MEF UNIVERSITY

ONLINE SHOPPING PURCHASING PREDICTION

Capstone Project

İdil Kazezyılmaz

İSTANBUL, 2021

MEF UNIVERSITY

ONLINE SHOPPING PURCHASING PREDICTION

Capstone Project

İdil Kazezyılmaz

Advisor: Asst. Prof. Dr. Evren Güney

İSTANBUL, 2021

MEF UNIVERSITY

Name of the project: Online Shopping Purchasing Prediction
Name/Last Name of the Student: İdil Kazezyılmaz
Date of Thesis Defense: 30/08/2021

I hereby state that the graduation project prepared by İdil Kazezyılmaz has been completed under my supervision. I accept this work as a "Graduation Project".

30/08/2021
Asst. Prof. Dr. Evren Güney

I hereby state that I have examined this graduation project by İdil Kazezyılmaz which is accepted by his supervisor. This work is acceptable as a graduation project and the student is eligible to take the graduation project examination.

30/08/2021

Prof. Dr. Özgür Özlük

Director
of
Big Data Analytics Program

We hereby state that we have held the graduation examination of _____ and agree that the student has satisfied all requirements.

THE EXAMINATION COMMITTEE

Committee Member	Signature
1. Asst. Prof. Dr. Evren Güney	…….
2. Prof. Dr. Özgür Özlük	……..

ACADEMIC HONESTY PLEDGE

I promise not to collaborate with anyone, not to seek or accept any outside help, and not to give any help to others.

I understand that all resources in print or on the web must be explicitly cited.

In keeping with MEF University's ideals, I pledge that this work is my own and that I have neither given nor received inappropriate assistance in preparing it.

Name	Date	Signature
İdil Kazezyılmaz	31.08.2021	

EXECUTIVE SUMMARY

ONLINE SHOPPING PURCHASING PREDICTION

İdil Kazezyılmaz

Advisor: Asst. Prof. Dr. Evren Güney

AUGUST, 2021, 49 pages

This project aims to understand the purchasing behavior of the consumers and make predictions about purchasing according to website metrics such as page values, bounce rates.

An existing dataset is used in this project. This dataset is available in the collection of data from an e-commerce website by Google Analytics, which consists of 10 numerical and 8 categorical attributes coming from 12,330 sessions. The 'Revenue' attribute is used as the class label. The attributes that have high impact on the prediction are; "Administrative", "Administrative Duration", "Informational", "Informational Duration", "Product Related" and "Product-Related Duration". They represent the number of different types of pages visited by the visitor in that session and the total time spent in each of these page categories.

The "Bounce Rate", "Exit Rate" and "Page Value" features represent the metrics measured by Google Analytics for each page in the e-commerce site. The "Special Day " feature indicates the closeness of the site visiting time to a specific special day (e.g. Mother’s Day, Valentine's Day) in which the sessions are more likely to be finalized with a transaction.

Since the purpose of this project is to predict potential purchasing using existing data, in the prediction part several machine learning algorithms such as decision trees, random forests will be applied to compare the models. The most suitable model will be chosen among these algorithms.

Key Words: E-commerce, online shopping, user behavior, shopping intention, machine learning, real-time shopping behavior, shopping purchase prediction

ÖZET

ONLİNE ALIŞVERİŞ TAHMİNLEMESİ

İdil Kazezyılmaz

Proje Danışmanı: Dr. Öğr. Üyesi Evren Güney

AĞUSTOS, 2021, 49 Sayfa

Bu proje, tüketicilerin satın alma davranışlarını anlamayı ve sayfa değerleri, hemen çıkma oranları gibi web sitesi metriklerine göre satın alma ile ilgili tahminlerde bulunmayı amaçlamaktadır.

Bu proje için hazır bir veri seti kullanılmıştır. Veri seti, Google Analytics aracılığıyla e-ticaret web sitesindeki verilerin toplanmasıyla oluşturulmuştur. Veri seti 10 sayısal ve 8 kategorik veriden oluşmaktadır. Veri setinde 12.330 değer bulunmaktadır. 'Satın Alma' özelliği, sınıflandırma etiketi olarak kullanılmaktadır. Tahmin üzerinde önemli etkisi olacak veriler; "Administrative", "Administrative Duration", "Informational", "Informational Duration", "Product Related" ve "Product-Related Duration". Bu veriler ziyaretçinin o oturumda ziyaret ettiği farklı türdeki sayfaların sayısını ve bu sayfa kategorilerinin her birinde harcanan toplam süreyi göstermektedirler.

"Hemen Çıkma Oranı", "Çıkış Oranı" ve "Sayfa Değeri" verileri, e-ticaret sitesindeki her sayfa için Google Analytics tarafından ölçülen metrikleri temsil etmektedir. "Özel Gün" verisi, site ziyaret saatinin, oturumların bir işlemle sonuçlanma olasılığının daha yüksek olduğu belirli bir özel güne (örneğin Anneler Günü, Sevgililer Günü) yakınlığını gösterir.

Bu projenin amacı, mevcut verileri kullanarak potansiyel satın alma tahminini yapmak olduğundan, tahmin bölümünde, modelleri karşılaştırmak için karar ağaçları, rastgele ormanlar gibi çeşitli makine öğrenme algoritmaları uygulanacaktır. Bu algoritmalar arasından en uygun model seçilecektir.

Anahtar Kelimeler: E-ticaret, online alışveriş, kullanıcı davranışı, alışveriş yapma eğilimi, makine öğrenmesi, gerçek zamanlı alışveriş davranışı, alışverişte satın alma tahmini

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

In recent years, consumers have started to shop more and more on e-commerce sites. With the effect of the pandemic, we can say that this habit has increased even more in the last 2 years. According to Statista Research Department (2021), e-commerce in the United States will increase almost 20% from 2021 to 2025.

In Turkey companies like Trendyol.com, Hepsiburada.com, N11.com have become the most revenue-generating e-commerce platforms. According to TUBİSAD retail e-commerce increased by 42.5% in 2019 compared to the previous year and reached 44.9 billion TL and its share in total retail trade increased to 6.2%.

Again according to the estimation made by Statista, the number of online shoppers in Turkey by 2023 will increase to 44.4 million. And average annual orders of consumers are expected to increase to $436.

This shows us that more brands will head online and as a consequence, the competition will ramp up. For that, it will be crucial to understand consumers' intentions. This paper tries to understand the purchasing behavior of the consumers and make predictions about purchasing according to website metrics such as page values, bounce rates.For this project supervised learning methods will be used. Machine learning algorithms such as Random Forest, Gradient Boosting, XGBoost, LightGBM, Logistic Regression, Support Vector Machine will be applied and compared. For comparison the models will be applied to the imbalance dataset, afterwards the data will be reproduced synthetically to see the difference. Additionally the statistical analysis and visualizations will be applied to the dataset.

2. LITERATURE SURVEY

There are several types of research about online shopping purchasing prediction. Karim Baati & Mouad Mohsil's (2020) study tries to predict real-time online shopping behavior with the session and the visitor information. For machine learning methods they have used the Naive Bayes classifier, C4.5 decision tree, and random forest. Among these models, they have a conclusion that the random forest model is the most suitable method to solve this problem and uses significantly higher accuracy and F1 Score than the other models. They have proved that to forecast the visitor's shopping intent as soon as the e-commerce website is visited.

On the other hand, Xiao (2020) tried models like Decision Tree, k-NN, Logistic Regression, and Naïve Bayes for online shopping intentions. The research compares the models to their accuracy, recall, precision, and F1 score. For accuracy, the best model is logistic regression but accuracy is not a very valid metric for the dataset. Because of the imbalanced data they choose the Naive Bayes model for its recall metric.

It is important to find a suitable model for the project. But another important factor is feature selection. Since not every feature will be as equally important and may harm the accuracy of the models a proper feature selection process should be applied. Poel and Buckinx (2005) used AUC scores to identify the importance of the variable types. Even though in their study the results highlight that predictors from all four categories are important, they have a conclusion that not every attribute is equally valuable; some of the features are the most important ones.

There are seven types of supervised learning methods. Random forest is one of them which is used as a classifier in the study of Joshi et al. (2018). In their study, they tried to understand different metrics that can affect the online buying behavior of Indian customers. They have used Random Forest models for each product category to set up the customer behavior effect on multichannel retailers. Their research contributes to the theoretical domain in terms of the interplay of the factors and their impact on online and offline buying behaviors.

Another study that used the Random Forest model is the study of Beck (2021). Besides the Random Forest model, a linear classification model was applied to the data. Although the linear classification model has a higher AUC score, in accuracy again the Random Forest model is more suitable for this research. But there is another model which is applied after the feature selection process. Both the Random Forest model and XGBoost's Classifier applied after the

feature selection process and the result is that with the XGB classifier the accuracy score is almost 90%.

For modeling with imbalanced datasets Dataman, D. (2021) provides several options in the article. In this project two of the methods on that article were used for modeling. These are SMOTHE and Random OverSampling. SMOTHE is used for all models, Random Over is just used for Random Forest for the comparison.

The SMOTHE method is commonly used in classification imbalance. Like in the article of A. Amin et al. (2016) this method is used to overcome the overfitting issue while reproducing the minority class samples. The reproduced data is learned from the existing dataset while it is created randomly.

As Last, F. et al. (2017) mentioned in the study other methods besides SMOTHE can generate noise in the model. Unlike other models SMOTHE is simple and very effective to overcome this imbalanced data issue.

After the oversampling method the main discussion is which model is suitable for the project. Previous studies examined Random Forest, Logistic Regression and XGBoost. But there aren't any other boosting models. For evaluating the performance of the models Daoud, E. A (2019) compared XGBoost, LightGBM and CatBoost. In the study it has shown that the LightGBM model has better performance in terms time and accuracy.

Evaluation metrics for the models are also important. There are accuracy, recall, precision and F1 scores. As Shung, K. P. (2020) explained in his article this will depend on the problem to be solved but simply one metric might not be enough to evaluate the model. For classification problems, it is important to know how models are classified data. That is why like in this study, besides the accuracy, recall and precision should be taken into consideration.

3. DATA ANALYSIS

3.1. Data Information

The dataset for this project is found from Sakar, Polat & Katircioglu's (2018) study. This ready dataset is used in their research so it was formed data. With this, each session would belong to a different user in 1 year to avoid any tendency to a specific campaign, special day, user profile, or period. In the dataset, there are 12,330 sessions. There are 10 numerical and 8 categorical features. The 'Revenue' feature is used as the class label.

There are 3 different page categories. These are "Administrative", "Informational" and "Product Related". Additionally, we have the duration information of the user for these pages. The dataset has "Bounce Rate", "Exit Rate" and "Page Value" features which are measured by Google Analytics. The "Special Day" feature represents the closeness of the site visiting time to a specific special day. For instance, let's take a look at Valentine'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. There is also information about the operating system, browser, region, traffic type. But this information is not expressed clearly. For example, in the region feature, we don't know if the user is coming from the Marmara region instead there are numbers to group these features. This information is limiting us to understand the users better.

Apart from that, there is a visitor type as returning or new visitor. And boolean values indicating whether the date of the visit is weekend, and month of the year.

3.2. Exploratory Data Analysis

The dataset is ready data and manipulated for the research there aren't any missing data or NA values. So first we can take a look at the statistics of the numerical values which can be seen in Table 1. At a glance, we can see that product-related duration is higher on average than the other page durations.

When we look at the categorical values we can say that there are 8 different operating systems, 9 different regions, 13 browser types, 20 traffic types, 3 visitor types. The observations seem to be held in 10 months. We have revenue as a class label which shows if the transaction

is completed or not. Finally, we can see a weekend attribute, a boolean value that shows whether the transaction is made at the weekend.

