality surrounding these problems, addressing the class balance anomaly is not a major priority, for now.

**3 - EDA**

***df.describe()***

Using describe() on the dataframe, for summary statistics of all the quantitative(numeric) variables.

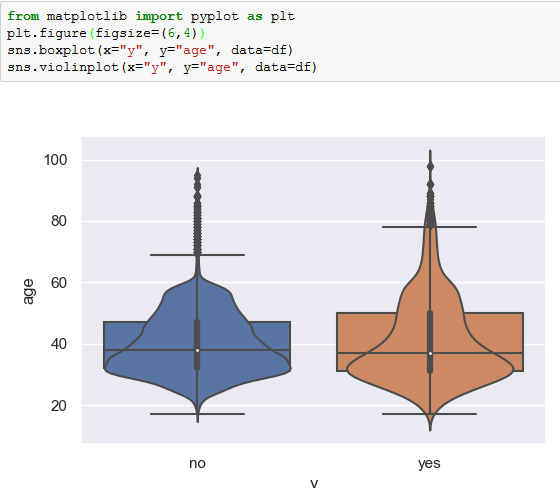
 The standard deviations of 'duration', 'pdays' and 'employees' are very large compared to other variables. These variables should be investigated to understand the reasons for this variability.

As part of Graphical EDA, I plot two graphs

* Histograms
* Violin plot

**3.1 EDA with Numeric Variables:**

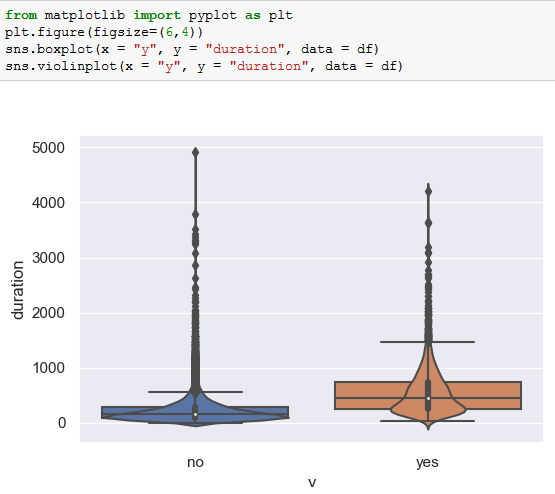
**‘age’:**

****

**Interpretation:** The variance of age of the customers who have rejected the offer is lower compared to that of the customers who have accepted. Even though most observations are around early 30's, the mean has been recorded around late 30's for both the classes.

There are significant number of outliers for both classes. However, the outliers for 'no' are widespread. Binning the 'age' variable with respect to 'job' category might provide us better insights.

**‘duration’:**

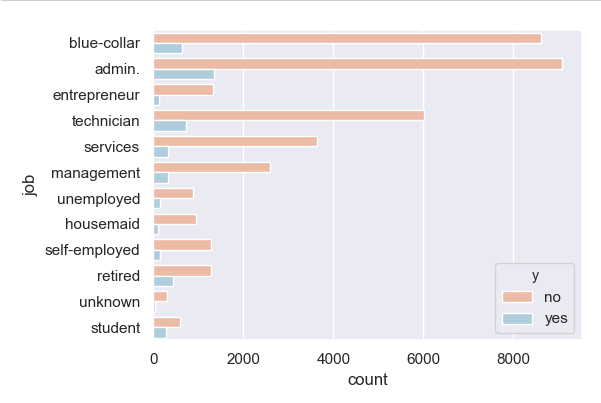
****

**Interpretation:** The variance of class 'no' of the response variable is less compared to that of 'yes' class. Outliers for 'no' are widespread than the outliers of 'yes'.

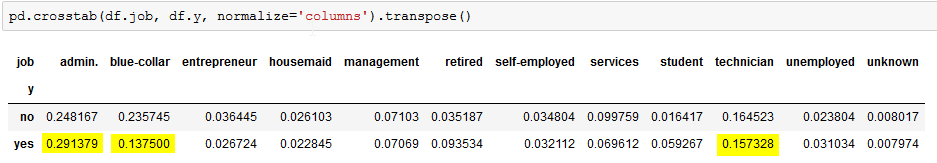
Since the data is widespread, it's a good idea to bin them and include upper bounds.

