**Figure List:**

[Figure 1: Correlation matrix between numerical variables and target variable 3](#_Toc91814163)

[Figure 2 Distribution of Month variable 4](#_Toc91814164)

[Figure 3 Distribution of Operating Systems variable 5](#_Toc91814165)

[Figure 4 Distribution of Browser variable 5](#_Toc91814166)

[Figure 5 Distribution of Region variable 6](#_Toc91814167)

[Figure 6 Distribution of Traffic Type variable 6](#_Toc91814168)

[Figure 7 Distribution of Visitor Type variable 7](#_Toc91814169)

[Figure 8a Percentage of total visits according to Weekend variable 7](#_Toc91814170)

[Figure 8b Percentage of only visits ended with transactions according to Weekend variable 7](#_Toc91814171)

[Figure 9 Distribution of Month – Revenue relationship 8](#_Toc91814172)

[Figure 10 Distribution of Special Day - Revenue relationship 8](#_Toc91814173)

[Figure 11 Distribution of Special Day sale - Month relationship 9](#_Toc91814174)

[Figure 12 Feature importance of Random Forest Model 10](#_Toc91814175)

[Figure 13 Feature importance of LightGBM Model 11](#_Toc91814176)

[Figure 14 Feature importance of AdaBoost Model 12](#_Toc91814177)

[Figure 15 Accuracy scores of Hyperparameter Tuned Models 13](#_Toc91814178)

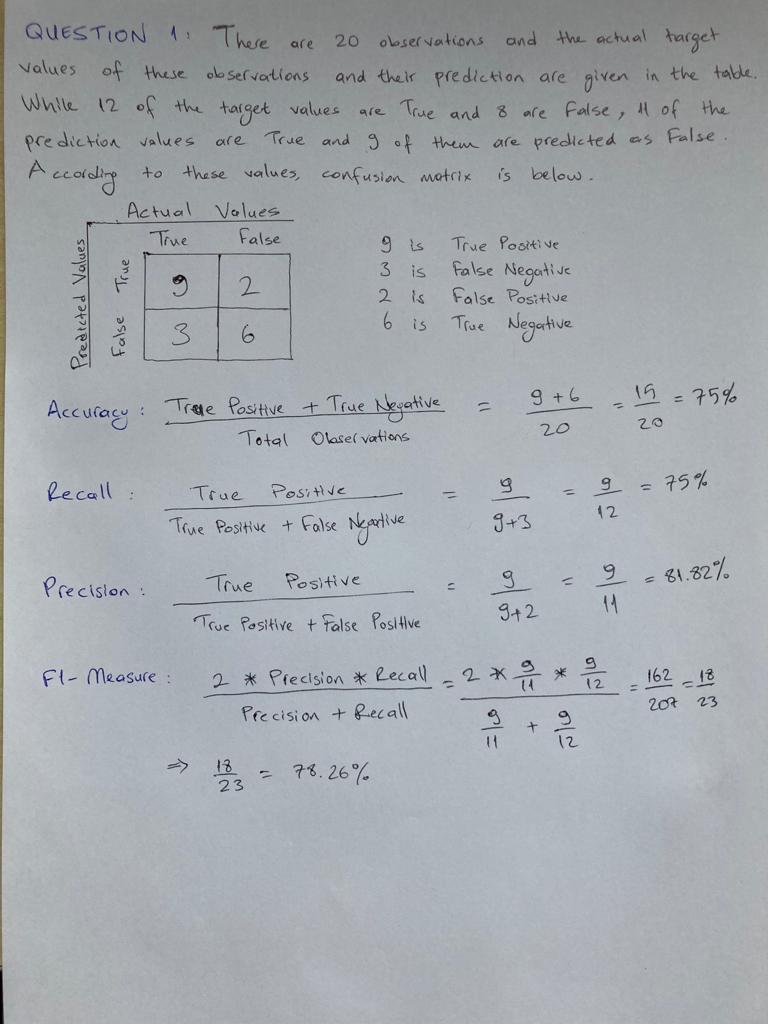
[Figure 16 Recall results of Hyperparameter Tuned Models 13](#_Toc91814179)