Table 1: Statistics of Numerical Values

	count	mean	std	min	25%	50%	75%	max
Administrative	12330.0	2.315166	3.321784	0.0	0.000000	1.000000	4.000000	27.000000
Administrative_Duration	12330.0	80.818611	176.779107	0.0	0.000000	7.500000	93.256250	3398.750000
Informational	12330.0	0.503569	1.270156	0.0	0.000000	0.000000	0.000000	24.000000
Informational_Duration	12330.0	34.472398	140.749294	0.0	0.000000	0.000000	0.000000	2549.375000
ProductRelated	12330.0	31.731468	44.475503	0.0	7.000000	18.000000	38.000000	705.000000
ProductRelated_Duration	12330.0	1194.746220	1913.669288	0.0	184.137500	598.936905	1464.157213	63973.522230
BounceRates	12330.0	0.022191	0.048488	0.0	0.000000	0.003112	0.016813	0.200000
ExitRates	12330.0	0.043073	0.048597	0.0	0.014286	0.025156	0.050000	0.200000
PageValues	12330.0	5.889258	18.568437	0.0	0.000000	0.000000	0.000000	361.763742
SpecialDay	12330.0	0.061427	0.198917	0.0	0.000000	0.000000	0.000000	1.000000

The revenue attribute has two classes: true and false. True shows that the transaction is completed and false shows that there is no transaction. When we count these two classes we can see that there are 10.422 incomplete transactions and 1.908 completed transactions.

If we look at the visits for weekdays and weekends. There are 9462 visits on weekdays and 2868 visits on the weekends. There are 10.551 returning visitors, 1.694 new visitors, and 85 other types of visitors for visitor type.

When we look at the number of transactions monthly it is clear that March, May, November, and December have the highest volume. Special days may have an increasing effect on the volume of these months. For instance, the 8th of March is World Women's Day, Mother’s Day on the first Sunday of May, Black Friday in November, and New Year in December (Figure 1).

Interestingly, in June, when Father’s Day takes place, there are much fewer transactions than in May when Mother's Day takes place. In this case, it can be said that on Mother’s Day users shop more rather than on Father’s Day.

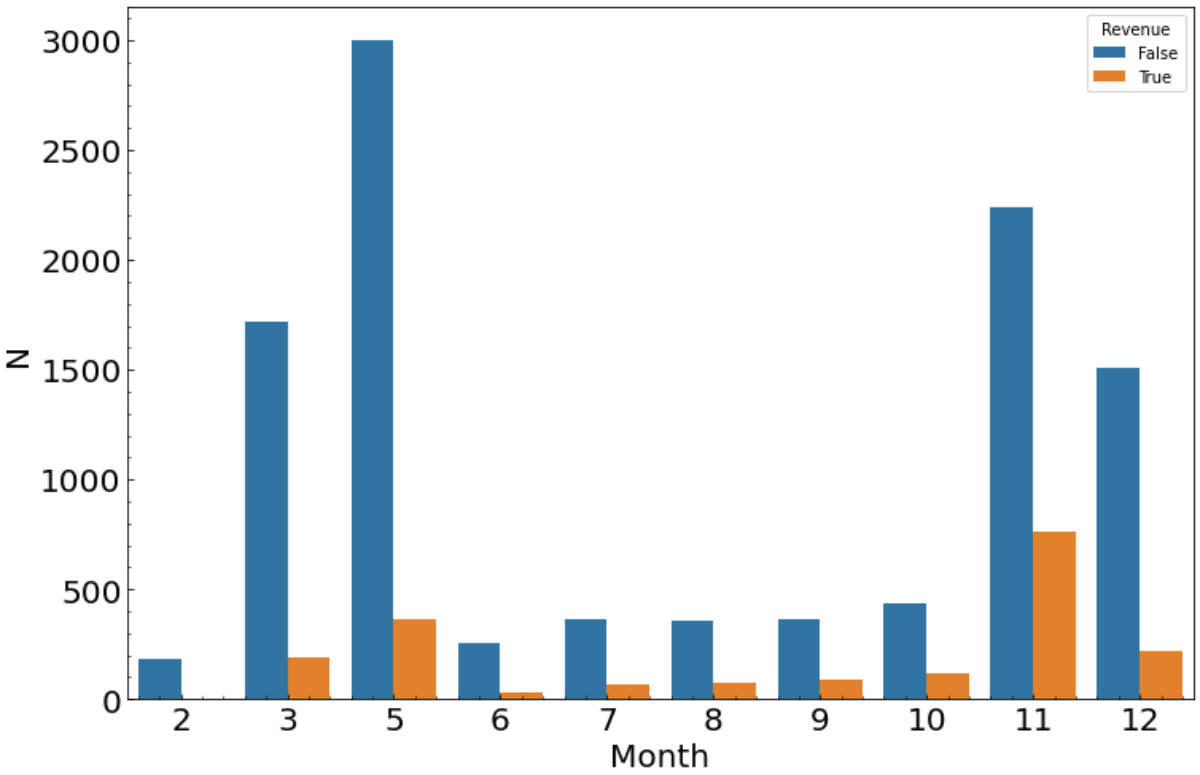


Figure 1: Number of Transactions per Month

When we look at the number of product-related pages visited per month we can see some parallelism with the number of transactions. In November, which is the highest rate of completed transactions, the number of product-related pages visited is significantly higher than the other months (Figure 2).

If we break the transactions according to the visitor types, we can say that returning visitors have more intention to purchase a product. Generally, users tend to research before they purchase a product. So it is expected to not have any transaction for both visitor types (Figure 3).

To see when the users make online shopping we can break it down as weekends and weekdays. As can be seen in Figure 4 on weekdays the volume of transactions is higher than on the weekend. A recent search conducted by RTB house also showed that Turkish consumers are shopping online more on weekdays than on the weekend.