**3.2 EDA on Categorical Variables:**

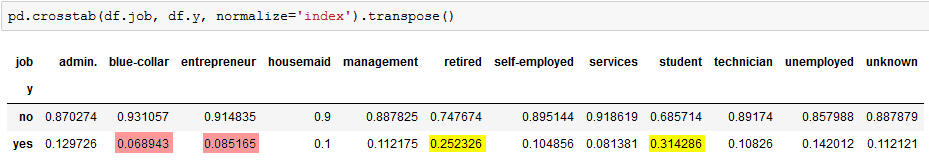
**‘job’:**



**Interpretation:** Admin category has the highest number of positive and negative responses while 'unknown' has the lowest for the both.



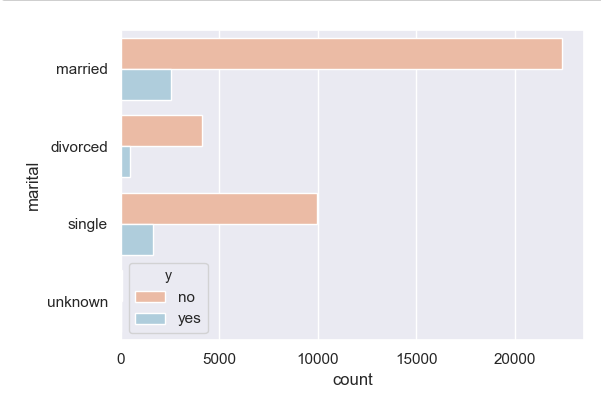
**Interpretation:** At category-level, Admin, Blue-Collar and Technicians contributed the highest percentage of positive response rate.



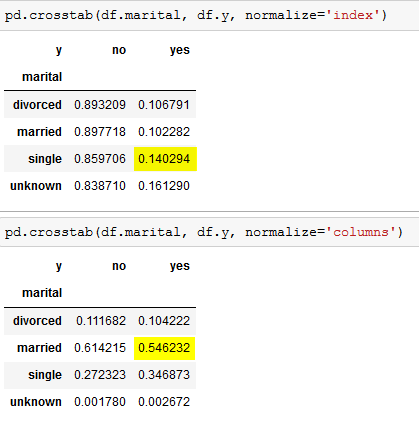
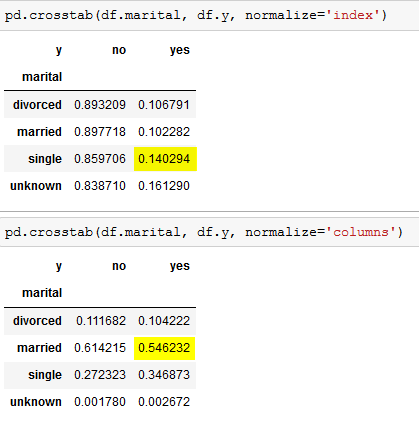
**Interpretation:** At class-level, ‘blue-collar’ and ‘entrepreneur’ (6% and 8%) had the lowest positive response rate while retired and students had high positive response rate (25% and 31%).

Thismeans that ‘admin’ and ‘blue-collar’ jobs were contacted frequently than any other job. However, the highest positive response rate, is among ‘retired’ and ‘student’, rather not ‘blue-collar’ and ‘entrepreneur’.

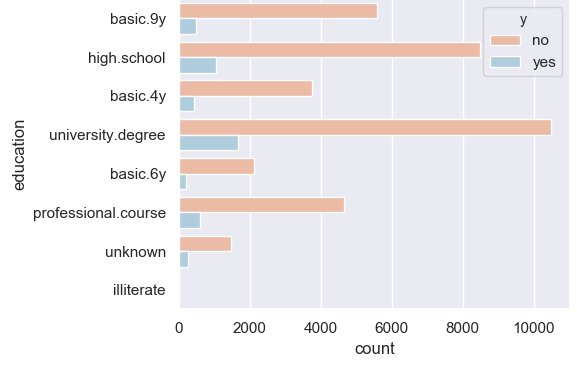
**‘marital’:**



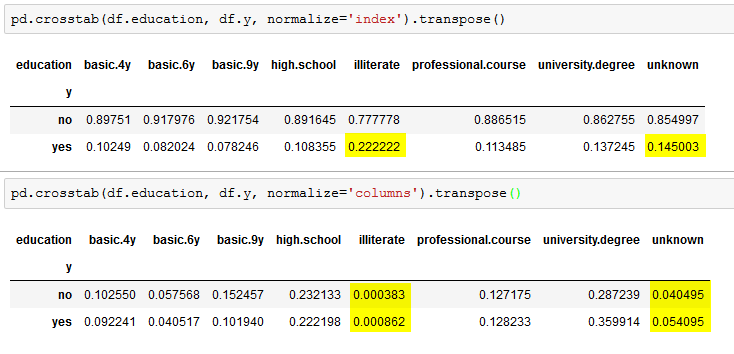
**Interpretation:** Married category has the highest number of positive responses. Around 60% of the people considered for this survey belong to either 'married' category.

