

PROJECT REPORT



# Customer Retention by Bank Telemarketing Analytics



**“submitted towards partial fulfilment of the criteria for award of PGPDSE by GLIM”**

**Submitted by**

|  |  |
| --- | --- |
| **Student Name** | **SIS ID** |
| Kamalesh reddy | 2IIFDPEA7Y |
| Soumik Chatterjee | R5YINJG7V2 |
| Ramadurgam Rajeev Kasyap | 7RJI9JW9HD |
| Subhashree Panda | 9ORPRVXB2S |
| Sadashivuni Sushant | T4EVKUEHH7 |

**Batch: DSE\_HYDERABAD\_JUN2019**

**Mentor:** **Mr.** **Srikar Muppidi**



**Abstract:**

# Abstract & keywords

### Retaining a customer is so much cheaper than acquiring a new one. This holds true for banks and financial services providers. According to Customer Think, the costs of acquiring a banking customer are estimated to be around $200 while the typical customer generates only $150 in revenue each year. That means the relationship does not become profitable for the bank until well into the second year. In banking, the annual churn rates on new customers hover in the 20 to 25 percent range during the first year, with half not making it past the first 90 days after opening their accounts. Banks that let their customer experience decline risk losing up to 12.5 percent of their share of deposits; at the same time banks with high customer experience metrics increased their share of deposits by 16.5 percent. As these numbers suggest, providing an effective, easy, meaningful, and effortless experience is one of the keys to reducing customer churn. Therefore, to improve customer retention in banking, organizations must focus on becoming more customer-centric — and delivering experiences that wow, delight, and inspire loyalty. Hence, the current research focuses on managing customer churn in the banking sector. The extracted features from page view data kept track during the visit along with some session and user information are fed to machine learning classification methods to build a model. Oversampling and feature selection pre-processing steps are used to enhance the performance and scalability of the classification methods. The results show that Light Gradient Boosting algorithm produces significantly higher accuracy and F1- score than Logistic Regression, Decision Tree and Random forest.

**Keywords:** Customer Retention, Machine Learning, Oversampling



Acknowledgements

At the outset, we are indebted to our Mentor Mr. Srikar Muppidi for his time, valuable inputs and guidance. His experience, support and structured thought process guided us to be on the right track towards completion of this project.

We also thank all the course faculty of the DSE program for providing us a strong foundation in various concepts of analytics & machine learning.

Last but not the least, we would like to sincerely thank our respective families for giving us the necessary support, space and time to complete this project.

We certify that the work done by us for conceptualizing and completing this project is original and authentic.

Kamalesh reddy

Soumik Chatterjee

Ramadurgam Rajeev Kasyap

Subhashree Panda

Sadashivuni Sushant

Date:14th Nov 2019 Place: Hyderabad



# Certification of completion

I hereby certify that the project titled “Customer Retention by Bank Telemarketing Analytics” was undertaken and completed under my guidance and supervision by Kamalesh reddy, Soumik Chatterjee, Ramadurgam Rajeev Kasyap, Subhashree Panda, Sadashivuni Sushant, students of the June 2019 batch of the Post Graduate Program in Data Science & Engineering, Hyderabad.

Date:14th Nov 2019 Place: Hyderabad



# Table of Contents

[Chapter 1 - Project overview](#_bookmark0) [08](#_bookmark0)

[Need for study](#_bookmark0) [08](#_bookmark0)

[Problem statement & project scope](#_bookmark1) [08](#_bookmark1)

[Data sources](#_bookmark1) [09](#_bookmark1)

Dataset Description [10](#_bookmark1)

[Statistical tools & techniques](#_bookmark2) [11](#_bookmark2)

[Model performance measures used for evaluating models 12](#_TOC_250001)

[Chapter 2 - Exploratory data analysis](#_bookmark3) [14](#_bookmark3)

Data Analysis 14

[Understand data distribution](#_bookmark3) [16](#_bookmark3)

[Chapter 3 -](#_bookmark4) Model Building [28](#_bookmark4)

Data Preprocessing [28](#_bookmark4)

Feature Selection [28](#_bookmark4)

Feature Engineering [29](#_bookmark4)

Base Model [29](#_bookmark4)

Resampling techniques [34](#_bookmark4)

Chapter 4 –Conclusion 34

Abbreviations used

|  |  |
| --- | --- |
| **Abbreviation** | **Expansion** |
| LR | Logistic Regression |
| DT | Decision Tree |
| AUC | Area Under the Curve |
| RF | Random Forest |
| LGBM | Light Gradient boosting method |
| Bag\_DT | Bagging Decision Tree |
| Boost\_DT | Boosting Decision Tree |
| FNR | False Negative Rate |
| FPR | False Positive Rate |
| mRMR | Minimum Redundancy and maximum relevance |
| SMOTE | Synthetic Minority Oversampling Technique |
| URL | Uniform Resource locator |



Executive summary

**Background & need for study**: Retaining a customer is so much cheaper than acquiring a new one. This holds true for banks and financial services providers. According to Customer Think, the costs of acquiring a banking customer are estimated to be around $200 while the typical customer generates only $150 in revenue each year. That means the relationship does not become profitable for the bank until well into the second year. In banking, the annual churn rates on new customers hover in the 20 to 25 percent range during the first year, with half not making it past the first 90 days after opening their accounts. Banks that let their customer experience decline risk losing up to 12.5 percent of their share of deposits; at the same time banks with high customer experience metrics increased their share of deposits by 16.5 percent. As these numbers suggest, providing an effective, easy, meaningful, and effortless experience is one of the keys to reducing customer churn.

**Scope & Objectives**: To improve customer retention in banking, organizations must focus on becoming more customer-centric — and delivering experiences that wow, delight, and inspire loyalty. Hence, the current research focuses on managing customer churn in the banking sector.

**Approach & methodology:** After processing the dataset and cleaning the inconsistencies, the numerical and categorical features used in the purchasing intention prediction model is generated. Various Classification algorithms are used to predict online consumer commercial intent based on set of independent variables like traffic type, visitor type, duration on administration pages, informational pages and product pages along with technology used. The predictive models are also used to identify the variables that strongly influence the conversion using variable importance and probabilistic approaches. The models are evaluated using relevant model performance measures to arrive at the most robust models for prediction.



# Chapter 1 - Project overview

The banking industry is data-intensive with typically massive graveyards of unused and unappreciated ATM and credit processing data. As banks face increasing pressure to stay profitable, understanding customer needs and preferences becomes a critical success factor. New models of proactive risk management are being increasingly adopted by major banks and financial institutions, especially in the wake of Basel II accord. Through Data mining and advanced analytics techniques, banks are better equipped to manage market uncertainty, minimize fraud, and control exposure risk. But in order to discover the set of critical success factors that will help banks reach their strategic goals, they need to move beyond standard business reporting and sales forecasting. By applying data mining and predictive analytics to extract actionable intelligent insights and quantifiable predictions, banks can gain insights that encompass all types of customer behavior, including channel transactions, account opening and closing, default, fraud and customer departure.

The best bank customer retention strategy for existing customers is to classify each type of customer (silent attrition, ideal and unhappy) and create appropriate initiatives to change their behavior. For instance, customers in “silent attrition” are those that have reduced or stopped using a product, but where the account is still open. Examples for instance are credit card accounts with little or no spending. For these customers, you must determine why they are no longer using your product (are you are their “back of wallet” card) and create initiatives to change their behavior.

Customers that are Exiting are those customers that have started the process of moving their business to another company or are in the process of considering that move. The first step in creating bank customer retention strategies for Exiting customers is to identify which customers are in each camp. For customers in the process of moving their business you will need to understand the product drop cycle, i.e. the order in which customers drop your products before leaving. With this information you can create effective customer retention strategies to target those customers.

Hence by working on the provided dataset we will be reducing the Customer Churn and provide suggestions to the telephone directory about the customers who willing opt for further services from the bank.

## 

.

## Data sources

The data is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed.

The dataset gives you information about a marketing campaign of a financial institution in which you will have to analyze in order to find ways to look for future strategies in order to improve future marketing campaigns for the bank. We need to find the best strategies to improve for the next marketing campaign. How can the financial institution have a greater effectiveness for future marketing campaigns? In order to answer this, we have to analyze the last marketing campaign the bank performed and identify the patterns that will help us find conclusions in order to develop future strategies. Marketing campaigns are characterized by focusing on the customer needs and their overall satisfaction. Nevertheless, there are different variables that determine whether a marketing campaign will be successful or not. There are certain variables that we need to take into consideration when making a marketing campaign.

Let’s understand the 4 Ps of a successful marketing strategy: -

1. Segment of the Population: To which segment of the population is the marketing campaign going to address and why? This aspect of the marketing campaign is extremely important since it will tell to which part of the population should most likely receive the message of the marketing campaign.
2. Distribution channel to reach the customer's place: Implementing the most effective strategy in order to get the most out of this marketing campaign. What segment of the population should we address? Which instrument should we use to get our message out? (Ex: Telephones, Radio, TV, Social Media Etc.)
3. Price: What is the best price to offer to potential clients? (In the case of the bank's marketing campaign this is not necessary since the main interest for the bank is for potential clients to open deposit accounts in order to make the operative activities of the bank to keep on running.)
4. Promotional Strategy: This is the way the strategy is going to be implemented and how are potential clients going to be address. This should be the last part of the marketing campaign analysis since there has to be an in-depth analysis of previous campaigns (If possible) in order to learn from previous mistakes and to determine how to make the marketing campaign much more effective.