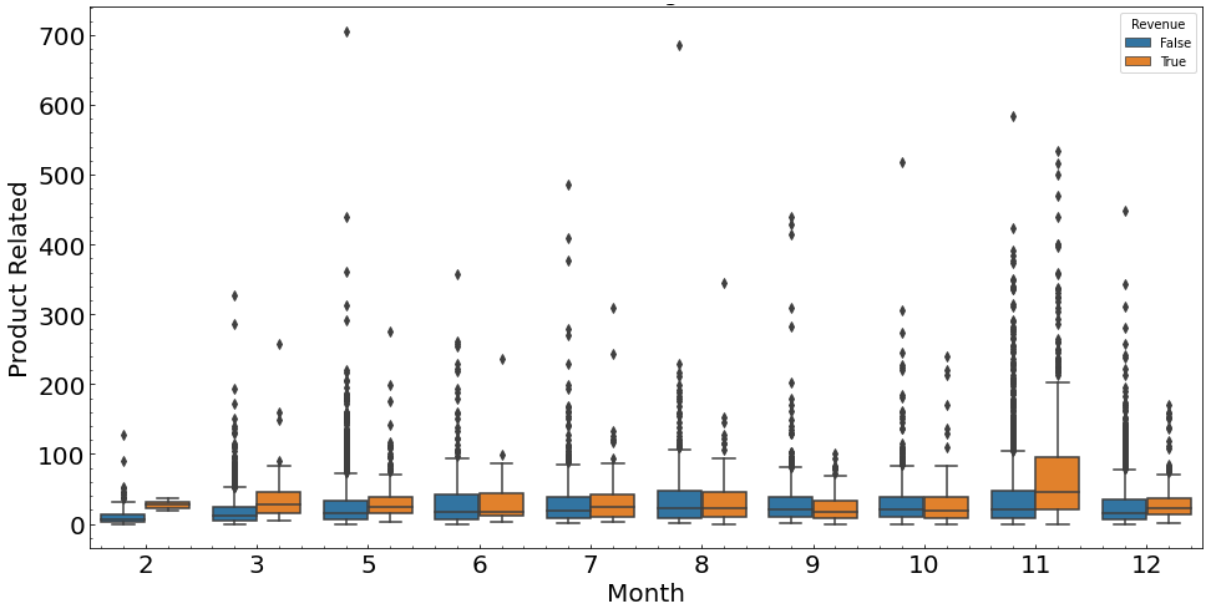


Figure 2: Number of Product-Related Pages Visited per Month

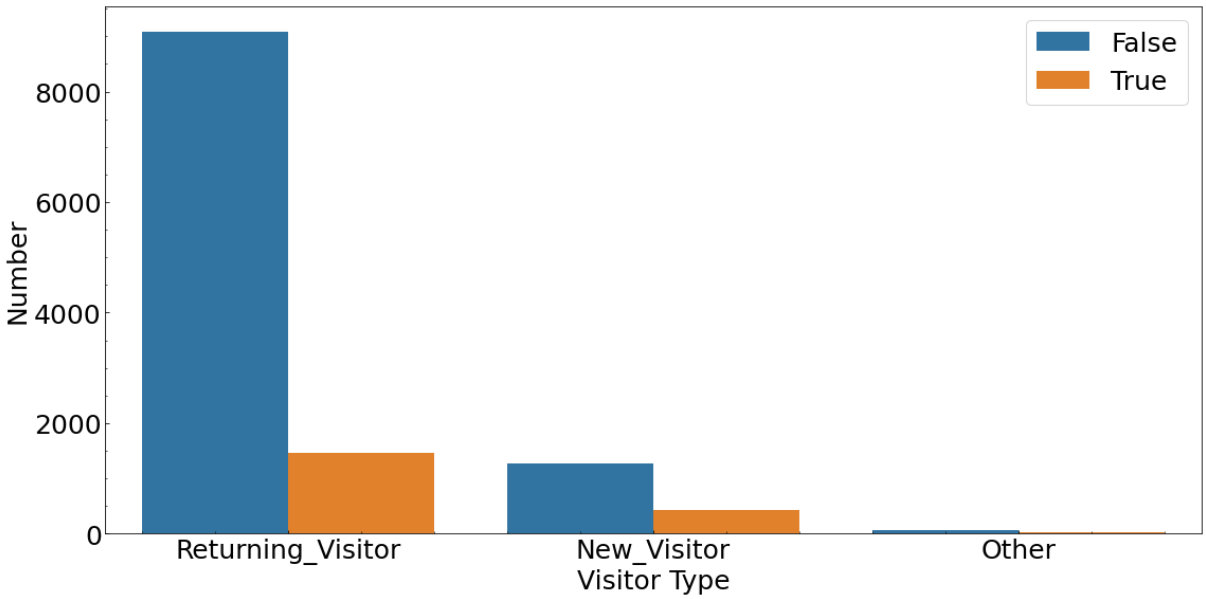


Figure 3: Number of Transactions According to Visitor Type

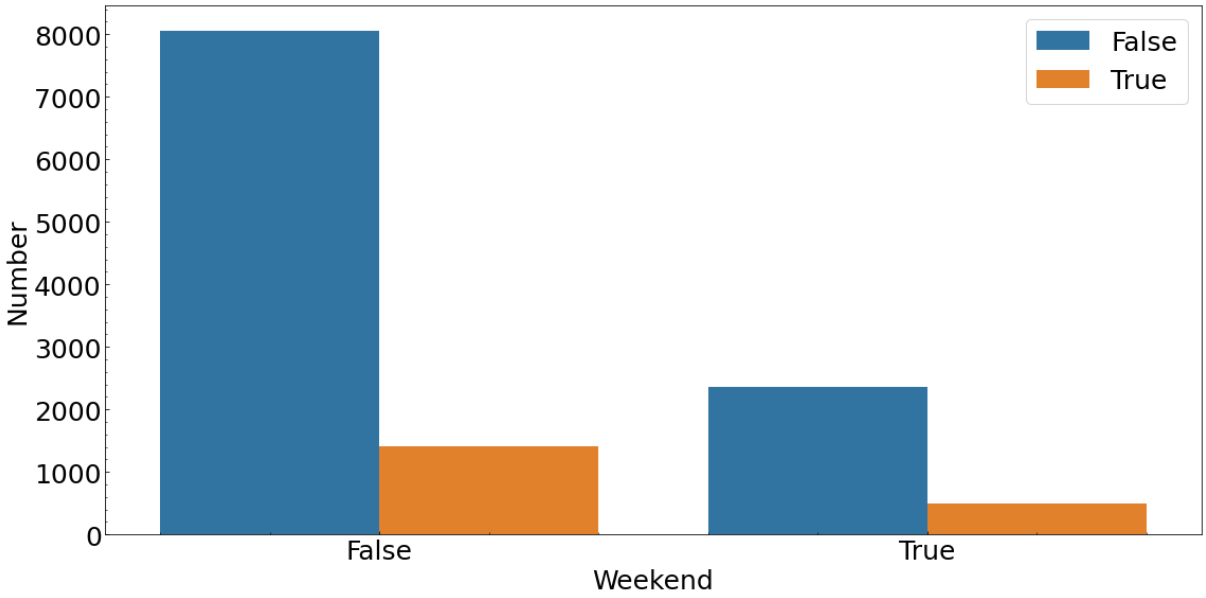


Figure 4: Transactions Based on Weekend

To further understand the user behavior we examine the different page categories. When we look at the administrative pages, which can be considered as the page where users can check older orders and payment methods, users who do not make purchases don't need to check these pages. For some of the users who make purchases also don't check administrative pages or mostly they have viewed one page (Figure 5).

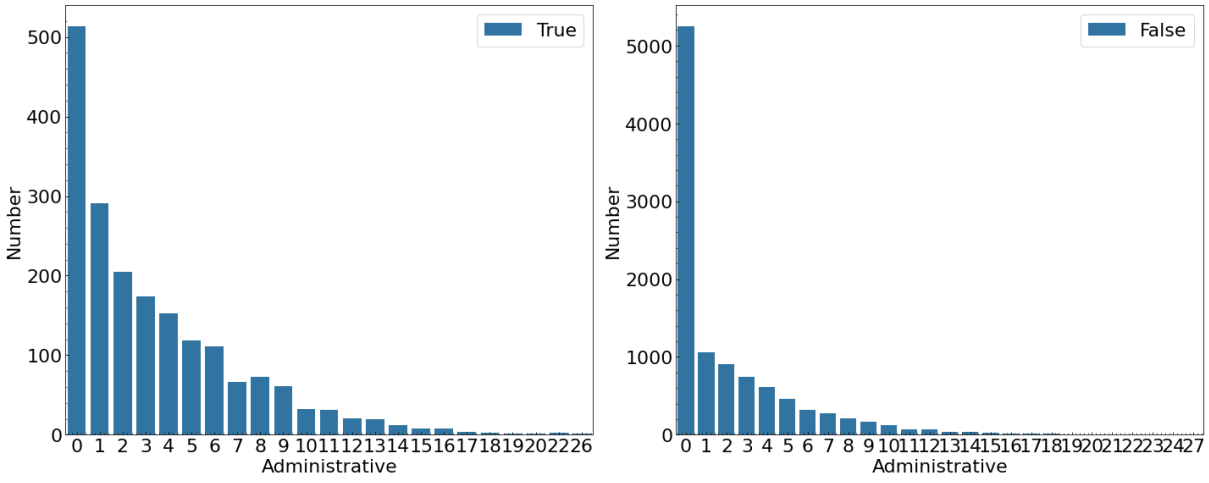


Figure 5: Number of Administrative Pages Visited

And users who visited this page and did not make any purchase spent almost 1.2 minutes on these pages. But users who visited this page and made a purchase spent about 2 minutes on these pages (Figure 6).

If we look at the informational pages, which can be considered as the page where users can check the contact information for the website if there is no transaction users tend to visit this page less. This can be considered as either user is checking the credibility of the website or the transaction has been made before and users would like to return the products (Figure 7).

That is why users with no transaction spent almost 30 seconds on informational pages. But the users who had a transaction spent almost 1 minute on these pages (Figure 8).

For product-related pages, surprisingly users who do not have transactions tend to visit product-related pages more than the users who have transactions. This can be explained as users like to have more information about the product before they purchase them (Figure 9).

There is a parallelism between the time spent in the pages and causality explained above. When users don't make transactions they tend to spend more time on product-related pages (Figure 10).

When we take a look at the bounce rates, users with no transaction tend to bounce from the page more than the users who have made a purchase. For e-commerce websites, it is expected to have a bounce rate on average of 20-40%. If it is higher than 35-55% it can be considered as there might be a problem with the website. In here we have 20% for those who didn't make a purchase. So it can be considered as a normal rate (Figure 11).

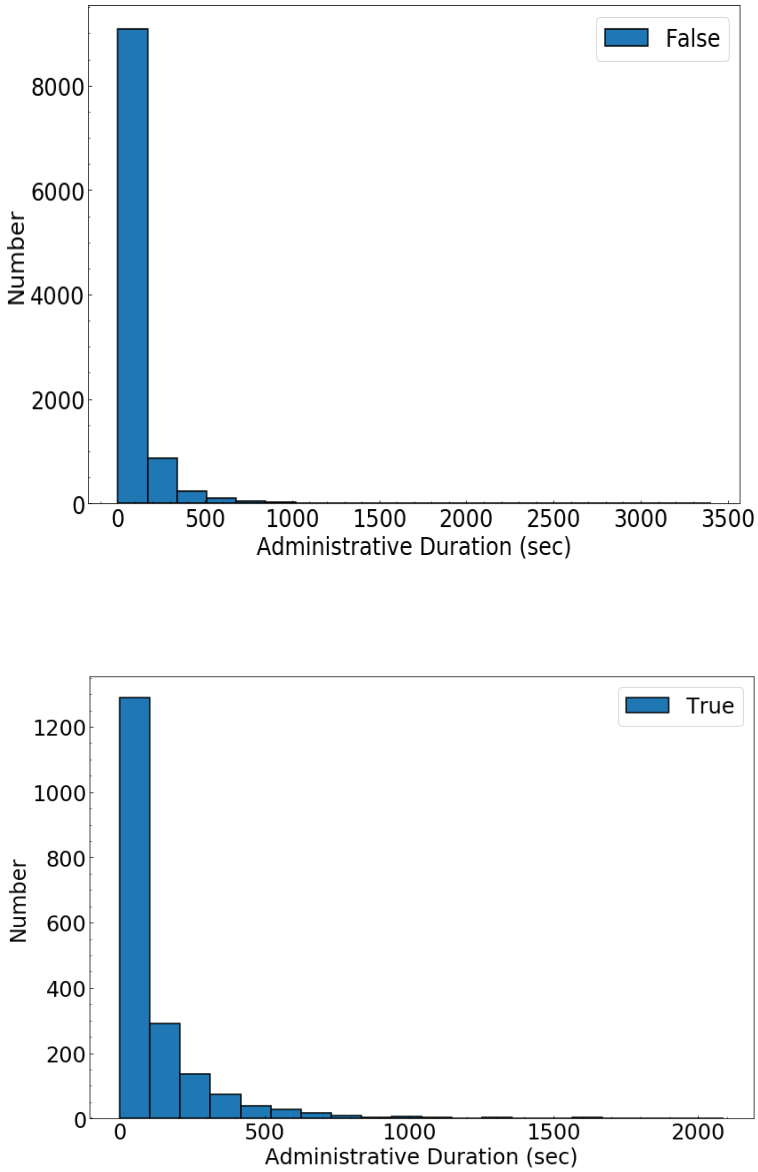


Figure 6: Administrative Page Duration

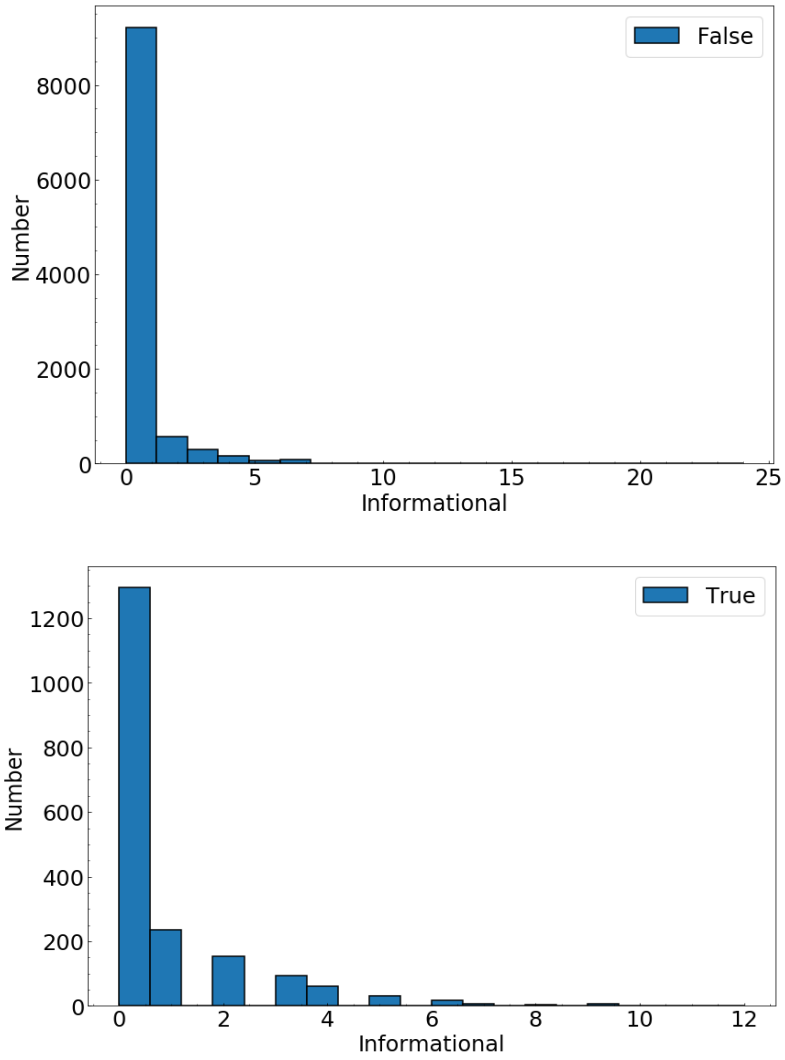
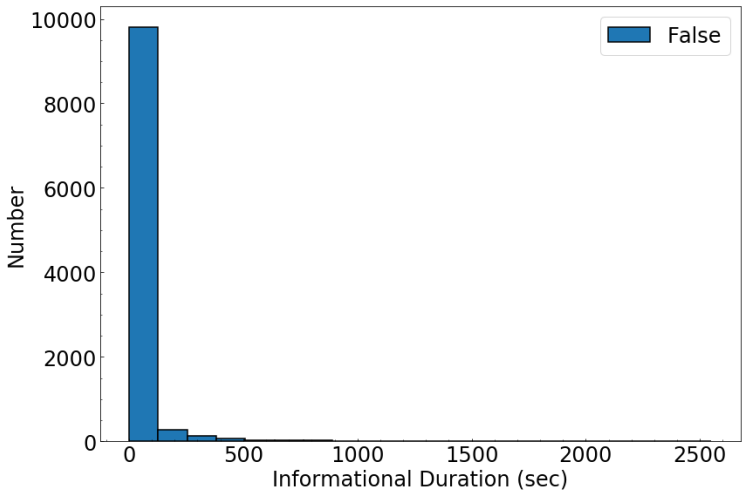