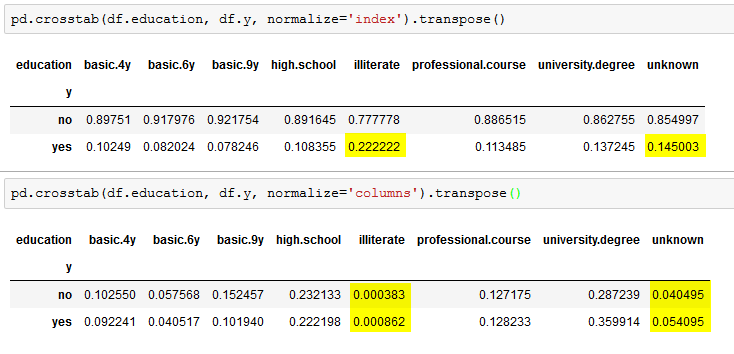
 **Interpretation:** At class-level, 'unknown' has the highest positive response rate. At category-level, with 55% 'married' contributed the highest percentage of positive response rate.

**‘education’:**



**Interpretation:** Around 5% of the datapoints has an education level of either 'illiterate' or 'unknown'. Though 'illiterate' and 'unknown' contribute to only 5% of the total datapoints, they have the highest positive response rate within categories.



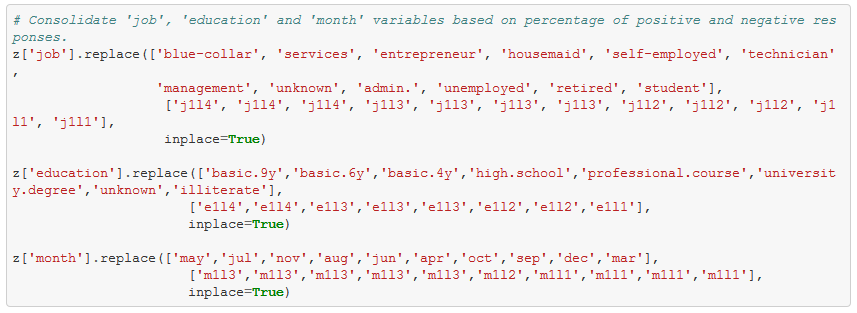


**Interpretation:** The distinction between the education descriptions is very minimal which makes it hard to combine similar classes in the category.

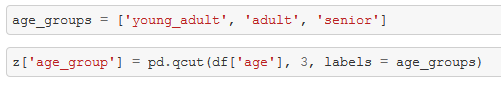
**4. Feature Engineering**

**4.1 Consolidate category classes:**

I consolidate category classes into various levels based on the percentages of ‘yes’ class.



**4.2 Binning the age:**

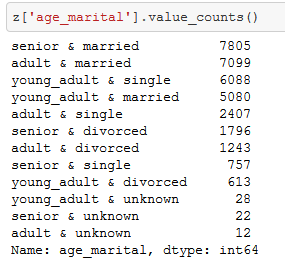


Given the data is highly imbalanced, 'age' is categorised into bins based using 'qcut' rather than 'cut'

**4.3. Categorize 'day' with 'weekday\_1', 'weekday\_2' and 'weekend' classes:**



**4.4 Merging 'marital' and 'age' variable:** 



**4.5 Inclusion and Exclusion of 'duration' column:**

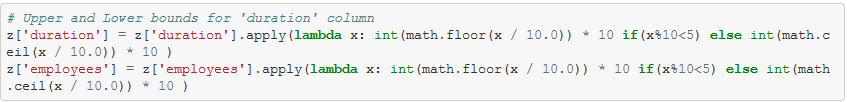
‘duration’ highly affects the output, I create two dataframes (one with 'duration' column, one without).