[Figure 17 Precision results of Hyperparameter Tuned Models 14](#_Toc91814180)

[Figure 18 F1 scores of Hyperparameter Tuned Models 14](#_Toc91814181)

[Figure 19 Comparisons for ROC Curve of Hyperparameter Tuned Models 15](#_Toc91814182)

**QUESTION 1:**



**QUESTION 2:**

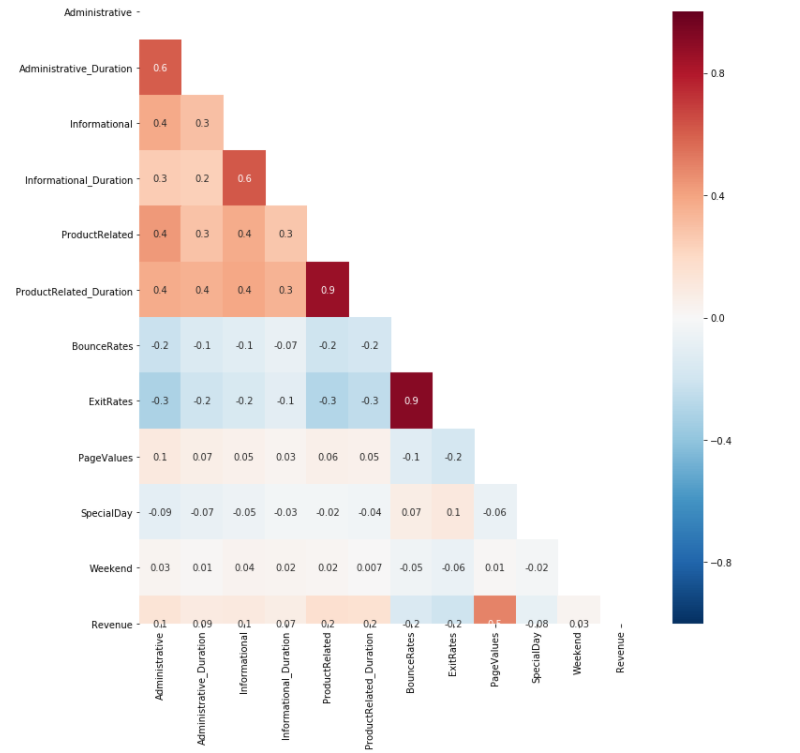
**Abstract**

The Online Shoppers Intention dataset includes feature vectors belonging to 12,330 observations. In order not to show any trend towards a certain campaign, special day, user profile or duration, the data set was created in a way that each session belongs to a different user in a 1-year period. The target variable in the dataset is the Revenue variable. Of the 12,330 observations belong to Revenue variable, 84.5% (10,422) have been negative class samples that were completed without shopping, and the rest (1908) have been positive class samples completed with shopping. This report includes a summary and discussion of the results of the study, which included data exploration, data pre-processing, model implementation, and performance evaluation.

**Data Exploration**

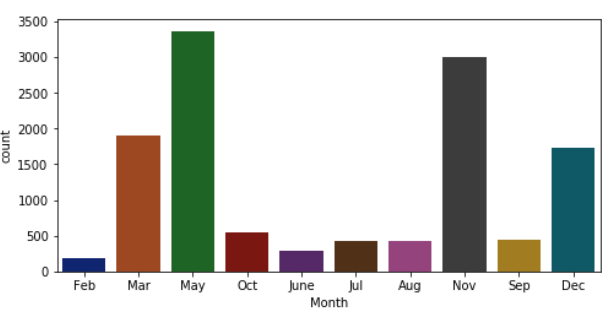
First of all, when the numerical and categorical data in the dataset are examined, it seems that 14 features are of numerical data type, but the feature description actually shows that some of these features are of categorical type. After investigating which data type these features really belong to, it was decided to convert them into categorical data types.