Figure 7: Number of Informational Pages Visited



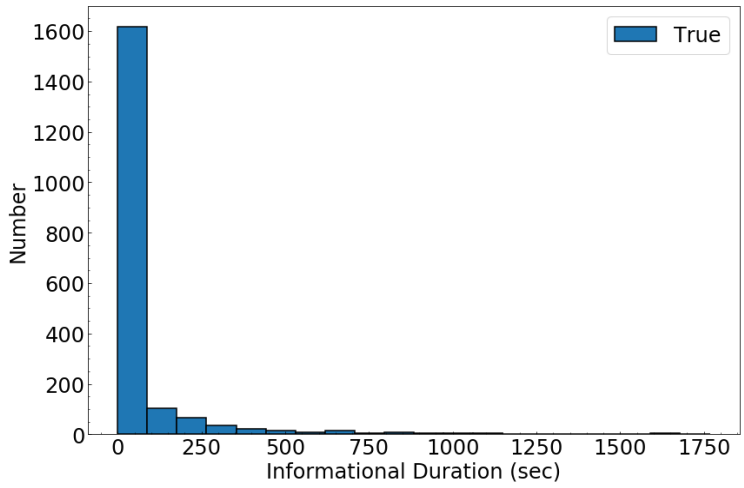


Figure 8: Informational Page Duration

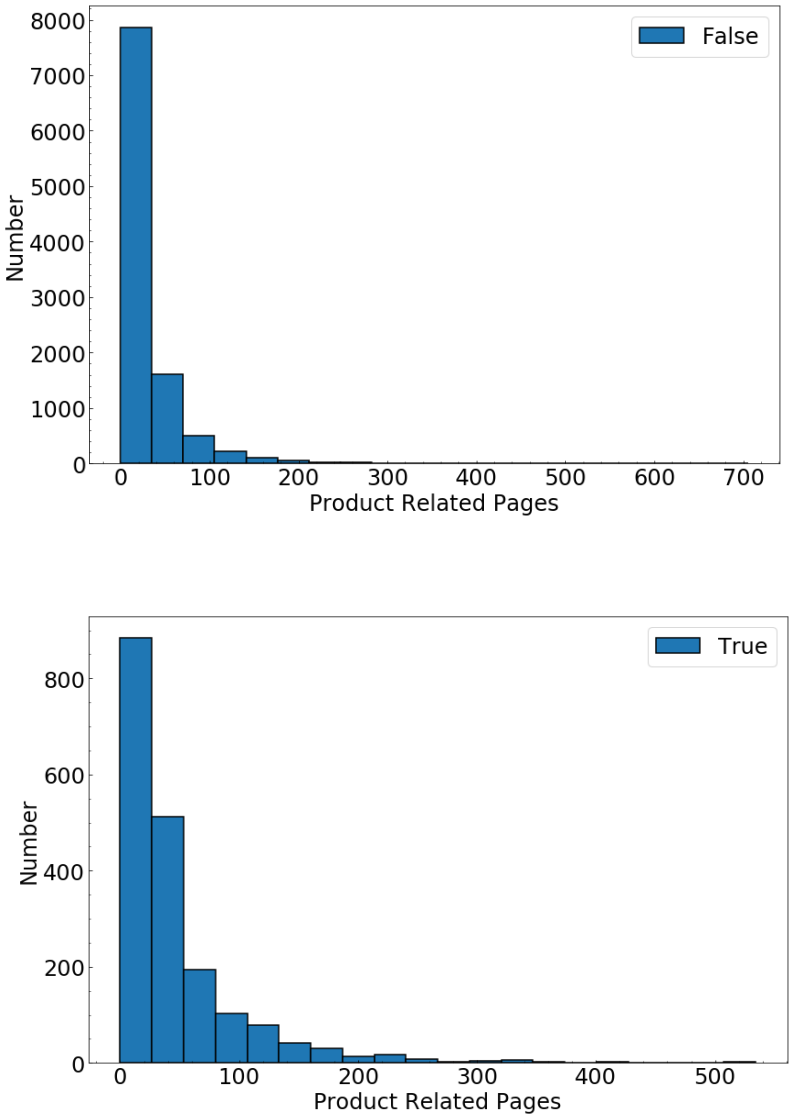


Figure 9: Number of Product-Related Pages Visited

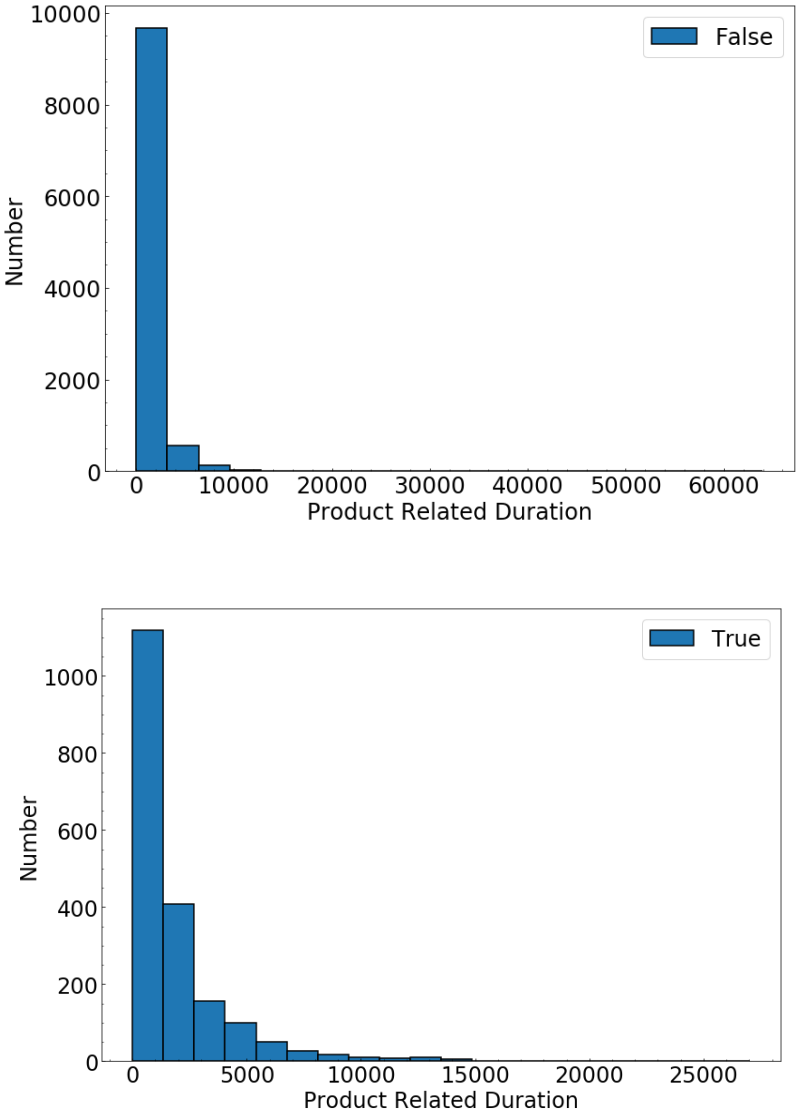


Figure 10: Product Related Page Duration

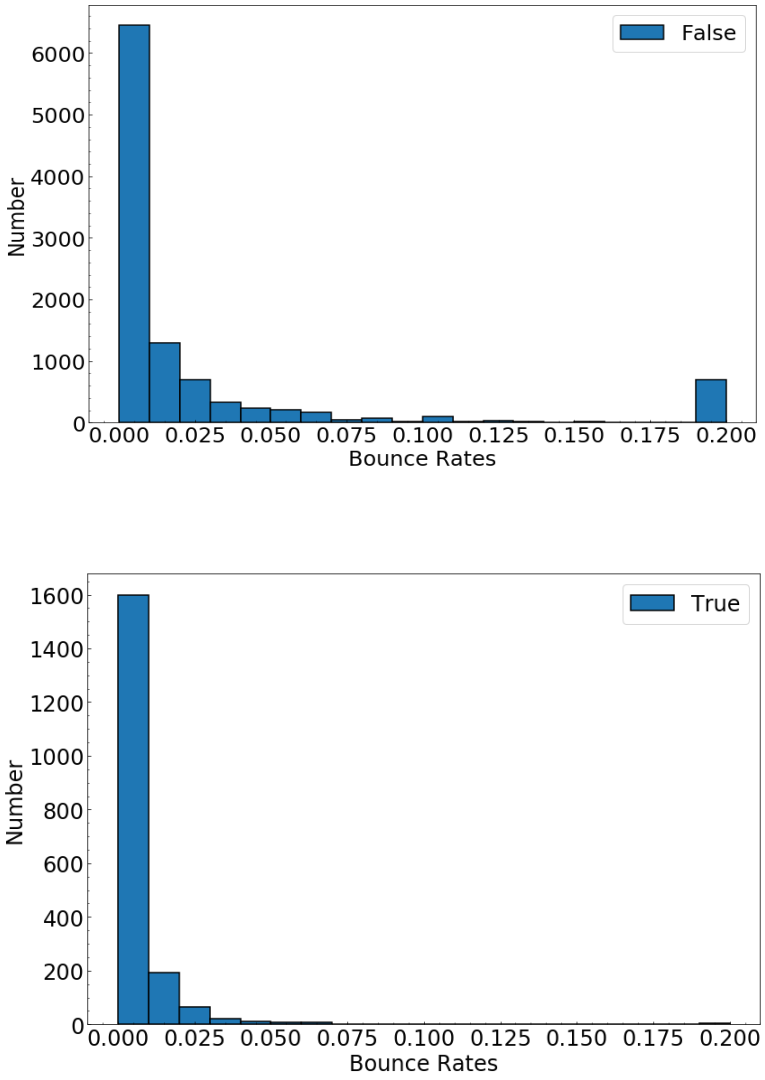


Figure 11: Bounce Rates for Purchasing Intentions

4. MODELING

In this chapter there are 6 different types of models tested on the preprocessed dataset. Each model is examined in order of their accuracy and recall rates. Their confusion matrix and ROC curves are plotted. Also, for each model, feature of importance graph are created.

4.1 Preprocess for Modeling

In order to apply the model to the dataset, we need to transform the data. The first step is to process string data. For the features weekend, visitor type we give values 1, 2, 3 for their categories. And we added a label column. After this process, we split data into train and test datasets. Train dataset size is 0.7 and test dataset size is 0.3.

4.2 Random Forest Model

The first model that is applied is random forest. After splitting data into train and test sets this model was applied. The accuracy rate is 89.86. To check the accuracy rate confusion matrix was created for this model in the figure below.

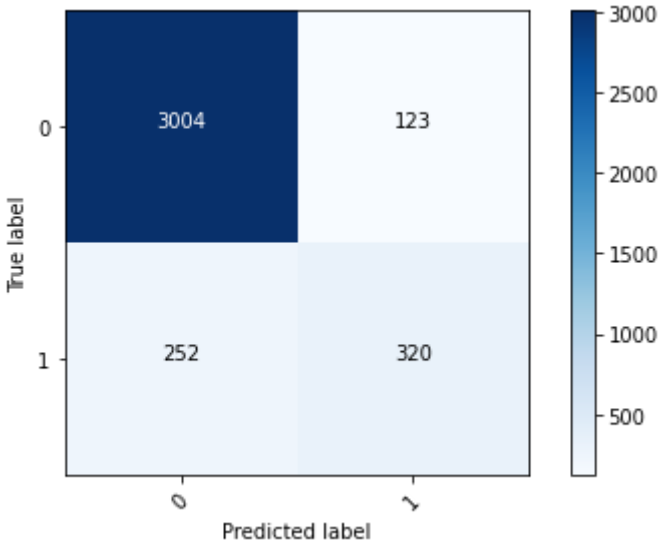


Figure 12: Random Forest Confusion Matrix

As can be seen in the figure the model is very successful for predicting the non-transactions. But for the transactions, it has almost 50% accuracy. This shows that the model's accuracy rate is not a good metric to measure the performance by itself. This can be explained as the result of imbalanced data. In Table 2 it can be seen that the recall rate is 0.56. This shows that even if the model's accuracy rate is high, the model can make wrong predictions.