**4.6 Treating Outliers:**

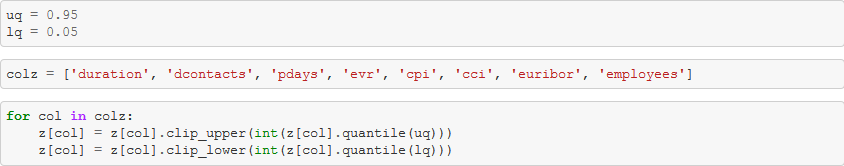
**Idea:**

(a). Replace valid outliers with logarithmic transformation  
 (b). Replace invalid outliers (human-error) with 90th percentile or upper bounds.

**4.6.1. Applying Upper and Lower bounds to 'duration' and 'employees' variable**

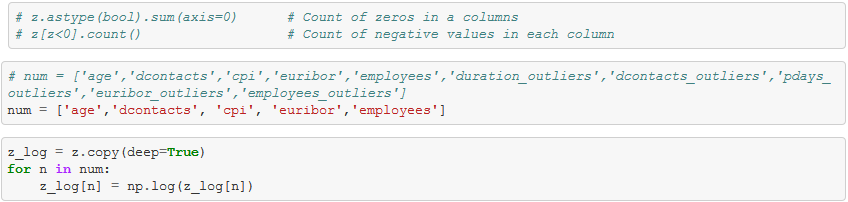


**4.6.2. Applying 90 percentiles and 5 percentiles for the lower and upper outliers**



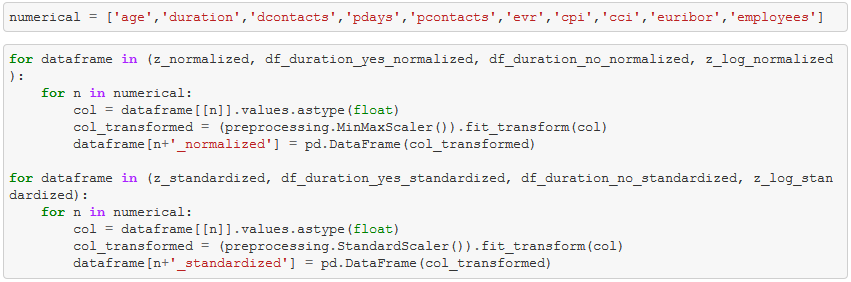
**4.6.3. Apply Logarithmic transformations to invalid outliers**

Creating a new dataframe to apply logarithm transformations. From all the numerical columns, logarithmic transformations is applied to only a few

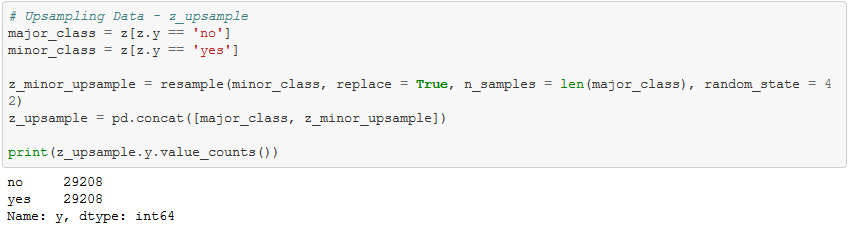


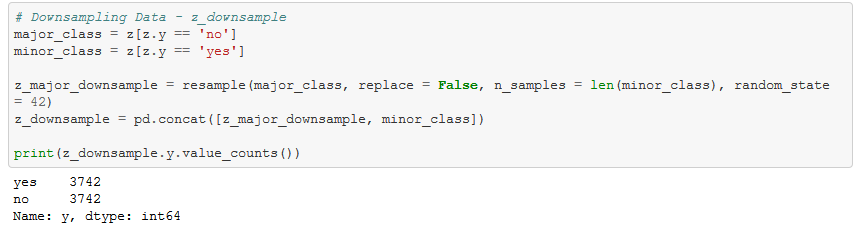
**5. Ready for Machine Learning**

**5.1 Standardization and Normalization**Two popular data scaling methods are normalization and standardization.  
 1. Data Normalization  
 2. Data Standardization