Next, the total number of missing values for each variable was checked and it was observed that there was no missing value in the data.



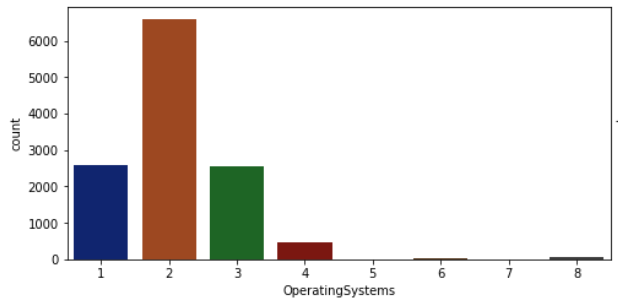
#### Figure 1: Correlation matrix between numerical variables and target variable

Then, in order to better understand the data set, it was tried to obtain deeper information about each feature by using data visualizations. Analysis of correlation was used to measure the relationship between the numerical data and the target variable. A positive correlation is a relationship between two variables in which both variables move in the same direction. A negative correlation is a relationship between two variables in which an increase in one variable is associated with a decrease in the other. The correlation coefficient shows the strength of the relationship between the two variables. According to the correlation table, while PageValues ​​was the variable with the strongest correlation between numerical data, it was observed that the negatively correlated ExitRates variable was an important variable for Revenue result with a correlation value of -0.207.



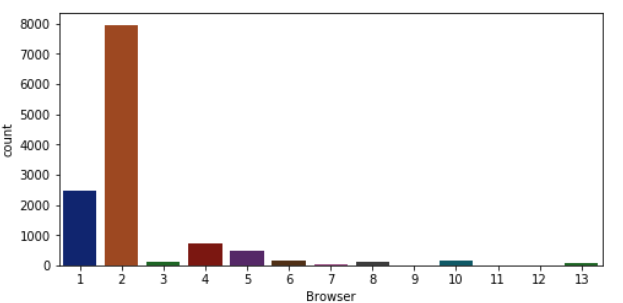
#### Figure 2 Distribution of Month variable

When the Month values ​​shown above are examined, there is data for 10 months, while there is no data for January and April. May and November are the months with the highest visits.



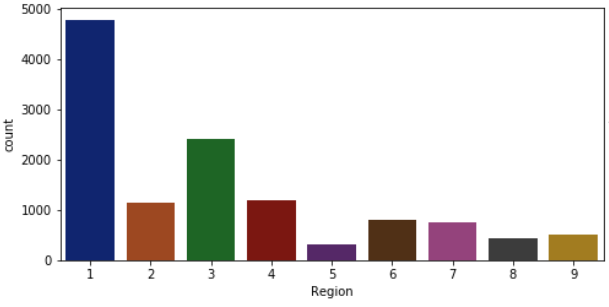
#### Figure 3 Distribution of Operating Systems variable

When the Operating system of the visitor values ​​shown above are examined, Operating system number 2 is the most used value, followed by number 1 and number 3 values respectively.



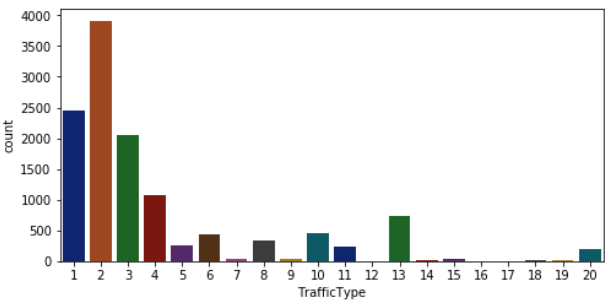
#### Figure 4 Distribution of Browser variable

When the Browser values ​​shown above are examined, the most used browser by far is the number two browser.