## Dataset Description

This dataset has 17 columns (features) and 45212 rows (records).

|  |  |
| --- | --- |
| **Feature Name** | **Feature Description** |
| age | Numeric |
| job | type of job (categorical: "admin.","unknown","unemployed","management","housemaid","entrepreneur","student","blue-collar","self-employed","retired","technician","services") |
| marital | marital status (categorical: "married", "divorced", "single"; note: "divorced" means divorced or widowed) |
| education | (categorical: "unknown", "secondary", "primary", "tertiary") |
| default | has credit in default? (binary: "yes", "no") |
| housing | has housing loan? (binary: "yes", "no") |
| loan | has personal loan? (binary: "yes", "no") |
| contact | contact communication type (categorical: "unknown", "telephone", "cellular") |
| day | last contact day of the month (numeric) |
| month | last contact month of year (categorical: "Jan", "feb", "mar", ..., "nov", "dec") |
| duration | last contact duration, in seconds (numeric) |
| campaign | number of contacts performed during this campaign and for this client (numeric, includes last contact) |
| pdays | number of days that passed by after the client was last contacted from a previous campaign (numeric, -1 means client was not previously contacted) |
| previous | number of contacts performed before this campaign and for this client (numeric) |
| poutcome | outcome of the previous marketing campaign (categorical: "unknown", "other", "failure", "success") |
| y | has the client subscribed a term deposit? (binary: "yes", "no") |

## Statistical tools & techniques

Various classification algorithms have been used to analyze customer purchase intention for conversion and to identify the extent to which each independent variable influence conversion. The independent variables can be broadly grouped as Visitor session information and visitor pageview information. The dependent variable is whether the customer will generate the revenue or not by his session navigation pattern.

The model building exercise has also considered cross validation and tuning techniques to ensure that the models built perform well when used for prediction.

The classification algorithms used for Commercial intent prediction include

* Logistic regression
* Decision Tree
* Random Forest
* Light Gradient Boosting

## Model performance measures used for evaluating models

The various models built, must be evaluated based on certain model performance measures to identify the most robust models. The choice of the right model performance measures is highly critical since the dataset is a highly imbalanced dataset. Model accuracy alone may not be enough to evaluate a model. Hence the following model performance measures have been used to evaluate the models, based on the confusion matrix built for the predictions on the training and test datasets:

|  |  |  |
| --- | --- | --- |
|  | **Negative (Predicted)** | **Positive (Predicted)** |
| **Negative (Observed)** | True Negative (TN) | False positive (FP) |
| **Positive (Observed)** | False negative (FN) | True positive (TP) |

### Accuracy

Accuracy is the number of correct predictions made by the model by the total number of records. The best accuracy is 100% indicating that all the predictions are correct.

Considering the response rate (conversion rate) of our dataset which is ~16%, accuracy is not a valid measure of model performance. Even if all the records are predicted as 0, the model will still have an accuracy of 84%. Hence other model performance measures need to be evaluated.

### Sensitivity or recall

Sensitivity (Recall or True positive rate) is calculated as the number of correct positive predictions divided by the total number of positives. It is also called recall or true positive rate (TPR).

For our dataset, it gives the ratio of actual customers who generated revenue by the total number of customers predicted who will generate the revenue.

### Specificity

Specificity (true negative rate) is calculated as the number of correct negative predictions divided by the total number of negatives.

For our dataset, specificity gives the ratio of actual customers who will not generate revenue by the number of customers who are predicted who will not generate revenue.

### Precision

Precision (Positive predictive value) is calculated as the number of correct positive predictions divided by the total number of positive predictions.

Precision tells us, what proportion of customers who generated revenue as customers actually generated revenue. If precision is low, it implies that the model has lot of false positives.

### F1-Score

F1 is an overall measure of a model’s accuracy that combines precision and recall A good F1 score means that you have low false positives and low false negatives, so you’re correctly identifying real threats and you are not disturbed by false alarms. An F1 score is considered perfect when it’s 1, while the model is a total failure when it’s 0.

### ROC chart & Area under the curve (AUC)

ROC chart is a plot of 1-specificity in the X axis and sensitivity in the Y axis. Area under the ROC curve is a measure of model performance. The AUC of a random classifier is 50% and that of a perfect classifier is 100%. For practical situations, an AUC of over 70% is desirable.

### Level of significance

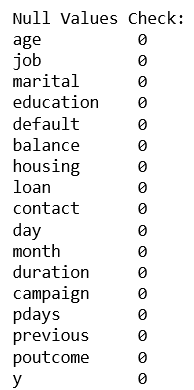
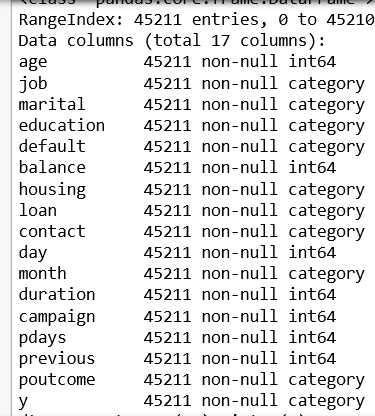
For all the hypothesis tests in the project, the level of significance is assumed as 5% unless specified otherwise.



# Chapter 2 - Exploratory data analysis

## Data Analysis

* Checking for the missing values and Datatypes:

*There are no null values in our dataset*

* Treatment of Outliers: ['age', 'balance', 'day', 'duration', 'campaign', 'pdays', 'previous']

Example summary snippet with respect to age:





*For the age below 40 there's the status retirement. Ideally retirement is for those who reaches the age 60.*

*Above 40 if the job status is retired, maybe it's because of the choice. The average retirement*

*age is 66 in Portugal. Let's compare it to Housing loan and personal loan. Younger people tend*

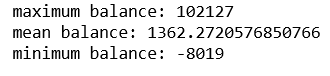
*to take personal loan. The middle aged financially established people will take housing loan.*

*Here the customers of age < 29 have the personal loans as yes and people around 30+ have the housing loan.*

*Also, People of age below 30 if the job status is retired maybe it'd because of health issues or*

*other constraints which can't be explained.*

* Checking the balance feature:

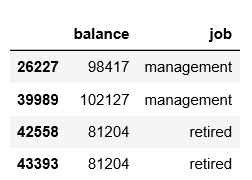


*So, the negative balance is the amount that customer owns to the bank. The average balance*

*in the bank for the customers is around 1400 euros. The maximum balance is 102,127 euros.*

*Let's check it with respect to job status.*

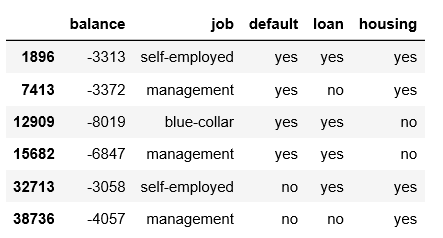




* Checking balance with respect to job It seems reasonable for a management and a retired customer to

have these numbers. May be just extreme values.





*Again, just a case of extreme values those with the negative balance lesser than -3000 are either*

*self-employed, management or blue collar and all of them either have one or the other form*

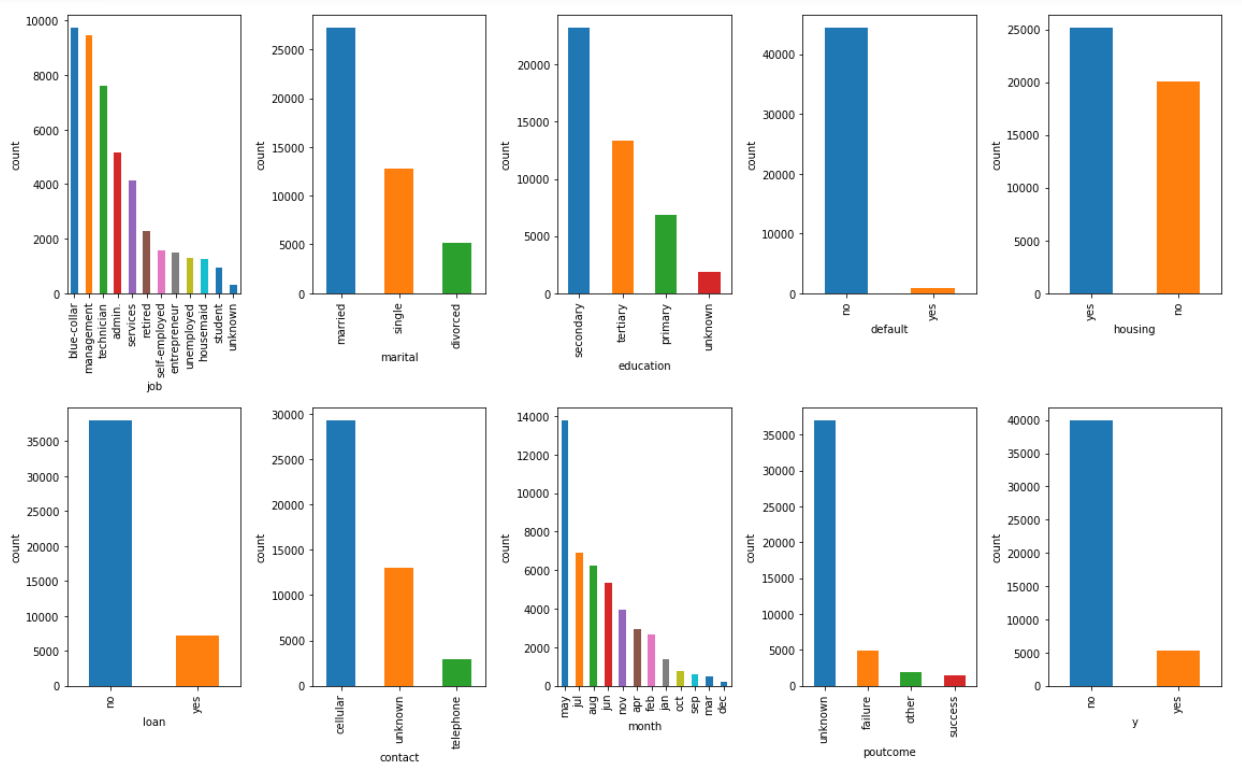
*of loan or being involved in credit default. A case of being not able to repay to the bank.*