Table 2: Random Forest Accuracy Rates

	precision	recall	f1-score	support
0	0.92	0.96	0.94	3127
1	0.72	0.56	0.63	572
accuracy			0.90	3699
macro avg	0.82	0.76	0.79	3699
weighted avg	0.89	0.90	0.89	3699

If we check the ROC curve the AUC is 0.91. Again a value parallels the accuracy rate. But this metric can be also misleading like the accuracy rate (Figure 13).

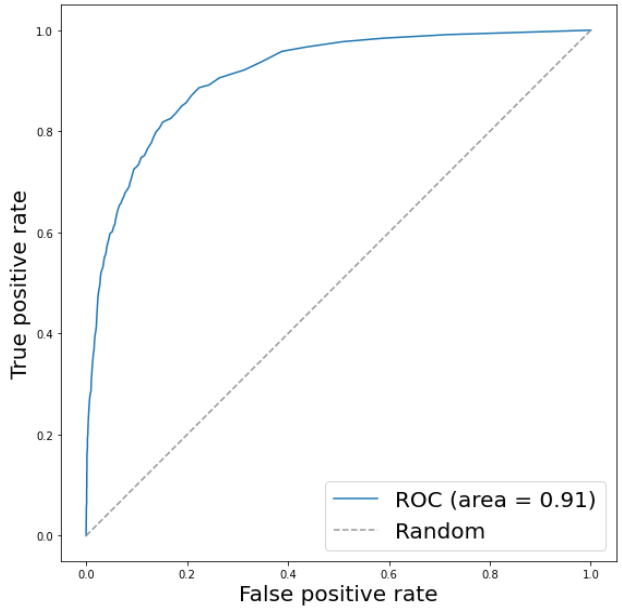


Figure 13: ROC Curve for Random Forest Model

When we check the feature importance for this model Page Value is the most important feature with 0.39 (Figure 14).

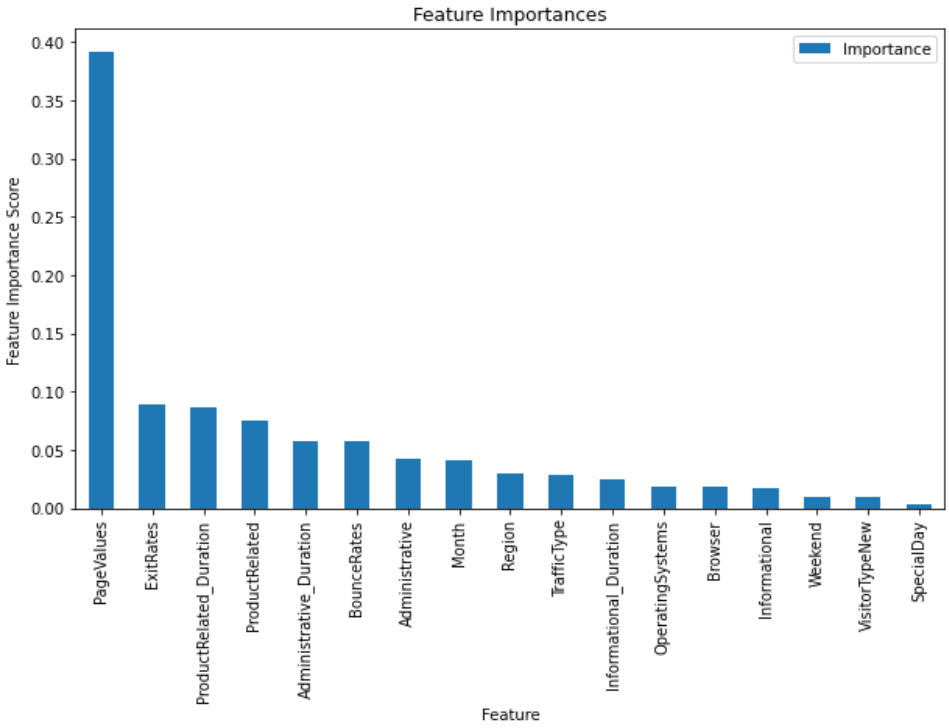


Figure 14: Feature Importance for Random Forest

4.3 Gradient Boosting Model

The second model tested is the gradient boosting model. This model is the longest time processed model in terms of performance. The accuracy rate is 89.84. In order to check the accuracy rate confusion matrix was created for this model in the figure below.

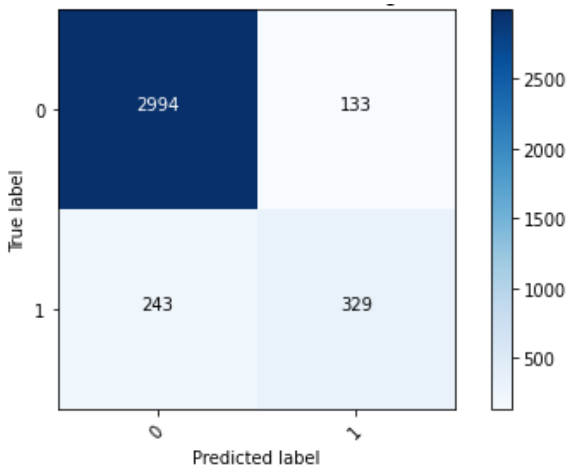


Figure 15: Confusion Matrix for Gradient Boosting

Again as can be seen in the figure the model is very successful for predicting the non-transactions. But for the transactions there is the same problem, the model is not performing well.

In the table below it can be seen that the recall rate is 0.58 which is slightly better then the Random Forest model but not enough for the prediction.

Table 3: Gradient Boosting Accuracy Rates

	precision	recall	f1-score	support
0	0.92	0.96	0.94	3127
1	0.71	0.58	0.64	572
accuracy			0.90	3699
macro avg	0.82	0.77	0.79	3699
weighted avg	0.89	0.90	0.89	3699

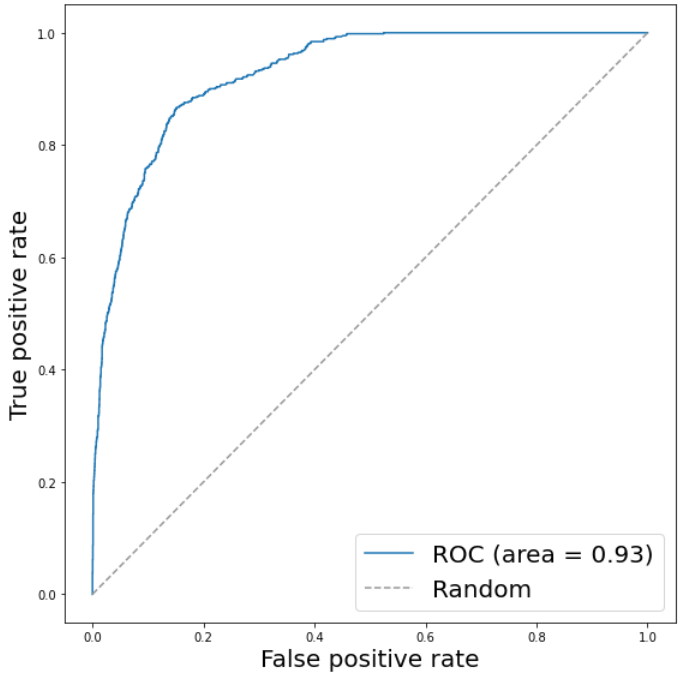


Figure 16: ROC Curve for Gradient Boosting

If we check the ROC curve the AUC is 0.93. Which is again slightly better than the Random Forest model (Figure 16).

And for this model, the most important feature is Page Value like the Random Forest model (Figure 17).

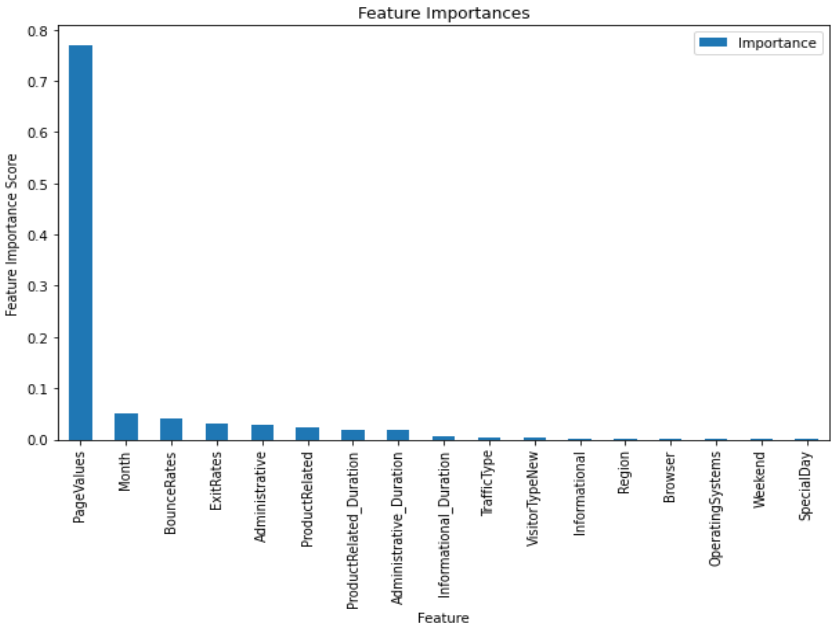


Figure 17: Feature Importance for Gradient Boosting

4.4 XGBoost Model

The third model tested is the XGBoost model. This model is much faster than the Random Forest and Gradient Boosting model in terms of performance. The accuracy rate is 89.97. To check the accuracy rate confusion matrix was created for this model in the figure below.

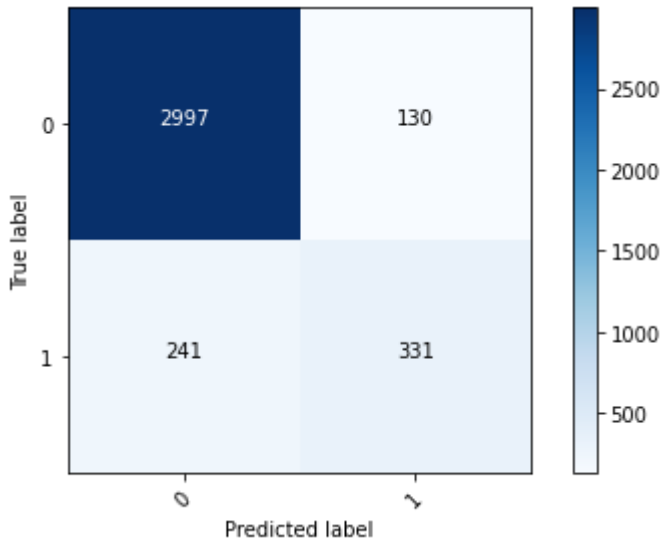


Figure 18: Confusion Matrix for XGBoost

Again as can be seen in the figure the model is not good at predicting users with transactions. Although its prediction rate is higher than Random Forest Gradient Boosting.