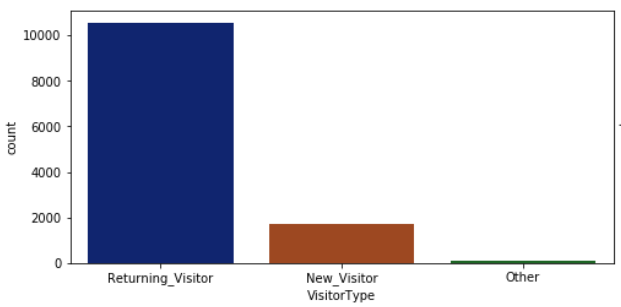
#### Figure 5 Distribution of Region variable

Examining the Region values ​​shown above, the first region is the geographic region where the session is started by the most visitors.



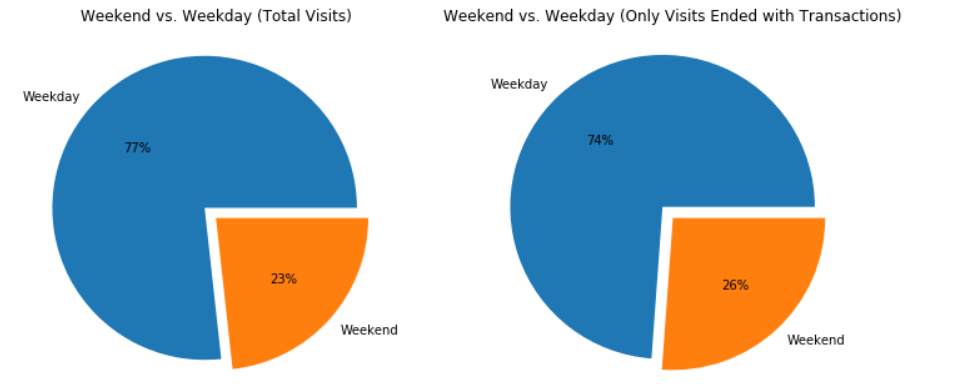
#### Figure 6 Distribution of Traffic Type variable

When the Traffic Type values ​​shown above are examined, the second source is the traffic source that the visitor reached the website the most.



#### Figure 7 Distribution of Visitor Type variable

When the Visitor Type values ​​shown above are examined, it was observed that the visitor type is by far the most Returning Visitor.

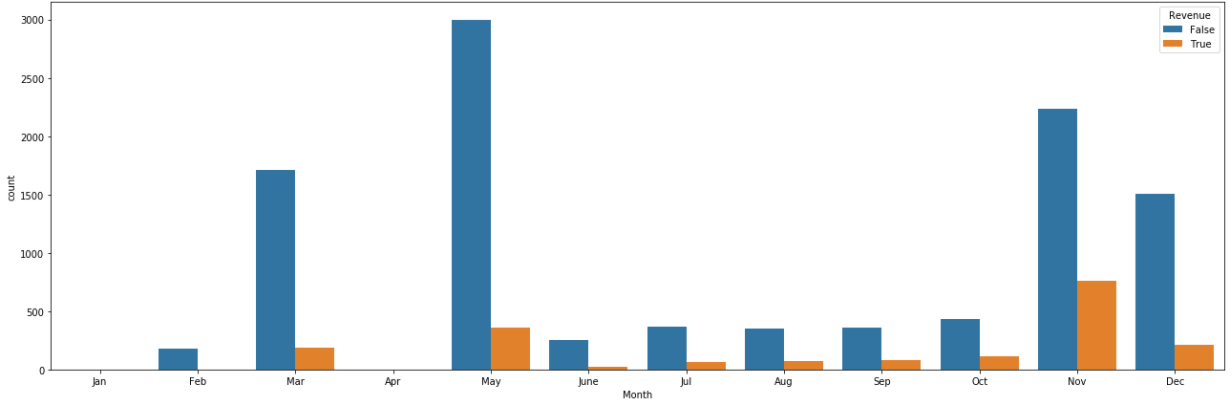


1. (b)