*Not going to remove any values from the dataset.*

## Understand data distribution

* Univariate Analysis:

Basic count plot on the categorical variables to look at the proportion



*Job: We mostly have blue-collar, management and technicians in the feature.*

*Marital: We mostly have married in the feature.*

*Education: We mostly have secondary level in the feature.*

*Default: We mostly have no credit defaulters in the feature.*

*Housing: We mostly have house loaners in the feature.*

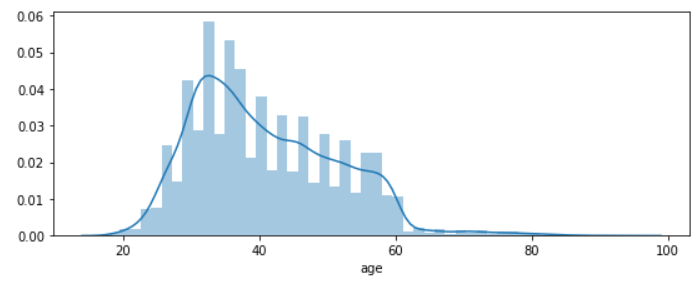
*Loan: We mostly have customers with no personal loans in the feature.*

*Contact: The campaign is and was mostly carried out through Cellular.*

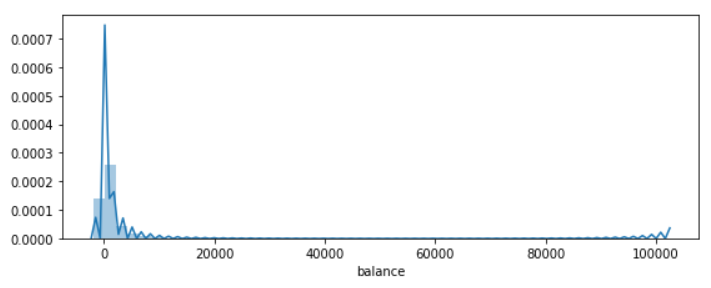
*Month: Maximum may month in the feature.*

*poutcome: The previous campaign outcome is mostly unknown.*

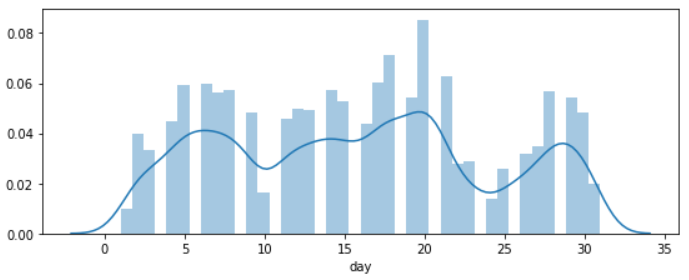
* Density plot to look at the numerical data:



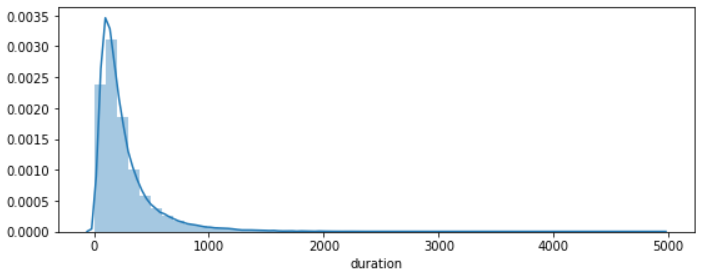
*Age: I see the age is mostly distributed around late twenties to late thirties*



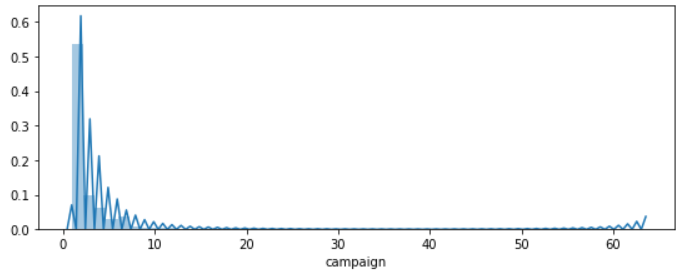
*Balance: Average yearly balance stacked around single digit thousands*



*Day: Last time campaigners contacted the customers for that month evenly distributed*



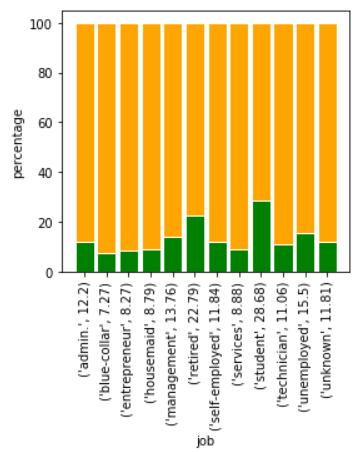
*Duration: last contact duration in seconds mostly around the hundred's*



*Campaign: Number of calls made during the campaign mostly below 10 calls but few went beyond 60 times. Please find below the graph.*

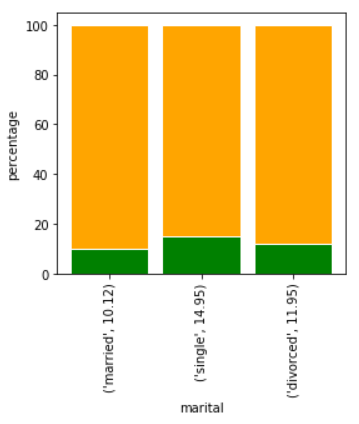
* Bivariate Analysis:

Percentage (%) Bar Graph for the categorical variables (the % of target variables in each feature):

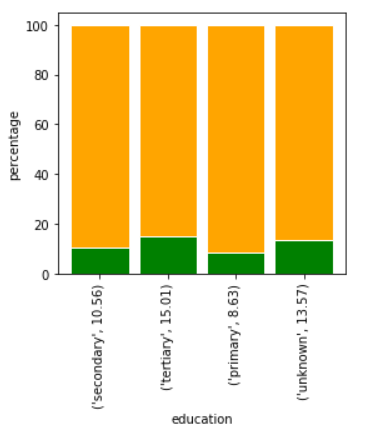


*Job: The maximum % of clients who subscribed for Term Deposit are students (28.6%) and*

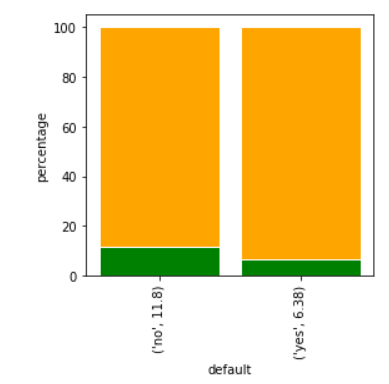
*retired people (17.4%)*



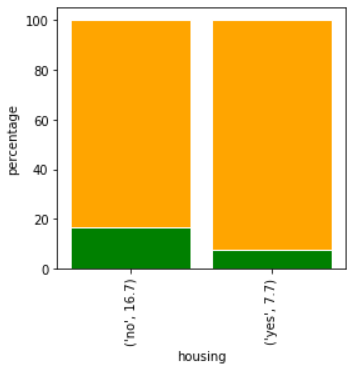
*Marital: The maximum % of clients who subscribed for Term Deposit are single people (14.9%)*



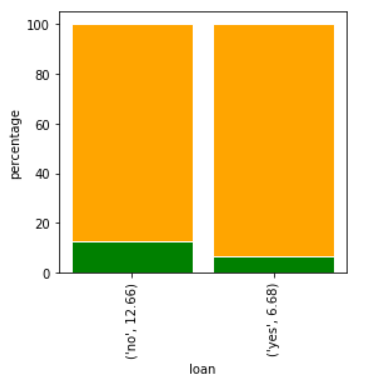
*Education: The maximum % of clients who subscribed for Term Deposit have a tertiary level (15.0%)*