In the table below it can be seen that the recall rate is again 0.58.

Table 4: XGBoost Accuracy Rates

	precision	recall	f1-score	support
0	0.93	0.96	0.94	3127
1	0.72	0.58	0.64	572
accuracy			0.90	3699
macro avg	0.82	0.77	0.79	3699
weighted avg	0.89	0.90	0.90	3699

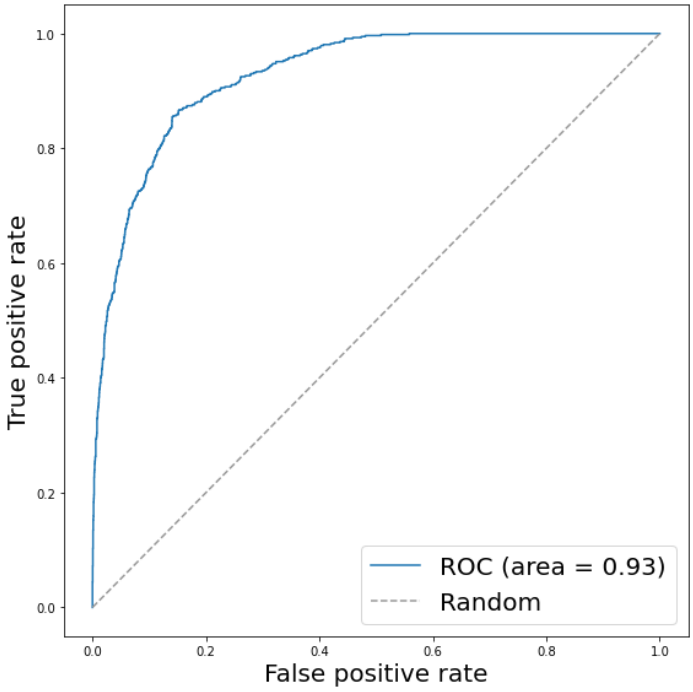


Figure 19: ROC Curve for XGBoost

If we check the ROC curve the AUC is 0.93. Which is the same with Gradient Boosting (Figure 19). And for this model, the most important feature is Page Value like the other two models (Figure 20).

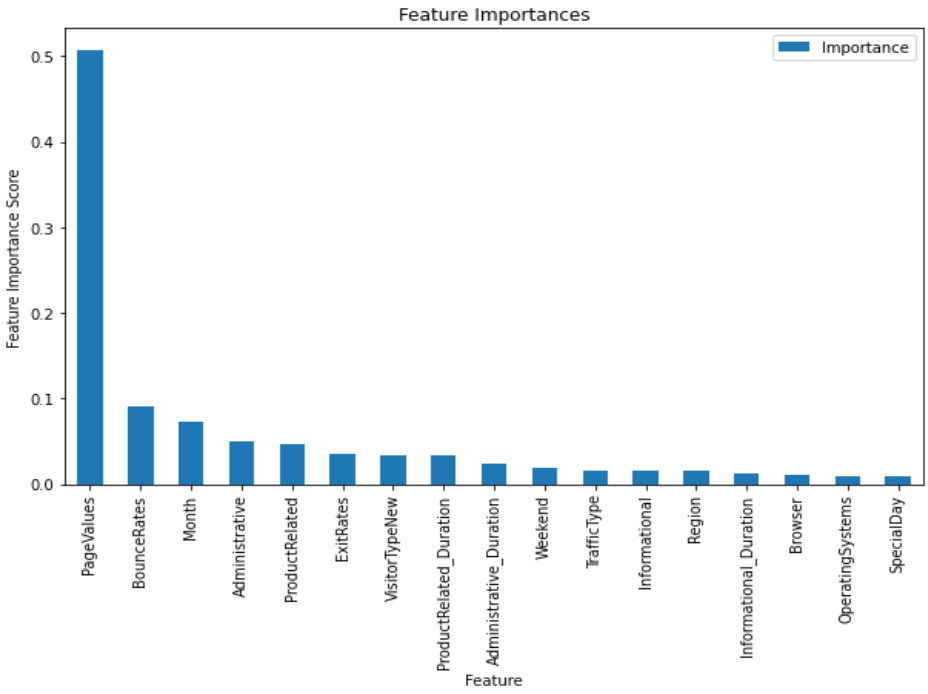


Figure 20: Feature Importance for XGBoost

4.5 LightGBM Model

The fourth model tested is the LightGBM model. The LightGBM model has the highest model performance among all the models. The accuracy rate is 89.86. To check the accuracy rate confusion matrix was created for this model in the figure below.

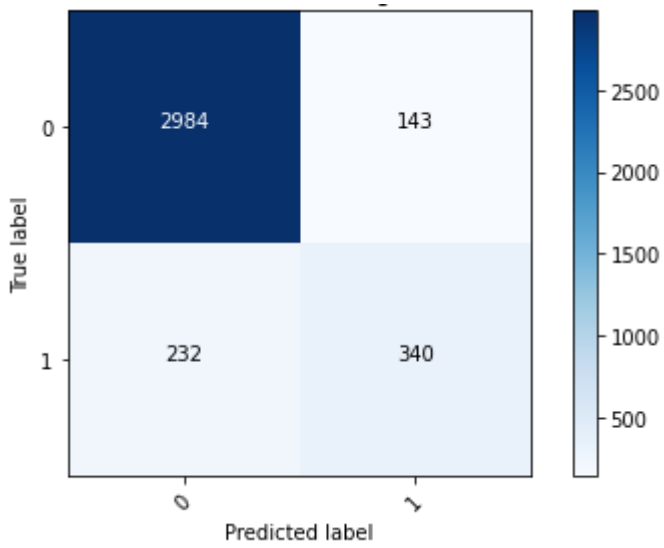


Figure 21: Confusion Matrix for LightGBM

As can be seen in the figure the model is not good at predicting users with transactions. But the true positive rate is higher than the XGBoost model.

In the table below it can be seen that the recall rate is again 0.59. Which is the highest rate among all four models.

Table 5: LightGBM Accuracy Rates

	precision	recall	f1-score	support
0	0.93	0.95	0.94	3127
1	0.70	0.59	0.64	572
accuracy			0.90	3699
macro avg	0.82	0.77	0.79	3699
weighted avg	0.89	0.90	0.90	3699

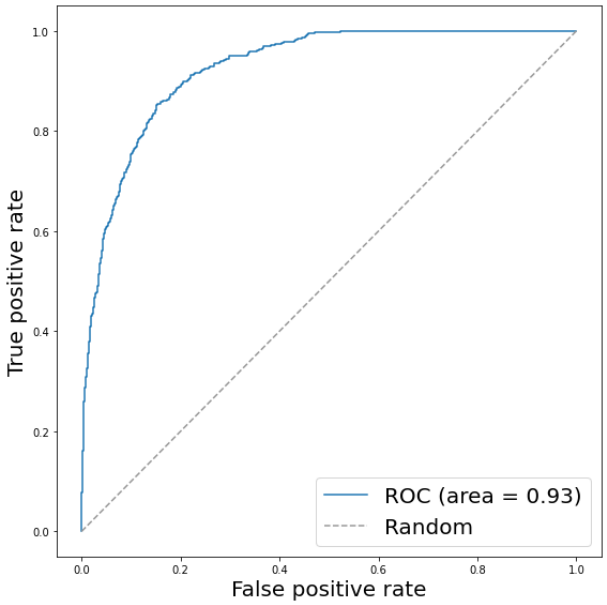


Figure 22: ROC Curve for LightGBM

If we check the ROC curve the AUC is 0.93. Which is the same with Gradient Boosting and XGBoost models (Figure 22).

And for this model the most important feature is Exit Rate which is different from the other three models (Figure 23).

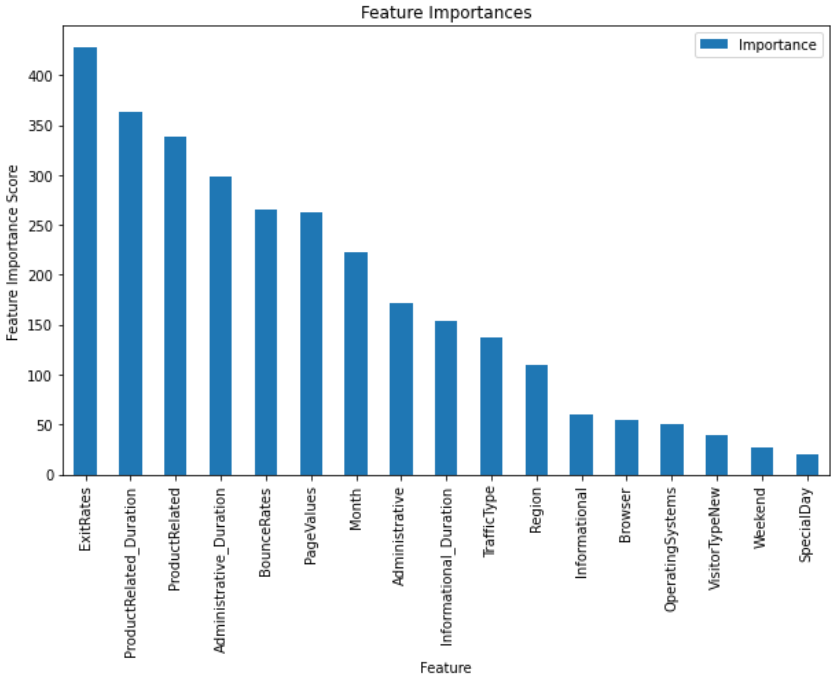


Figure 23: Feature Importance for LightGBM

To compare the models overall the below figure can be examined (Figure 24). As we can see from the graph the highest performance is the LightGBM model. XGboost has the highest accuracy rate. Gradient Boosting model has the lowest accuracy and model performance.

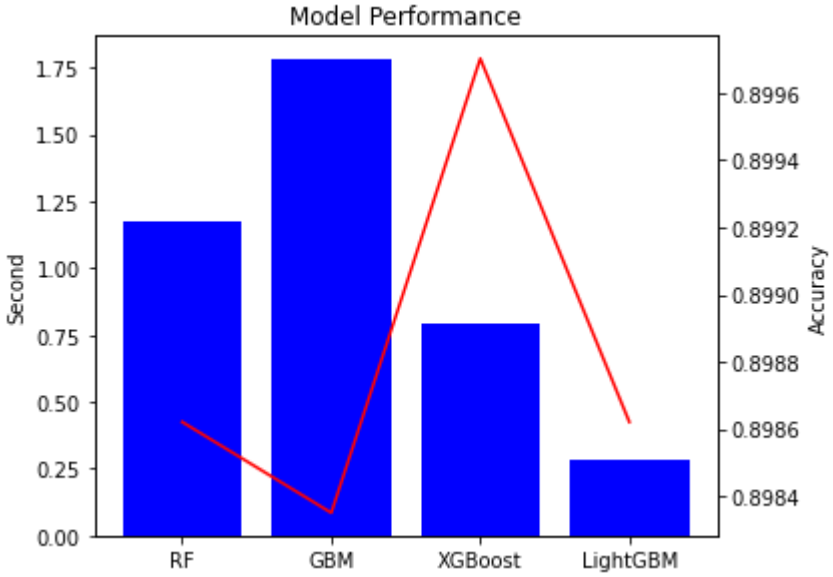


Figure 24: Model Performance vs Accuracy Rates.

4.6 Logistic Regression Model

The fifth model tested is the Logistic Regression model. The accuracy rate is 87.94. Which is lower than the other models. To check the accuracy rate confusion matrix was created for this model in the figure below.