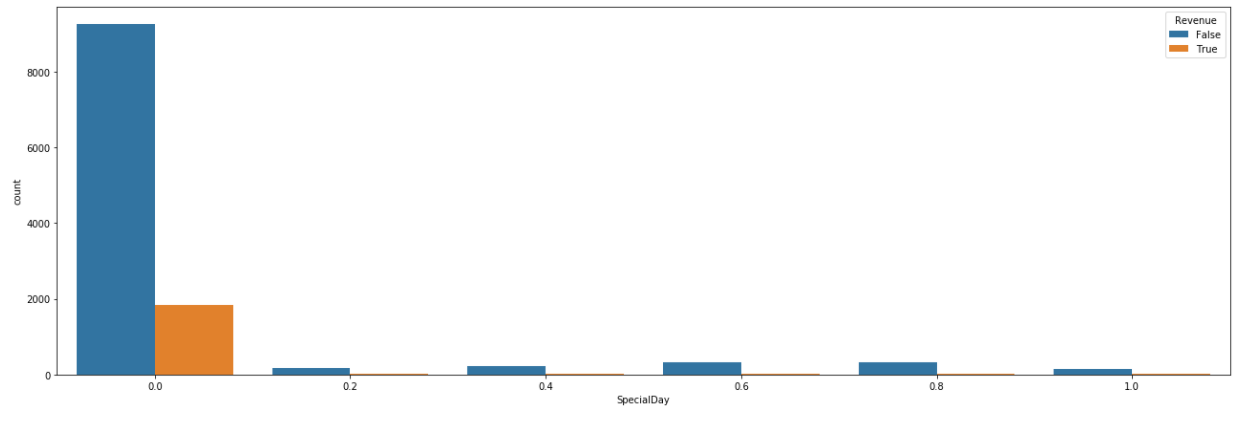
#### Figure 8a Percentage of total visits according to Weekend variable

#### Figure 8b Percentage of only visits ended with transactions according to Weekend variable

When the Weekend values ​​shown above are analyzed, it is observed that there are much more visitors during the weekday than weekend. Moreover, the majority of the times (74%) when the visit is completed with a transaction belong to the weekday.

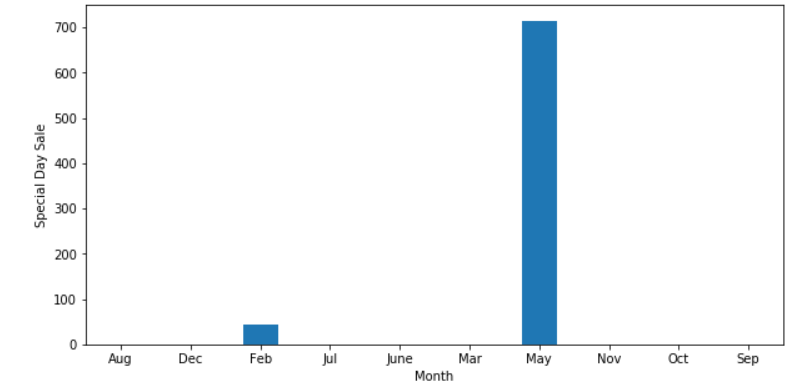


#### Figure 9 Distribution of Month – Revenue relationship

When the Month - Revenue relationship shown above is analysed, the month with the highest number of transactions and visits is November, followed by May. 

#### Figure 10 Distribution of Special Day - Revenue relationship

Examining the Special Day - Revenue relationship shown above, the value of "Special Day" feature is determined by considering the dynamics of e-commerce such as the duration between the order date and delivery date. For example, for Valentina's day, this value takes a nonzero value between February 2 and February 12, zero before and after this date unless it is close to another special day, and its maximum value of 1 on February 8. While it is very unlikely that the page is visited and transaction completion on special days, the number of visits and transaction completions is much higher when it is away from special days.