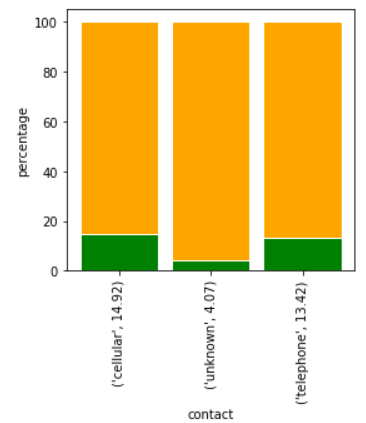
*Default: The maximum % of clients who subscribed for Term Deposit do not have a credit in default (11.8%)*



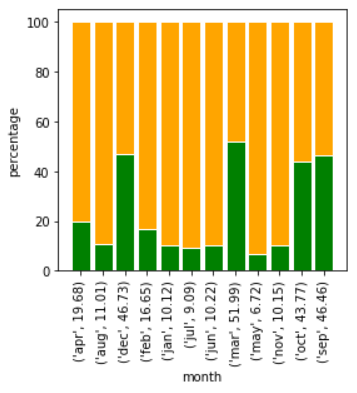
*Housing: The maximum % of clients who subscribed for Term Deposit have housing loan (16.7%)*



*Loan: The maximum % of clients who subscribed for Term Deposit do not have personal loan (12.6%)*

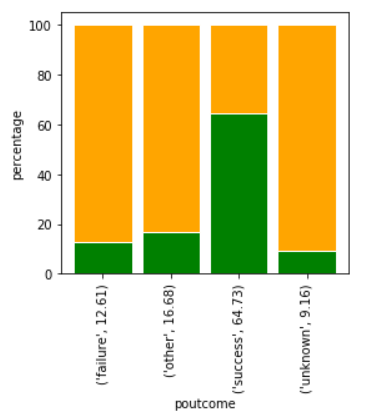


*Contact: The maximum % of clients who subscribed for Term Deposit were contacted through the cellular mode of communication, as a part of campaign by the bank (14.9%)*



*Month: The maximum % of clients who subscribed for Term Deposit were contacted last on the month of*

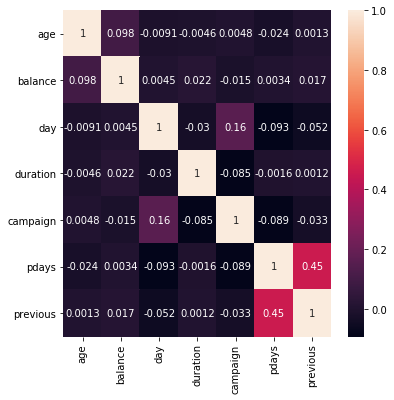
*March of that year (51.9%)*



*poutcome: The maximum % of clients subscribed for the Term Deposit also had subscribed during the previous campaign. Which means the campaign is working successfully in majority for the previous campaign where the outcome that's successful. The customer retention seems to be working*

*Here we are just comparing the 'Yes' proportion horizontally for the specific features. Any how the 'No' proportion leads compared to 'Yes' individually (vertical).*

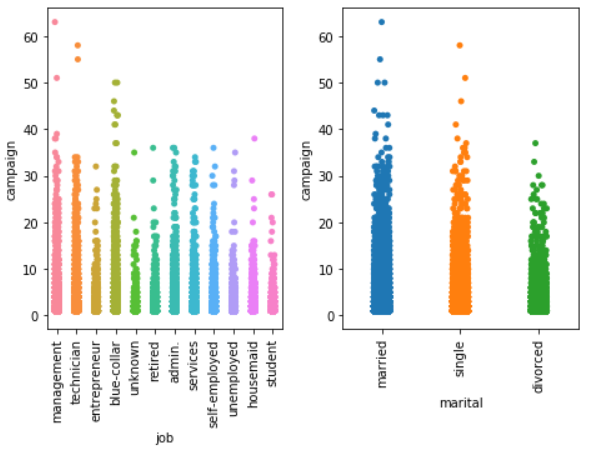
* Co-relation matrix with Heat map to check the co-relation between the numerical variables:

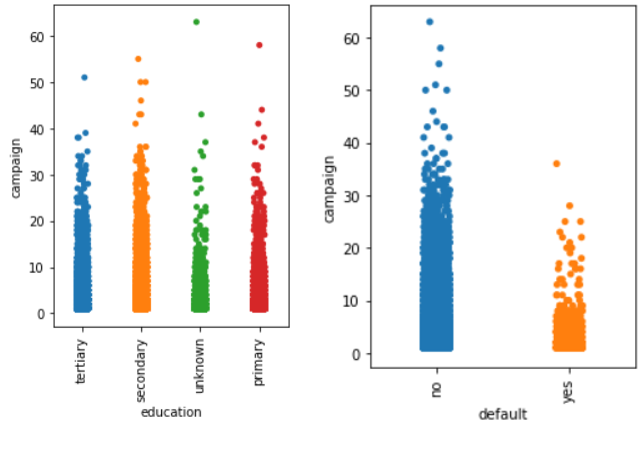


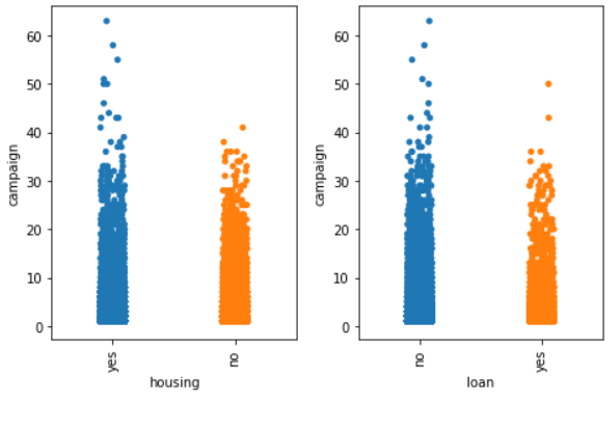
*We do not see any positive or a negative relation between any of the variables.*

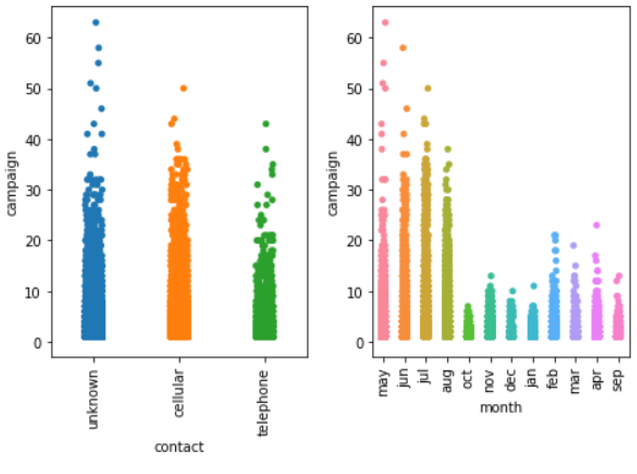
* Let’s see the number of contacts that’s being made through out the months and let’s take other

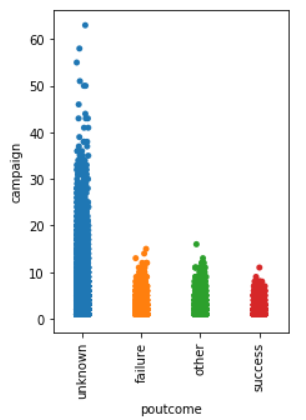
categorical features into the consideration:



**

**





*Observations:*

*Job: We see that most of the call are being made to those who are in services, management and admins*

*Marital: We see that most of the call are being made to married and singles*

*Education: We see the calls are spread evenly almost, but We see mostly being made to the one in the tertiary level* *and Notice one call being made more than 60 times in the primary category.*

*Default: Campaign calls are mostly carried out to those who do not have their credit in default.*

*Housing: Campaign calls are mostly carried out to those who do have housing loans.*

*Loan: Campaign calls are mostly carried out to those who do not have any personal loans.*

*Contact: We see that most of the call are being made through cellular and unknown but the frequency is more though the Cellular we see it crossed 60 calls.*

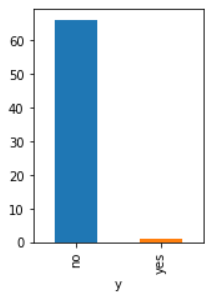
*Month: So, Here the campaign is mostly carried out in the months of May, July, August and June. Moderately*

*in the months of September, October and March. Less in the months of January, February, April, November*

*and December.*

*poutcome: The most calls are being made to depending up on the previous campaigning outcomes that is unknown.*

* Let’s check the calls made more than 30 times during campaign to see if it went successful:



*We are surprised even though the calls are made frequently, the campaigners could not make*

*customers subscribe for the term deposits.*

# Chapter 3 - Model Building

## Data Preprocessing:

## We started with checking for null values in the dataset. There were no null values. Then we check

## For dataset info, Grabbing the categorical columns and converting the object datatype to

## Category.

## Outlier removal is not being carried out since the bank cannot afford losing its valuable customer data.

## For example, we can see a lot of outliers in Age column. For the age below 40 there's the status

## retirement. Ideally retirement is for those who reaches the age 60. Above 40 if the job status is

## retired, maybe it's because of the choice. The average retirement age is 66 in Portugal. Let's compare

## it to Housing loan and personal loan. Younger people tend to take personal loan. The middle aged

## financially established people will take housing loan. Here the people of age < 29 have the personal

## loans as yes and people around 30+ have the housing loan. Also, People of age below 30 if the job

## status is retired maybe it'd because of health issues or other constraints which can't be explained.

## We performed scaling on the numeric features of the dataset.

## Feature Selection:

## We performed chi-square test for every categorical variable with the target 'y'

## categorical variable to check for the independence.

## Ho: There is no statistically significant relationship between Dependent and the Independent variable.

## Ha: There is a statistically significant relationship between Dependent and the Independent variable.

## All the features were showing equal importance.

## Feature Engineering:

## Here we are a creating another feature named Age Group and Weighted months, months weights with respect to the calls that are being made respectively.