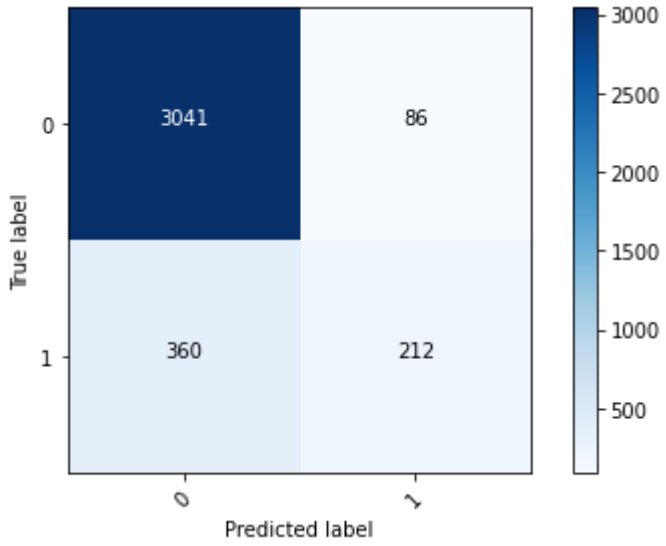


Figure 25: Confusion Matrix for Logistic Regression

As can be seen in the figure the model is performing lower than the others for predicting the users with transactions.

In the table below it can be seen that the recall rate is again 0.37. Which is a really low rate among other models.

Table 6: Logistic Regression Accuracy Rates

	precision	recall	f1-score	support
0	0.89	0.97	0.93	3127
1	0.71	0.37	0.49	572
accuracy			0.88	3699
macro avg	0.80	0.67	0.71	3699
weighted avg	0.87	0.88	0.86	3699

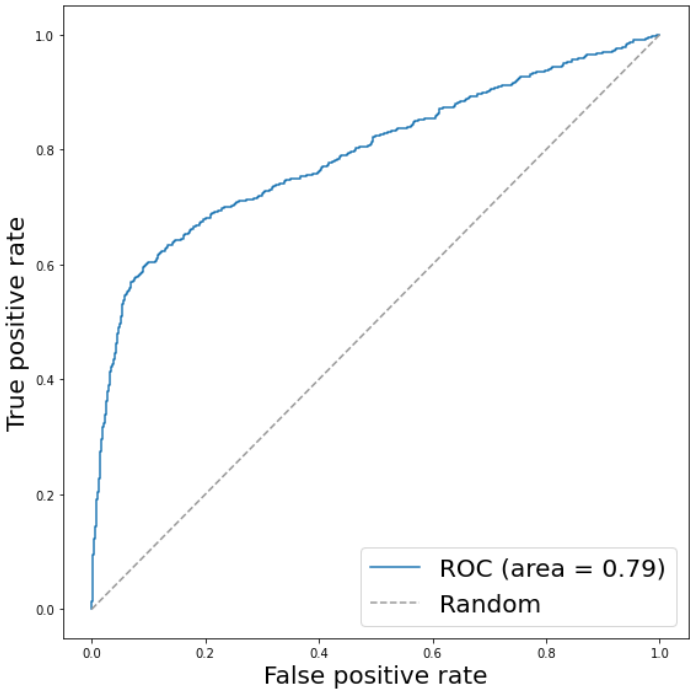


Figure 26: ROC Curve for Logistic Regression

If we check the ROC curve the AUC is 0.79. Which is the lowest rate among all other models and we can see that curve shape is corrupted. (Figure 26).

4.7 Support Vector Machine Model

The final model tested is the Support Vector Machine model. The accuracy rate is 84.64. Which is the lowest rate in all models. To check the accuracy rate confusion matrix was created for this model in the figure below.

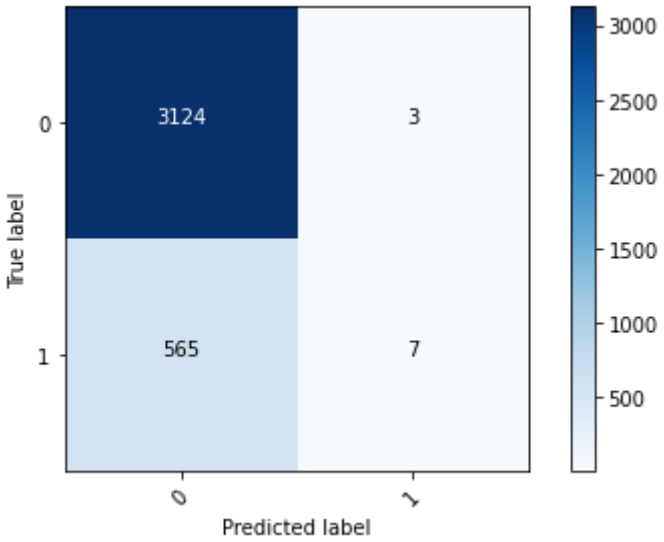


Figure 27: Confusion Matrix for Support Vector Machine

As can be seen in the figure the model is the worst performer for predicting the users with transactions.

In the table below it can be seen that the recall rate is again 0.01. Which is the lowest rate among all models.

Table 7: Support Vector Machine Accuracy Rates

	precision	recall	f1-score	support
0	0.85	1.00	0.92	3127
1	0.70	0.01	0.02	572
accuracy			0.85	3699
macro avg	0.77	0.51	0.47	3699
weighted avg	0.82	0.85	0.78	3699

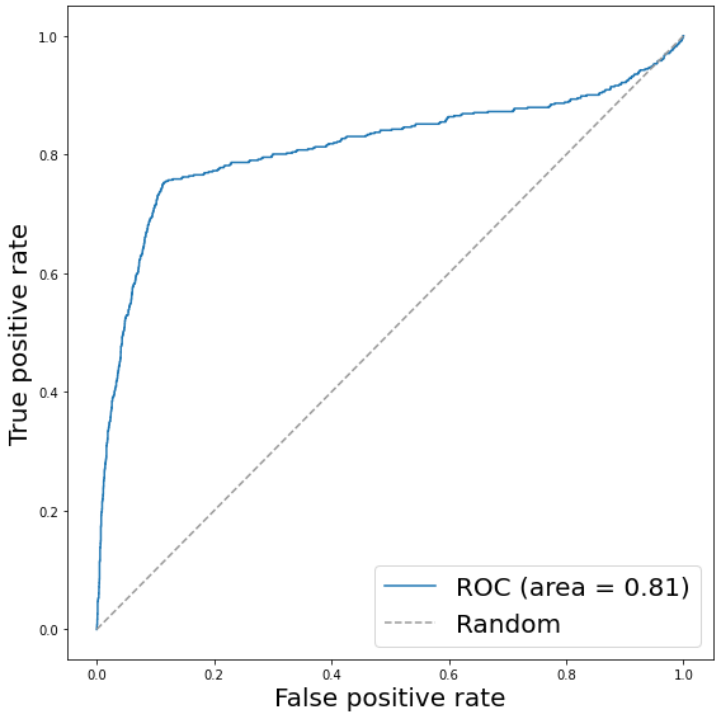


Figure 28: ROC Curve for Support Vector Machine

If we check the ROC curve the AUC is 0.81. And we can see that the curve shape is corrupted. (Figure 28).

After this part weighting method added to the models. For Gradient Boosting and XGBoost models sample weight was used. This process was used in the model fit part. When calculating weights “balanced” method was preferred. For Random Forest, LightGBM, Logistic Regression and Support Vector Machine models it is not necessary to use “sample weight” because it can be used as in the models class_weight = ”balanced”. Although the accuracy rate of some models have decreased as a result of the weight method, it is seen that the determination of the second class in these models is better than before. These are Gradient Boosting and XGBoost models. The models with the highest accuracy are below (Table 8). We can see that even though the accuracy rates of the GBM and XGBoost models are lower than the other models when we check the confusion matrices the determination of the second class is better in these models.

Table 8: Weighted Accuracy Rates

Model	Score
RF	89.889159
lightGBM	87.591241
GBM	86.726142
XGBoost	86.347662
LR	85.698838
SVC	67.937280

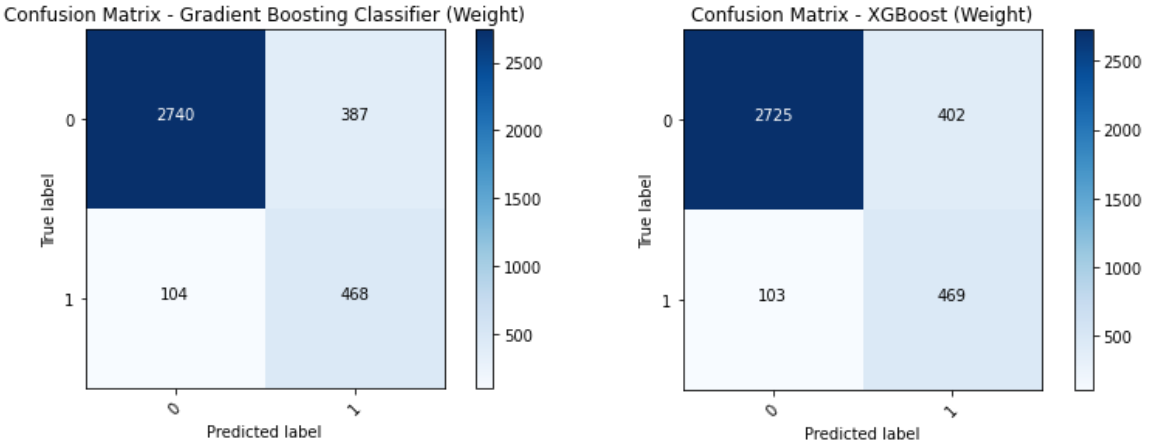


Figure 29: Best Performed Weighted Confusion Matrices

After weight method model tuning applied to the models. In the below table, the parameters used, best parameters and best accuracies can be seen. The best performing model is XGBoost with accuracy rate of 0.9062.

Table 9: Weighted Model Tuning Parameters

Weighted Model	**Parameters**	**Best Parameters**	**Best Accuracy**
Random Forest	params = { 'max_depth': [3, 6, 10, 20, None], 'max_features': ['auto', 'sqrt'], 'n_estimators': [100, 500, 1000]}	{'max_depth': None, 'max_features': 'auto', 'n_estimators': 500}	0,9020
LightGBM	params = { 'max_depth': [-1, 1, 5, 10], 'num_leaves': [20, 30, 40]}	{'max_depth': -1, 'num_leaves': 40}	0,8900
XGBoost	params = { 'max_depth': [3, 6, 10, 20, None], 'learning_rate': [0.01, 0.1, 0.2], 'n_estimators': [100, 500, 1000]}	{'learning_rate': 0.1, 'max_depth': 3, 'n_estimators': 100}	0,9062
Gradient Boosting	params = { 'max_depth': [3, 6, 10], 'learning_rate': [0.01, 0.1, 0.2], 'n_estimators': [10, 100, 500, 1000]}	{'learning_rate': 0.01, 'max_depth': 3, 'n_estimators': 500}	0,9049
Logistic Regression	params = {'penalty': ['l1','l2'], 'C': [0.01,0.1,1,10,100]}	{'C': 100, 'penalty': 'l2'}	0,8712
Support Vector Machine	params = {'C': [0.1,1, 10, 100], 'kernel': ['rbf', 'poly', 'sigmoid']}	{'C': 100, 'kernel': 'poly'}	0,8512

4.8 Oversampling

For this project, 6 different models were applied to the dataset. The common points of all models are that since the data is unbalanced all the models fail to predict users with transactions even though their accuracy rate is quite reasonable. To prevent this situation and increase the model's performance the data should be transformed into balanced data. To do that an oversampling method Smothe can be applied. This method will generate random data and increase the transaction so that the data will become balanced.
Another method is called Random Over Sampler. Both methods were applied for Random Forest and XGBoost models. For the other models, the Smote model was applied.