#### Figure 11 Distribution of Special Day sale - Month relationship

When the Special Day - Month relationship shown above is examined, almost all of the shopping made on special days is made in May, and a small amount is made in February.

**Data Pre-processing**

Outliers were checked for numerical data. Although outlier detection was tried to be performed using the first (0.25) and third (0.75) quartiles with the IQR (Interquartile range) method, an outlier could not be obtained due to the distribution of the data. Here, instead of increasing IQR values, it is planned to reach a solution by choosing models that can be less affected by outliers.

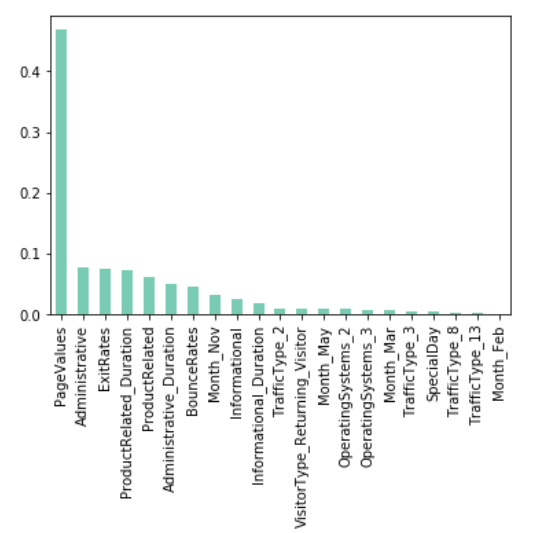
Later, to obtain numerical data from categorical data, encoding should be applied. When the distributions of categorical data are examined, One Hot Encoding was performed because it was thought that applying One Hot Encoding would bring better results. One of the newly formed variables was deleted for each main categorical variable to avoid the curse of dimensionality. From now on, new target feature is Revenue\_True and the number of input variables is 68.

The data were first divided into two as y and x, then the number of input variables was reduced to the 24 most important variables thanks to the threshold values ​​determined by correlation and chi-square statistical methods. The most important variables were preferred because the variables after them had a very low relationship with the target variable. This prevents the model from dealing with insignificant or less important data. Next, thanks to the standard scale, input variable values were reduced to certain ranges before applying machine learning algorithms to them. The y value and the scaled x value are split into train and test. If modelling was started directly after this stage, a higher accuracy score would be obtained due to the imbalance in the target variable, but the accuracy score may not be reliable for imbalanced data. It was aimed to obtain more reliable results by creating synthetic data thanks to SMOTE(Synthetic Minority Oversampling Technique) in train data.

**Model Implementation**

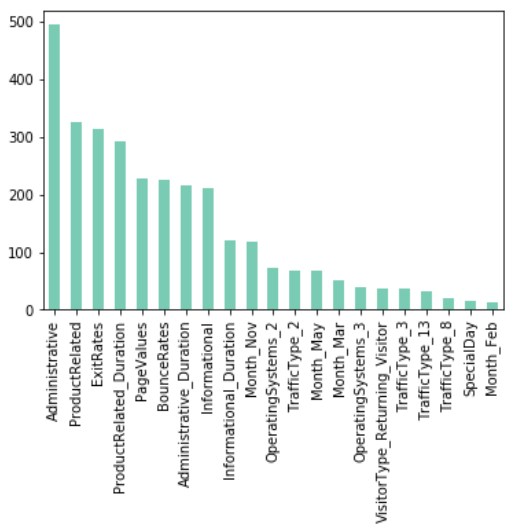
Although many models were tried during the modelling process, the models that were observed to be less affected by not excluding the outliers were preferred. Three models were chosen, Random Forest, LightGBM and AdaBoost. Ensemble learning models were chosen because in this dataset they accurately reflect the general belief that ensemble learning models are less affected by outliers. In addition, the selected models are very cost effective (RF took 32 minutes, LightGBM took 32 minutes, AdaBoost took 25 minutes on the run computer). Besides, the model metric results are quite successful. They will be compared in more detail in the performance evaluation section.