****

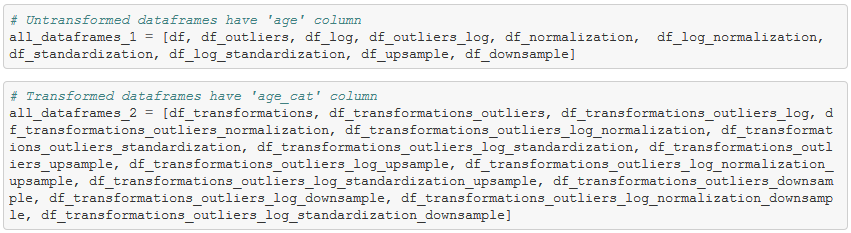
**5.2 Upsampling and Downsampling:**

****

****

**5.3 Dummy Variables:**

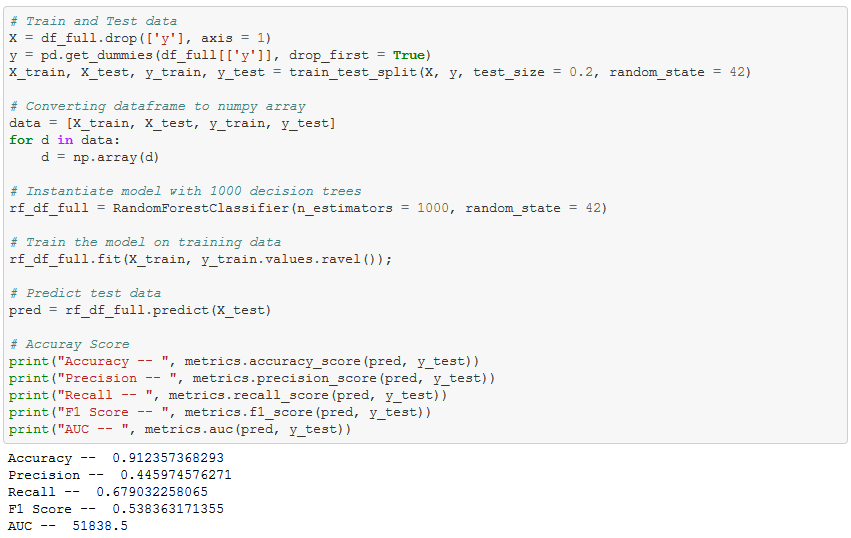
Since each dataframe has different categorical columns, all dataframes are divided into two lists.



**6. Machine Learning**

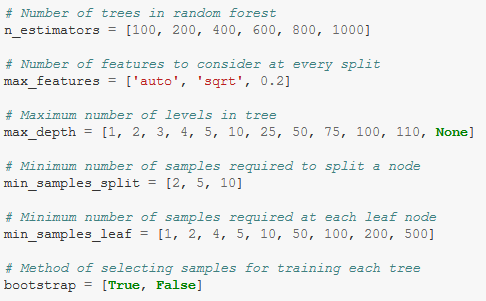
**6.1 Random Forests with untransformed data**

Initially the data is trained on the base model with no transformations

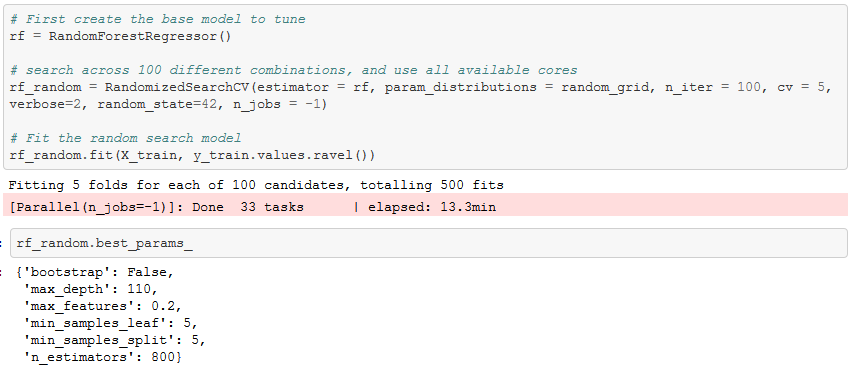
****

Accuracy – 0.912, Precision – 0.44, Recall – 0.67, F1 – 0.53, AUC - 518

**6.2 Randomized Search CV:**Efficient approach is to narrow our search to evaluate a wide range of values for each hyperparameter.

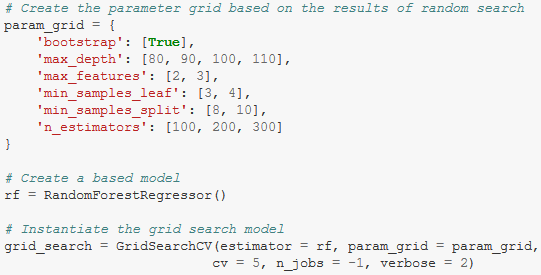


Training the base model with different sets of parameters to find the best set.