## Also, One-hot encoding: Nominal non-binary - Enabling drop\_first=True parameter to avoid dummy variable trap that leads to multicollinearity and Label Encoding: ordinal non-binary.

## Converted the seconds duration to minutes: data['duration’] =round(data['duration']/60,2)

## Base Model:

## We have taken our base model as logistic regression, because logistic regression is a method

## which is good at predicting binary class output (logistic regression works for multiclass classifiers

## as well), i.e. the target variable is binary in nature. It is a widely used technique because it is very

## efficient and does not require too many computational resources. It does not require input features

## to be scaled and there is no need for tuning. Because of the simplicity it can be implemented easily

## and quickly so logistic regression is a good baseline model.

## Logistic regression shows the probability of an event occurrence. For example, in our data for

## one record or for one person if predicted probability is 0.4 which means there is 40% change of subscribing for a term deposit or in other words, we can say that they are 60% i.e. (1-0.4) chance

## of not subscribing for a term deposit. Logistic regression works same as that as linear regression.

## W.K.T linear regression equation

## Y = b0+b1X1+b2X2+……. bnXn

## Logistic regression also uses the same equation but it uses sigmoid function to “y” so that all the

## values come under 0 to 1. We can set threshold in logistic regression so that the value above the

## threshold will be predicted as positive class or 1 and the value below the threshold will be predicted as negative class or 0. For example we have set threshold as 0.5 then if output of the sigmoid function is more than 0.5, we can classify the outcome as “yes” or 1.

## If it is less than 0.5, we can classify as “No” or 0.

## 

## Logistic regression works well when we remove features those are unrelated to the target variable.

## When there is a correlation between two independent variables, we can drop one. Therefore, feature engineering plays an important role in logistic regression. Now let’s get back to our data and try to fit

## the data to our base model. Our data consists of 10 categorical columns and 7 continuous columns. We cannot build a base model using categorical columns in our data. So, we need to convert categorical

## data in to numerical.

## Data is of 2 types

## 1)Nominal data

## 2)Ordinal data

## Nominal data is the data which has no order or sequence or hierarchy. That kind of categorical data is transformed to numerical by one hot encoding.

## Ordinal data is the data which follows order or hierarchy. We use label encoding for this kind of data.

## In our there are 10 categorical columns, they are Job, marital, education, default, housing, loan, contact, month, poutcome, y.

## We have done label encoding for education and month as they follow order, replaced 1 for yes and 0

## for no in default, housing, loan, y columns and done one hot encoding for remaining columns and we

## land up with 31 features

## 

## Train test split:

## We can consider all independent features as X and one dependent feature as y

## By using train test split we can split the data into 80/20 or 70/30 or whatever ratio we want i.e.

## for 70/30 70% of the random data goes in to training and remaining 30% of the data goes in to

## testing or validation set. Since we are having imbalance in the data, to capture it we can train the

## model on more training records i.e. 80/20.

## After splitting we are training the model with training data and evaluating model performance

## using validation set.

## 

## 

## By seeing 90 % training and testing accuracy we should not judge that our model is working well when

## our data is imbalanced. We need to check important metrics such as f1score, precision, recall

## (precision and recall can be referred from confusion matrix) and then come to a conclusion.

## Confusion Matrix:

## 

## From confusion matrix we can refer that

## True positive TP = 350 (correctly predicted as positive class)

## True negative TN = 7763 (correctly predicted as negative class)

## False positive FP = 217 (wrongly predicted as positive)

## False negative = 713 (wrongly predicted as negative)

## Recall or sensitivity or true positive rate is proportion of actual values that are identified correctly

## Out of all predicted positives how many are actually positive is called precision

## Specificity or true negative rate is defined as the proportion of actual negatives which got

## predicted as negative.

## Classification results:

## 

## F1 score is defined as the weighted harmonic mean of test’s precision and recall

## 

## F1 score depends on precision and recall, if both precision and recall are high then f1\_score will also be

## high, that’s why f1\_score is an important metric while dealing with classification problem. we can judge

## our model by looking at the f1\_score of our model. If f1\_score is too low our model is not at all good

## at predicting. We can get further information from classification report

## Classification report:

## 

## Classification report is the report which consists of metrics like precision, recall, f1\_score, accuracy of

## both the classes. We can infer that precision and recall is higher for class 0 and obviously f1\_score of

## class 0 is high which means our model is biased towards predicting class 0 (clients who are not

## subscribed for term deposit). Generally, when the binary class are around the proportion 70:30 the

## relation is expected to be captured. In our case target variable class is in the ratio of 88:12, in which

## model can’t capture the minority class data.

## There are a greater number of records of the people who aren’t subscribing for term deposit and a

## smaller number of records of the people who are subscribing for a term deposit. Since there is

## more imbalance in the data, our model is biased towards majority class and predicting the people who

## are subscribed as not subscribed for term deposit. That’s why precision and recall for minority class is

## low.

## To overcome this problem, we have to do re-sampling technique

There are two types of re-sampling techniques

1. Over sampling
2. Under sampling

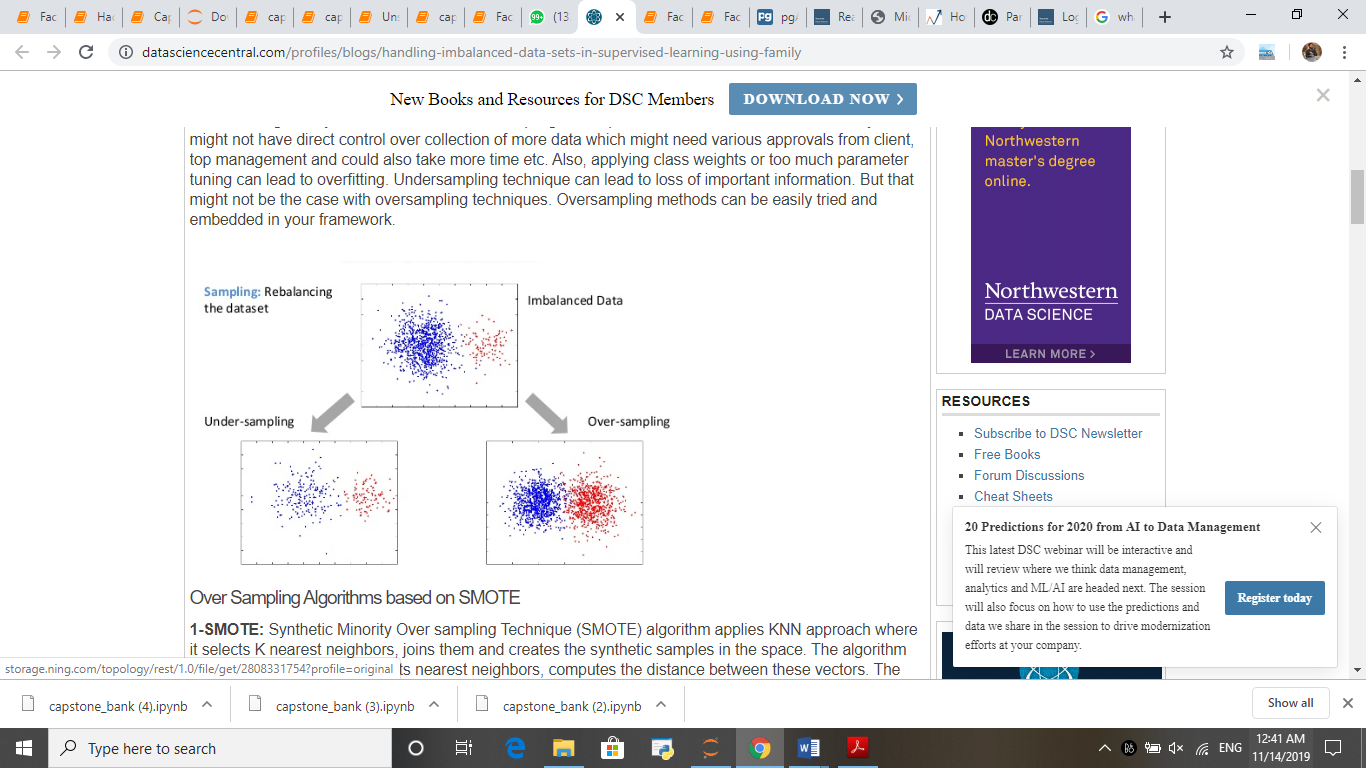
If we use under sampling the majority class (88%) will be reduced to the ratio which is equal to minority class (12%) which leads to loss of data and more variance error compared to over sampling method. So, we would like to implement over sampling which overcomes all the shortcoming that under sampling technique results in.