When the feature importance graphs of these models were analysed, the following results were obtained.



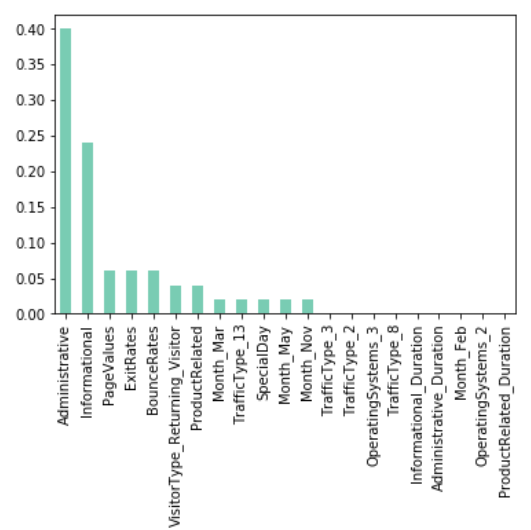
#### Figure 12 Feature importance of Random Forest Model

The most important variable for the Random Forest is the PageValues ​​variable, followed by Administrative and ExitRates.



#### Figure 13 Feature importance of LightGBM Model

The most important variable for LightGBM is the Administrative variable, followed by ProductRelated and ExitRates.

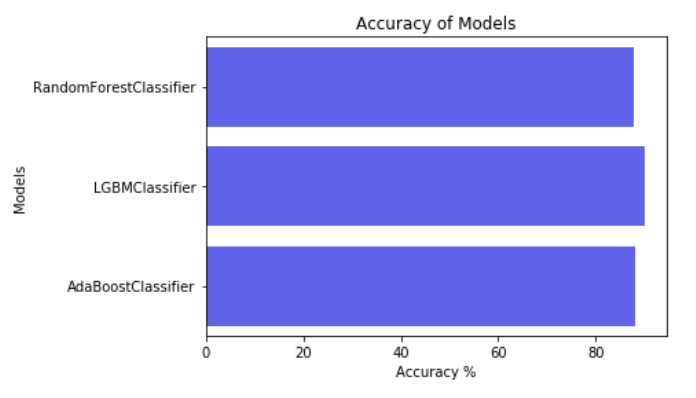


#### Figure 14 Feature importance of AdaBoost Model

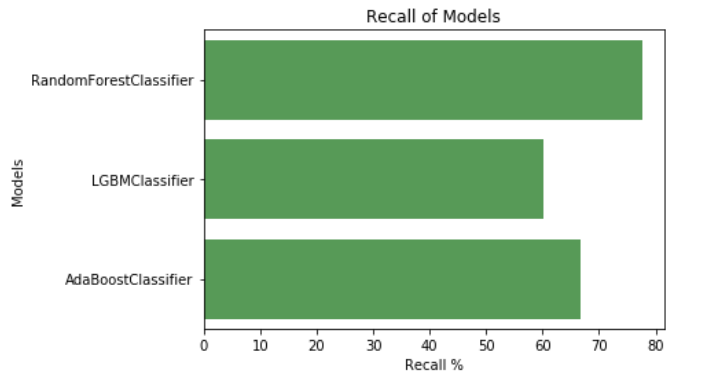
The most important variable for AdaBoost is the Administrative variable, followed by Informational and PageValues.

**Performance Evaluation**

Hyperparameters are essential as they manage the general behaviour of a machine learning model. The main purpose is to obtain an optimal solution of hyperparameters that minimizes a predefined loss function to present higher results. The results of models with hyperparameter tuning will be discussed below.