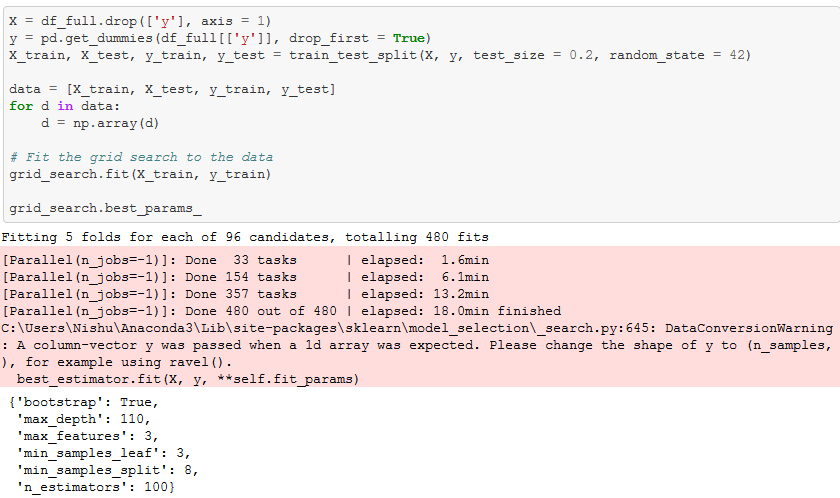


Bootstrap – False, max\_depth – 110, max\_features – 0.2, min\_samples\_leaf – 5, min\_samples\_split – 5, n\_estimators - 800

**6.3 Grid Search CV:**Using Grid Search CV to pick the best parameters. This gives us an idea where to concentrate our search.



Training the base model with different sets of parameters to find the best set.

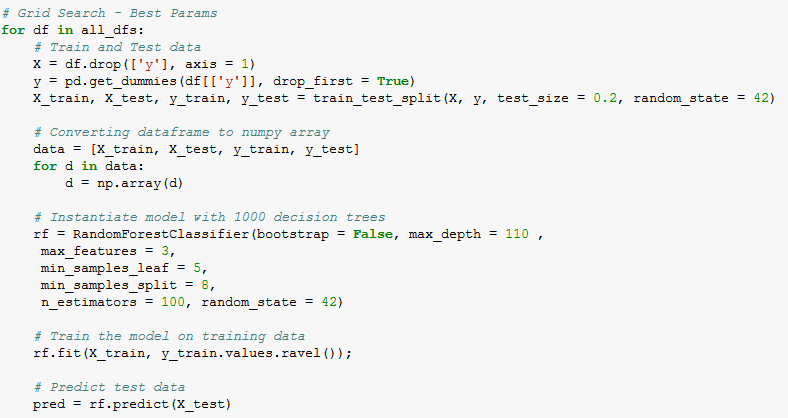


Bootstrap – False, max\_depth – 110, max\_features – 3, min\_samples\_leaf – 3, min\_samples\_split – 8, n\_estimators - 100

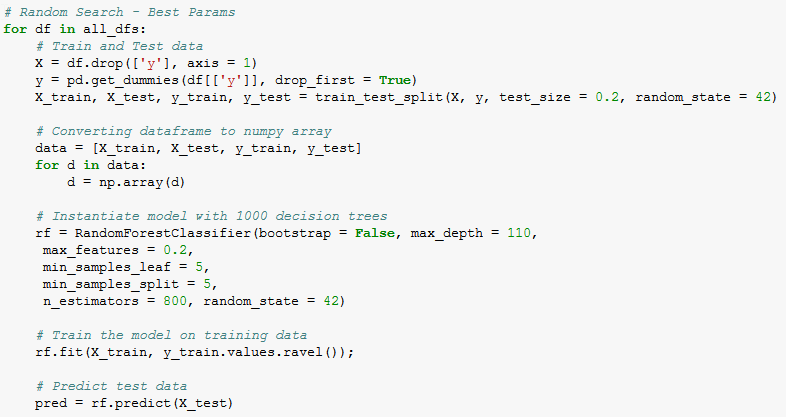
**Note:** Since the goal of this project is to minimize False Negatives (How many did we miss), we focus on getting a recall value close to 100% with a less bad precision value