## Resampling Techniques:

Data Imbalance:

An imbalanced dataset means instances of one of the two classes is higher than the other, in another way, the number of observations is not the same for all the classes in a classification dataset.

We have used 3 resampled techniques, Random Under sampling, smote (synthetic minority class over sampling technique), Smotetomek



Random Under sampling:

Random under-sampling is a non-heuristic method that aims to balance class

distribution through the random elimination of majority class examples. The rationale

behind it is to try to balance out the dataset in an attempt to overcome the idiosyncrasies

of the machine learning algorithm. The major drawback of random under sampling

is that this method can discard potentially useful data that could be important for the induction process. Another problem with this approach is that the purpose of machine learning is for the classifier to estimate the probability distribution of the target variable.

Smote:

SMOTE generates synthetic minority examples to over-sample the minority class.

Its main idea is to form new minority class examples by interpolating between several

minority class examples that lie together. For every minority example, its k (which is

set to 5 in SMOTE) nearest neighbors of the same class are calculated, then some

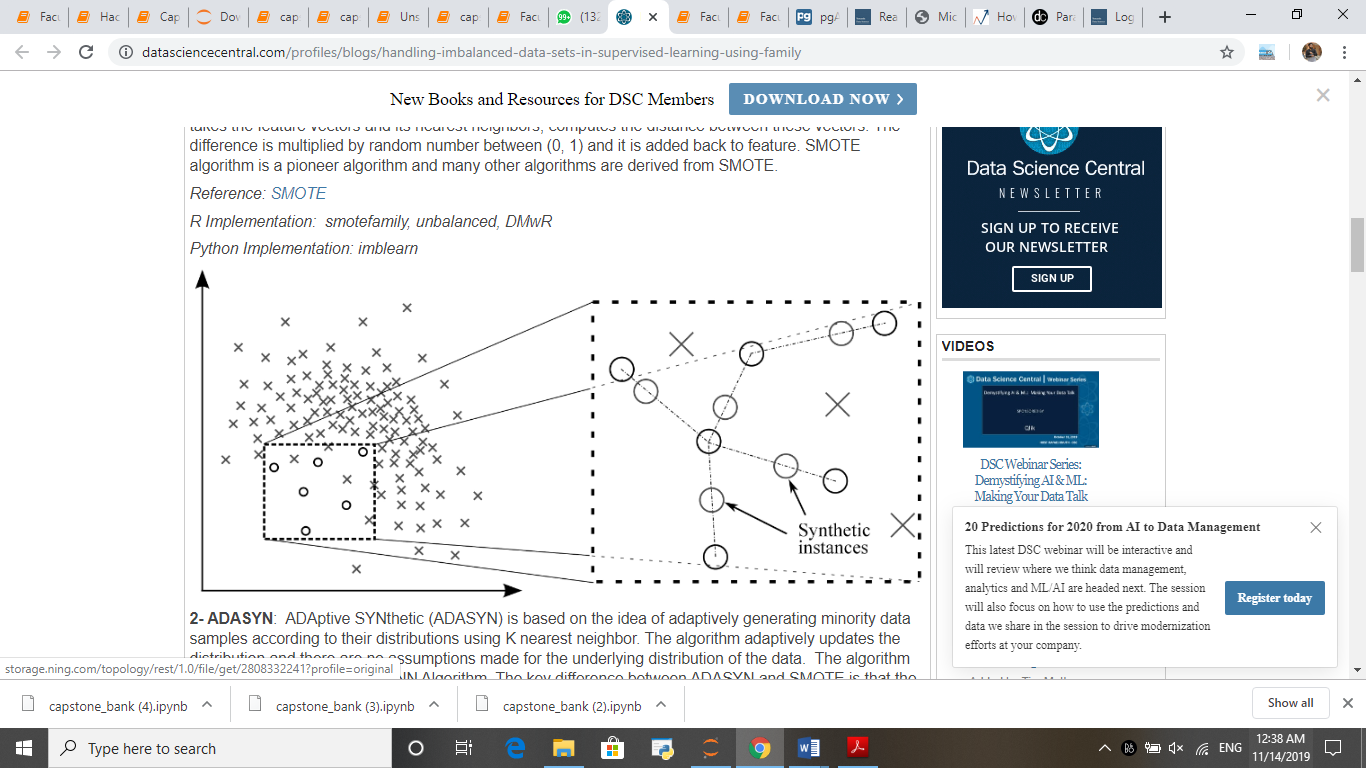
examples are randomly selected from them according to the over-sampling rate. After

that, new synthetic examples are generated along the line between the minority example

and its selected nearest neighbors. Thus, the overfitting problem is avoided and

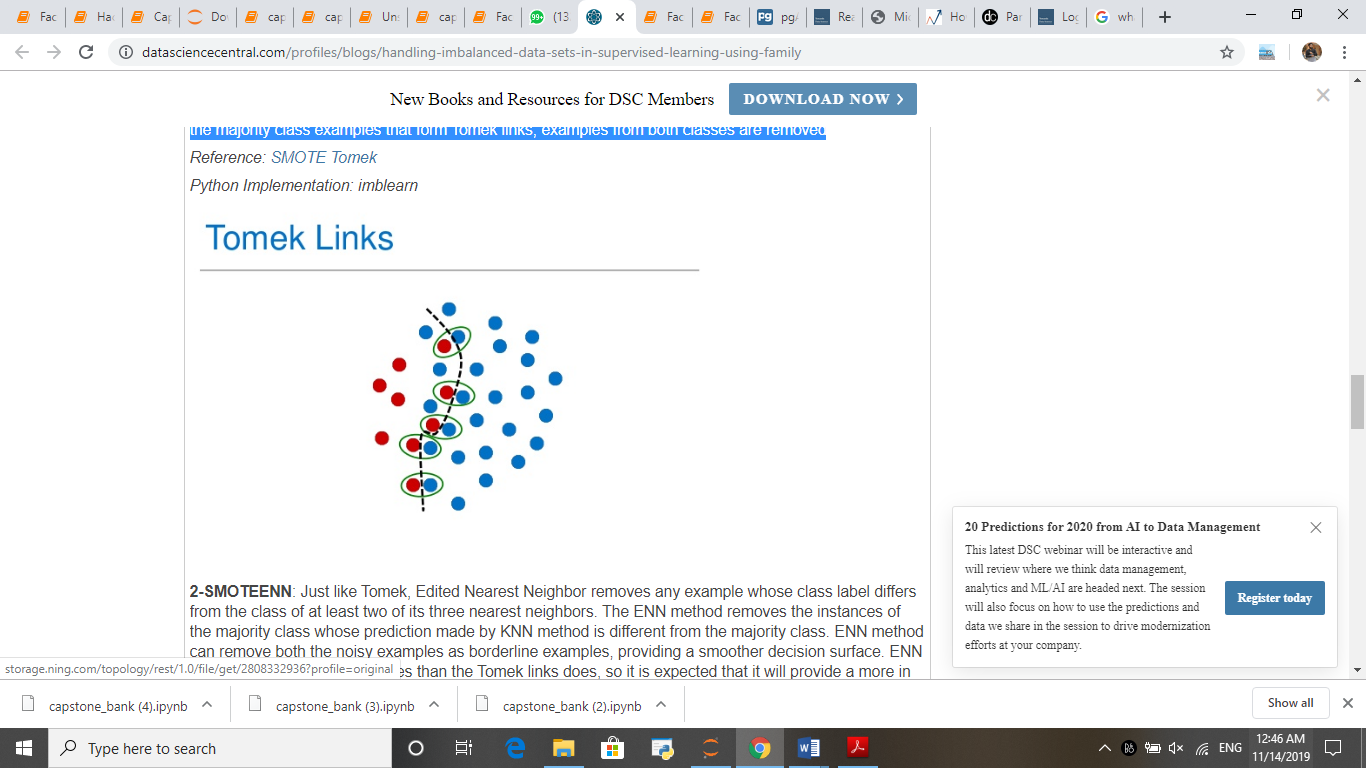
causes the decision boundaries for the minority class to spread further into the majority

class space.