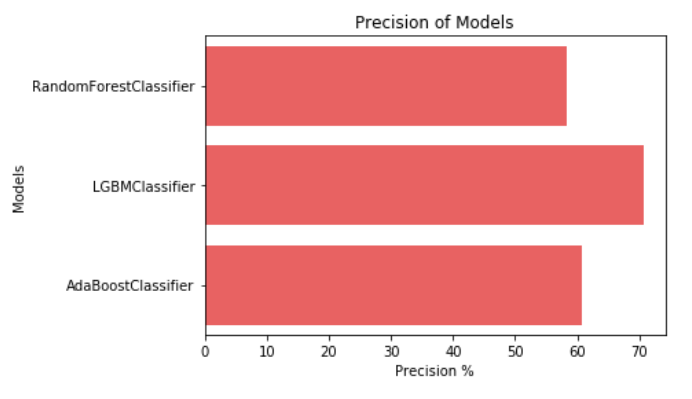


#### Figure 15 Accuracy scores of Hyperparameter Tuned Models

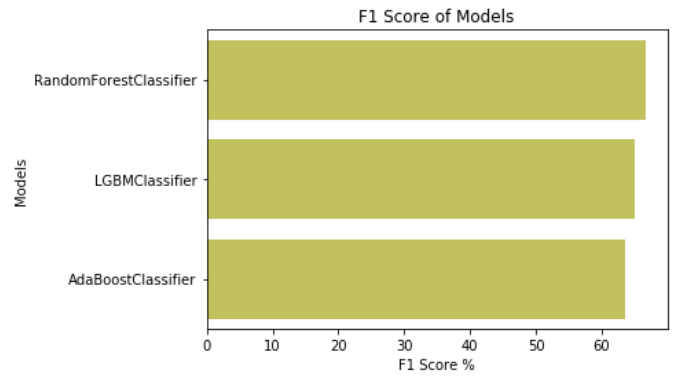
LightGBM(89.65%) has the best accuracy score by a small margin, followed by AdaBoost(87.74) and Random Forest(87.71).

#### Figure 16 Recall results of Hyperparameter Tuned Models

Given its recall values, RF(80.16%) predicts by far the best, followed by AdaBoost(65.03%) and LightGBM(60.33%).

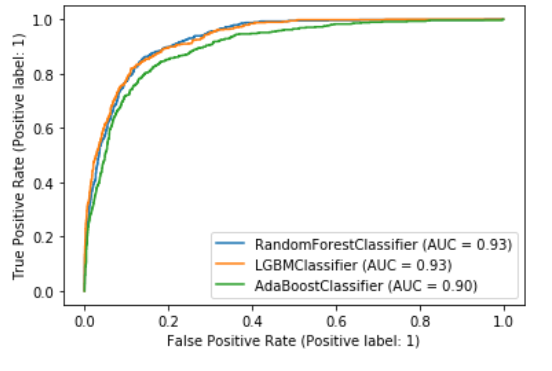


#### Figure 17 Precision results of Hyperparameter Tuned Models

Given the precision values, LightGBM(70.24%) estimates by far the best, followed by AdaBoost(60.57%) and RF(58.16%).

#### Figure 18 F1 scores of Hyperparameter Tuned Models

Considering the F1 score of models, RF(67.41%) has the best result, followed by LightGBM(64.91%) and AdaBoost(62.72%).



#### Figure 19 Comparisons for ROC Curve of Hyperparameter Tuned Models

Considering the ROC Curve of models, RF(93%) and LightGBM(93%) have the best result, followed by AdaBoost(90%).

**Conclusion**

Three different learning classifiers (Random Forest, LightGBM and AdaBoost) were tested and optimized and very close results were obtained. Random Forest model results gave the best results for ROC Curve, F-1 Measurement and Recall, while LightGBM had the best results on Accuracy and Precision metrics, followed by Adaboost.

In general, Random Forest and LightGBM gave the best results for this dataset. If the precision score is more important, the LightGBM model can be used, while if the recall score is more important, the Random Forest model will give the best results. Other metric results are very close.