**Training all the models with grid search CV best parameters:**

Training all models with best parameters of Grid Search CV

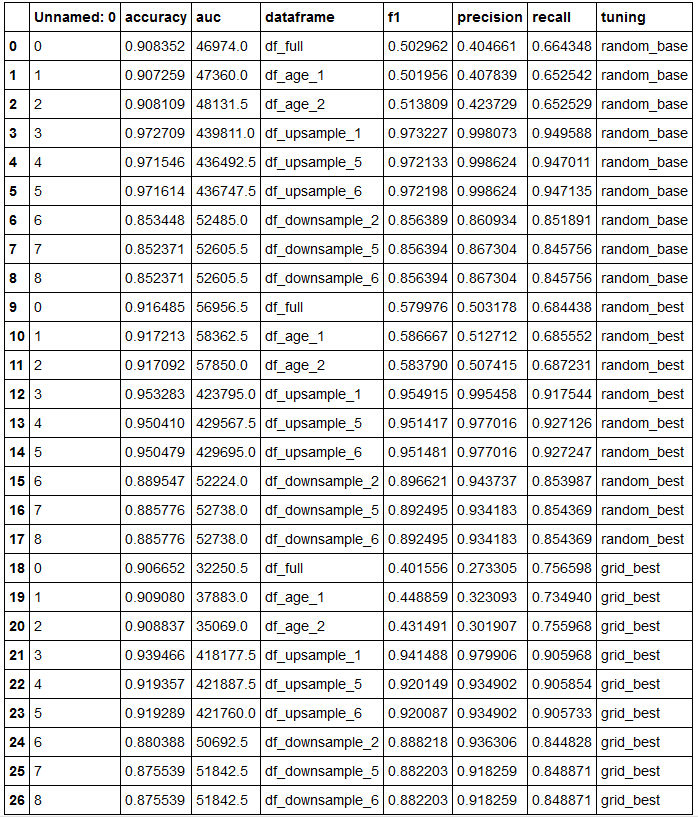


**Training all models with best parameters of Random Search CV**

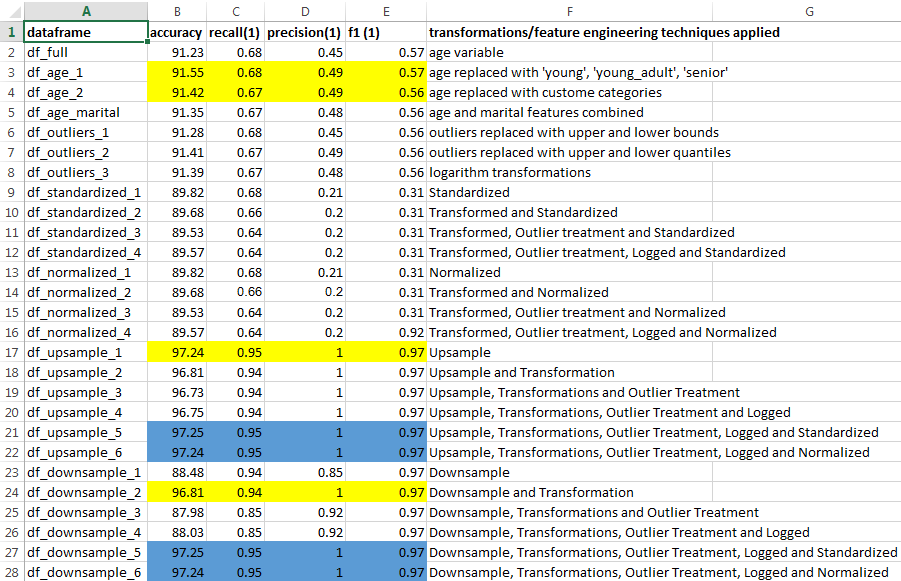


Best parameters of Grid Search CV are chosen over the best parameters Random Search CV considering the computational resources I have.

Below is the table with models are their respective metrics



Here is the list of dataframes tested on the best parameters of Grid Search CV and their respective metrics.



Highlighted are the models with optimal metrics.

**8. Choosing the best model**

From the models highlighted in the above screenshot, df\_age\_2 is the model that yields a better results on the test data.

Other models are not chosen (upsample and downsample) considering the weights each classes are given when the data is either upsampled/downsampled.

**9. Other potential data sets I could use**

The data provided could actually be considered very rich in terms of predicting the client’s behavior for a given campaign. However, given additional data pertaining to client’s financial spending such as income disposal, large credit purchases, demographic of the client.