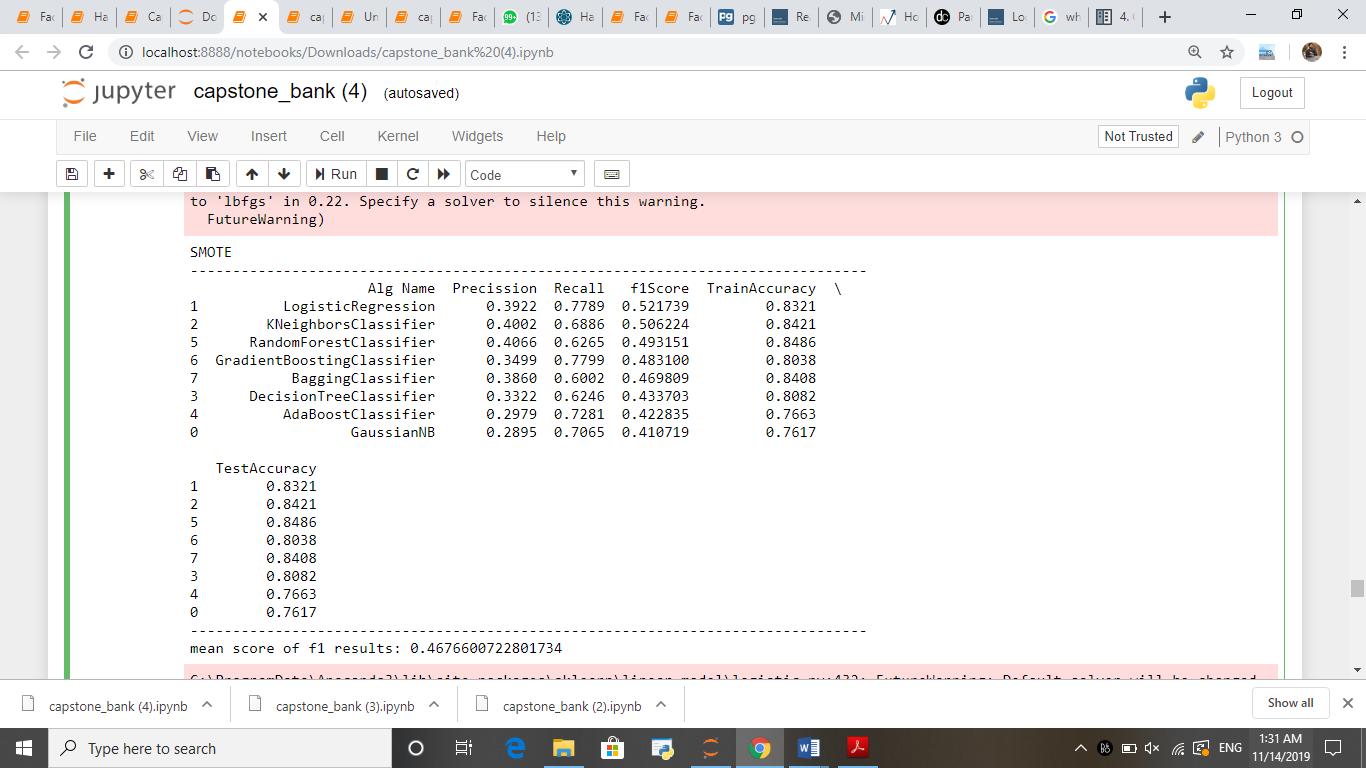


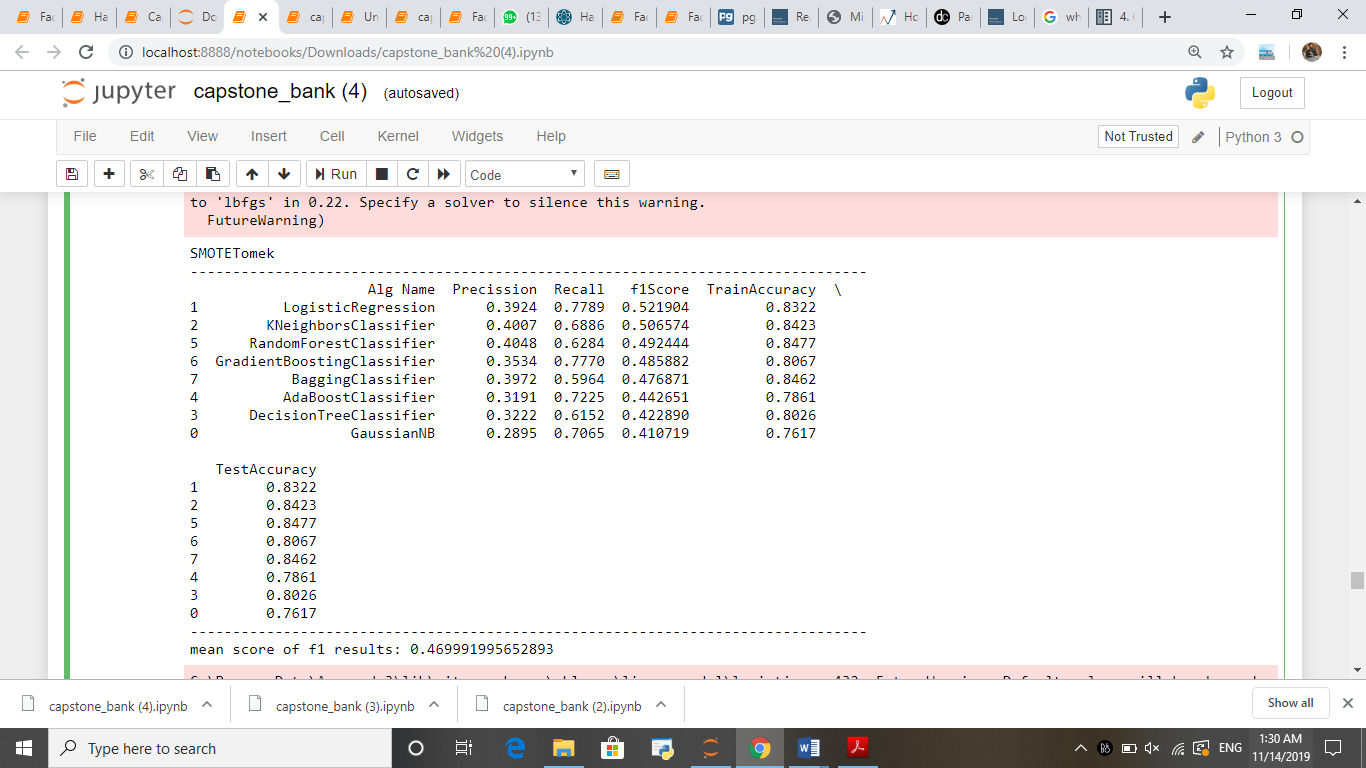
Smotetomek:

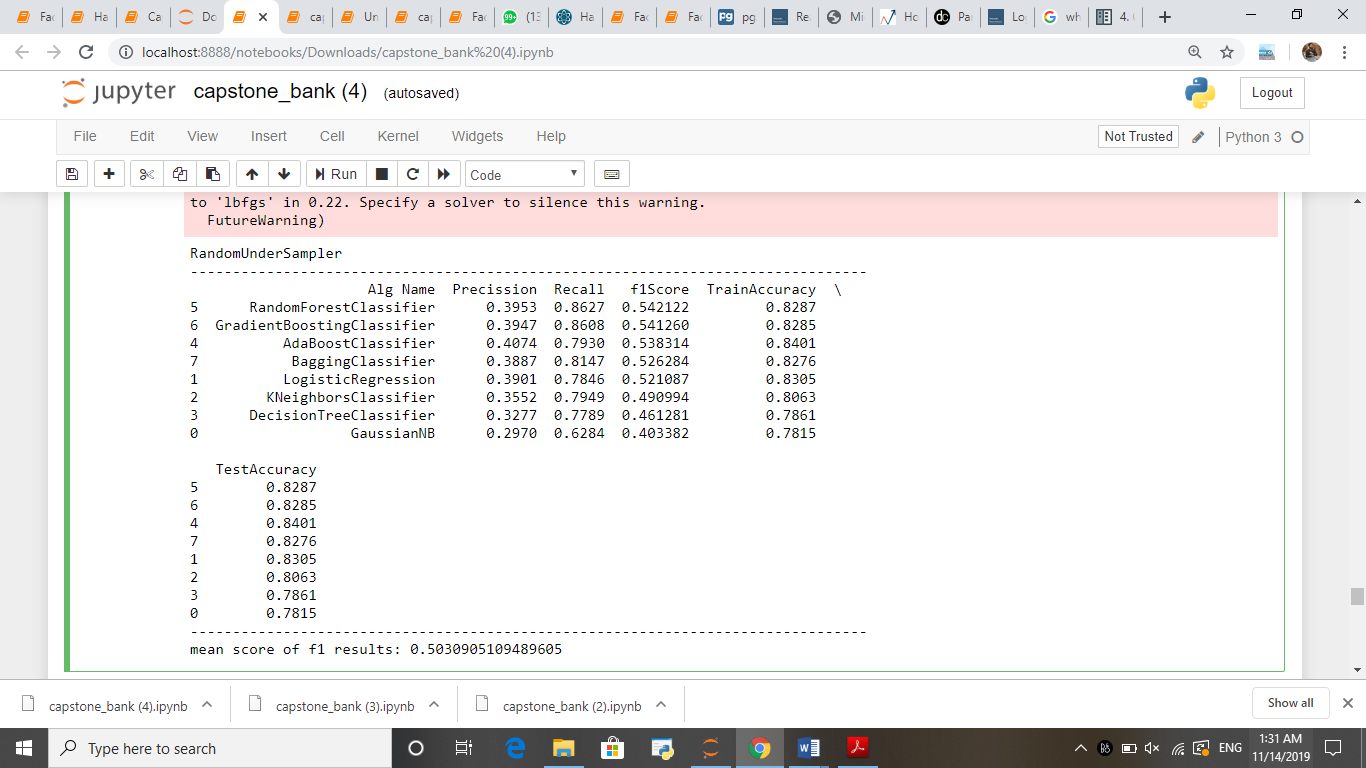
Tomek links can be used as an under-sampling method or as a data cleaning method. Tomek links the over-sampled training set as a data cleaning method. Thus, instead of removing only the majority class examples that form Tomek links, examples from both classes are removed.

****

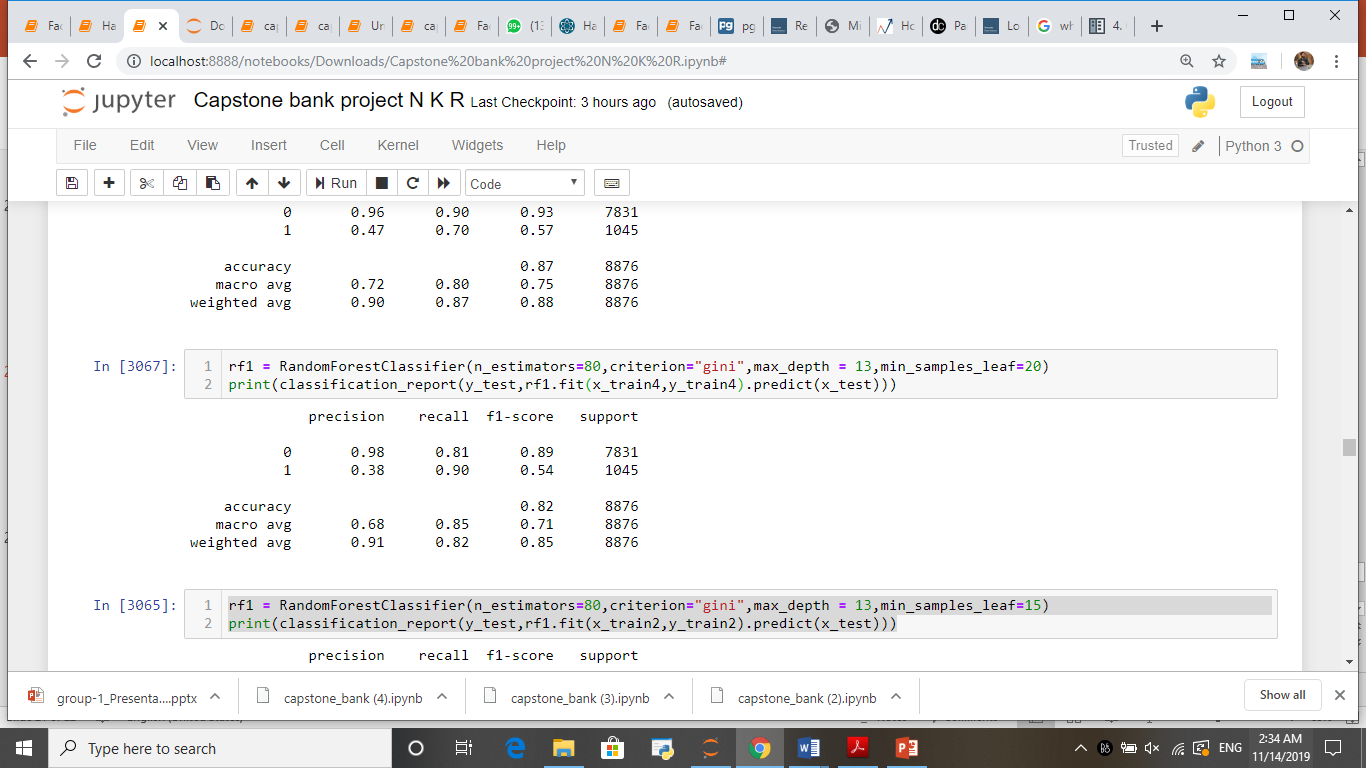
We have performed random under sampling, smote, smotetomek to logistic regression, KNN, Naïve bayes, cart and for ensemble also (Random forest, gradient boosting, adaboost, bagging classifier), checked with mean of f1\_score resulted as follows

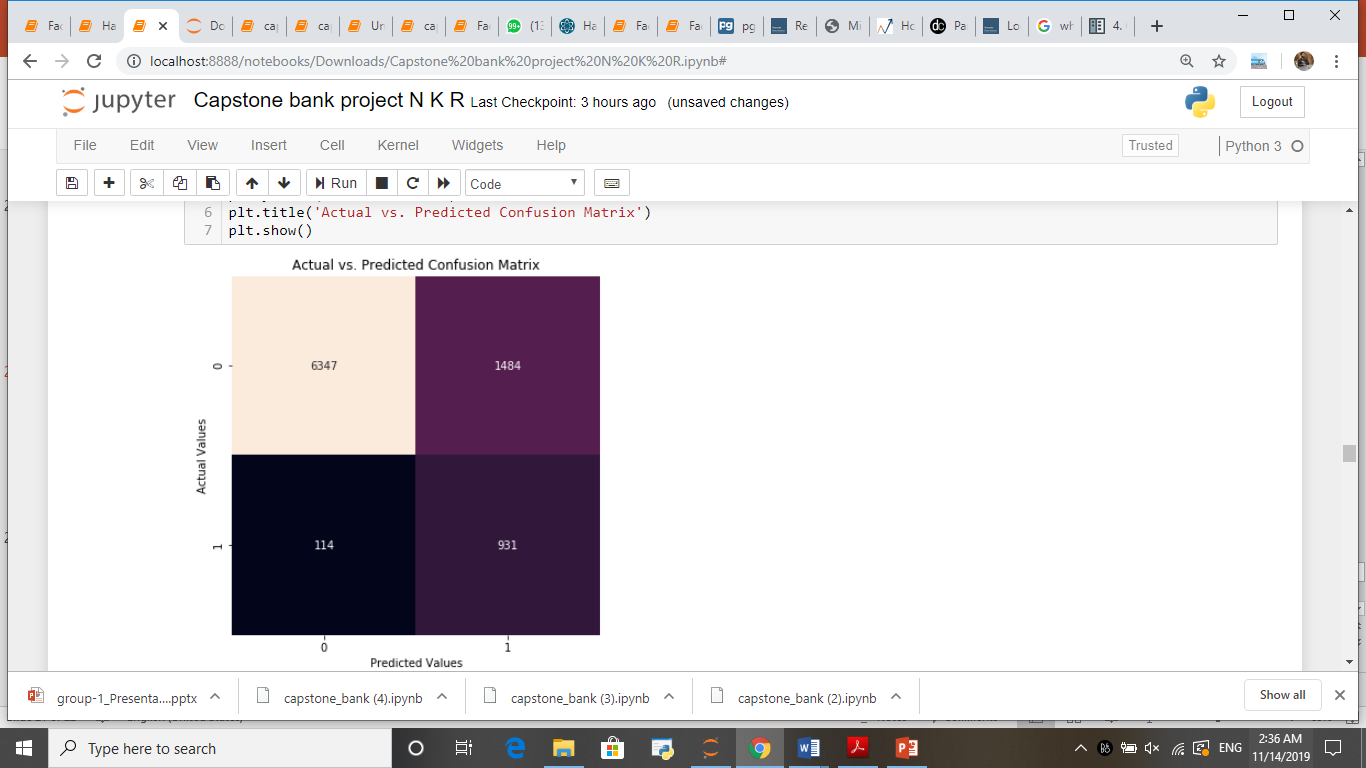






Out of all results we can see that mean f1\_score for random under sampling method is more that is 50.3. Now try resample our data with random under sampling technique. In all models we got random forest model yielding good precision and better f1\_score. So, by using grid search cv we tuned the best parameters in our model and got results which are shown below



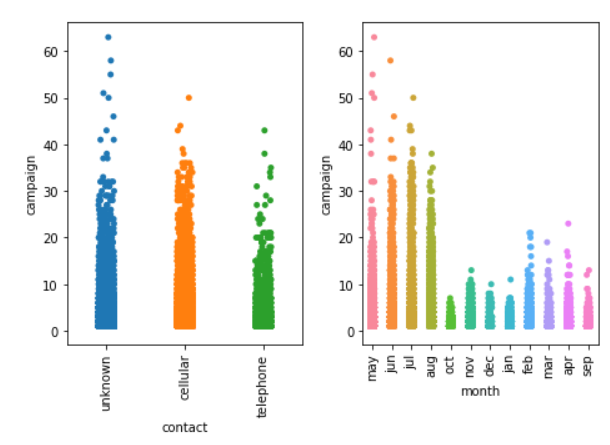
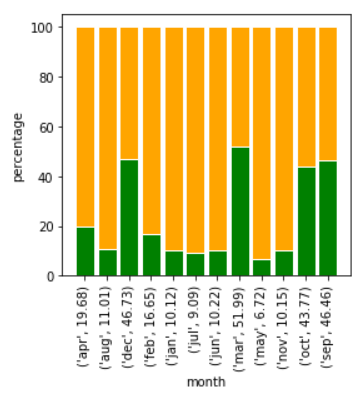


In our base model we got 713 false negatives (type 2 error) i.e. our model predicted the persons who got subscribed for term deposit as not subscribed which are costly in our condition. We focused on reducing those type 2 errors and have done feature engineering, random under sampling and applied ensemble model random forest classifier with hyper parameter tuning and landed with a recall of 90 and 114 false negatives.



# Chapter 4 – Conclusion

Business Solution:

* Though the campaign is in high volume for the month may. The customers did not subscribe much for the term deposits. Months March, October, September and December are the potential months. Future campaigns can be carried during these months.
* Can target old customers.
* An average contact duration of 500 seconds is resulting in positive outcomes.
* Target retired and students during the campaigns.