4.9 Oversampled Random Forest Model

After the Smote method is applied, the accuracy rate for the Random Forest Model is 93.78 which is higher than the normal dataset rate. The recall rate is normalized and becomes 0.94. But for the random over sampler method, the results are better. The accuracy rate is 95.70 and the recall rate is 0.99. In the random over sampler method, the model has higher accuracy in predicting users with transactions. Below there are the confusion matrices.

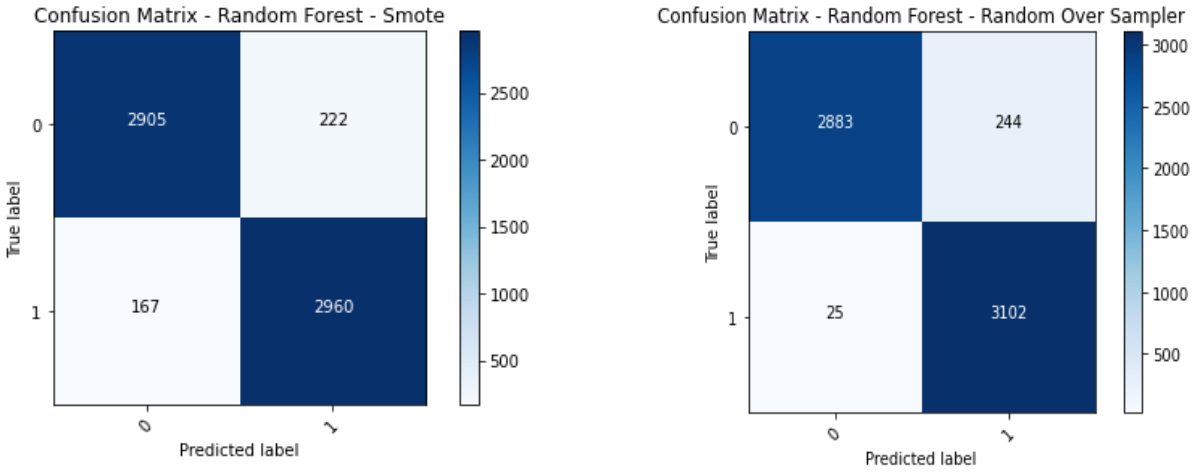


Figure 30: Confusion Matrices for Oversampling

Other metrics can be seen in the table below.

Table 10: Accuracy Rates for Smote Oversampled Random Forest

	precision	recall	f1-score	support
0	0.94	0.92	0.93	3127
1	0.92	0.94	0.93	3127
accuracy			0.93	6254
macro avg	0.93	0.93	0.93	6254
weighted avg	0.93	0.93	0.93	6254

Table 11: Accuracy Rates for Random Oversampled Random Forest

	precision	recall	f1-score	support
0	0.99	0.92	0.96	3127
1	0.93	0.99	0.96	3127
accuracy			0.96	6254
macro avg	0.96	0.96	0.96	6254
weighted avg	0.96	0.96	0.96	6254

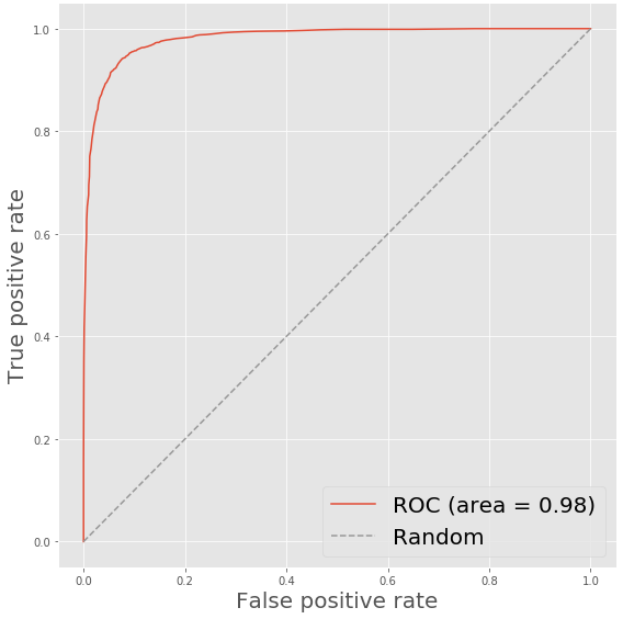


Figure 31: ROC Curve for Oversampled Random Forest

As can be seen, from Figure 30 the AUC is 0.98 which is close to 1. This shows the oversampling method is increasing the accuracy of the classification.

Below the important features can be seen. But as before for this model, the page value feature is the most important attribute.

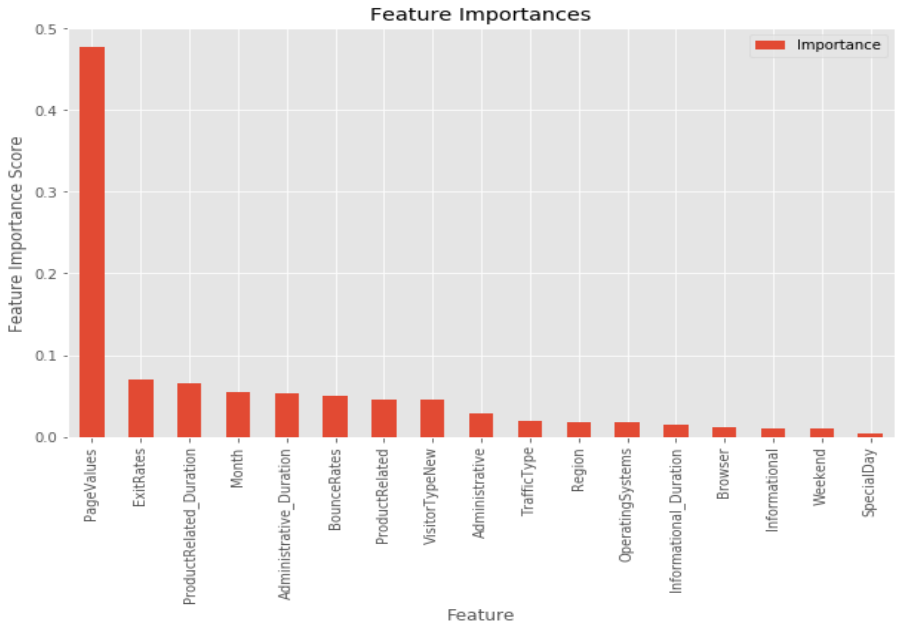


Figure 32: Feature Importance for Oversampled Random Forest

4.10 Oversampled Gradient Boosting Model

After the Smote method is applied, the accuracy rate for the Gradient Boosting Model is 87.13 which was 89.84 before. Although the accuracy rate decreased the recall rate was normalized and became 0.92. Below there is the confusion matrix.

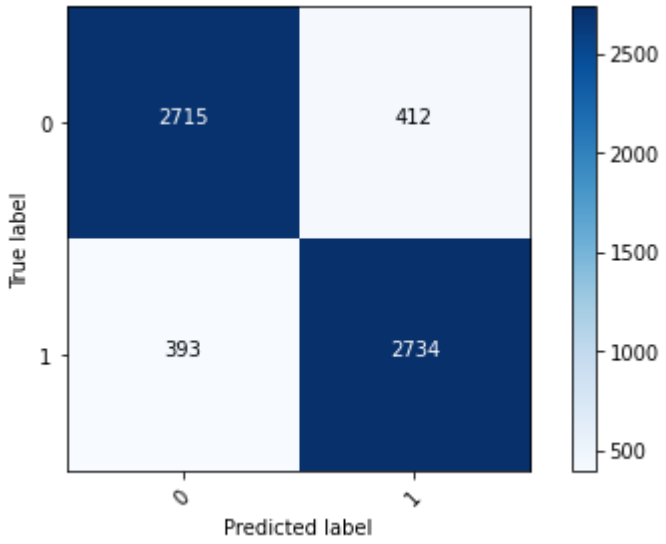


Figure 33: Confusion Matrix for Oversampled Gradient Boosting Model

From the figure, it can be seen that the classification for non-purchaser and purchaser users is quite balanced.

And in the table below there are the recall rates that are normalized.

Table 12: Accuracy Rate for Oversampled Gradient Boosting Model

	precision	recall	f1-score	support
0	0.92	0.91	0.91	3127
1	0.91	0.92	0.92	3127
accuracy			0.92	6254
macro avg	0.92	0.92	0.92	6254
weighted avg	0.92	0.92	0.92	6254

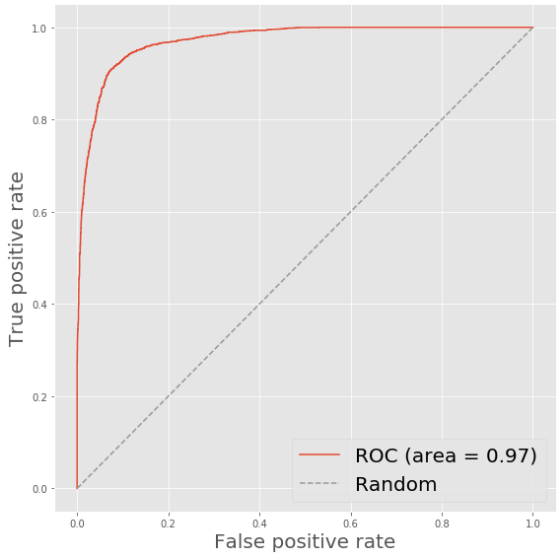


Figure 34: ROC Curve for Oversampled Gradient Boosting Model

The AUC metric is increased to 0.97 from 0.93. And for this model again the page value feature is the most important attribute.

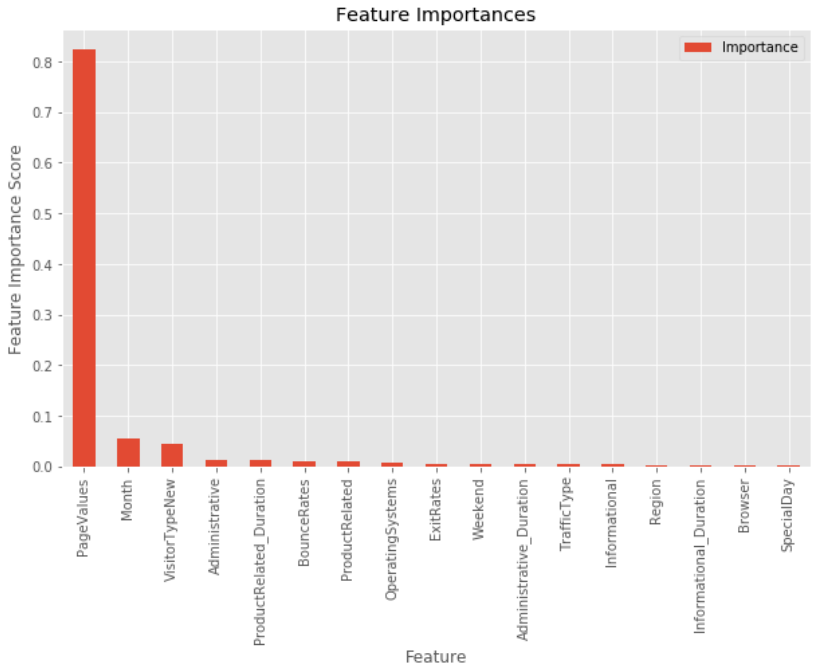


Figure 35: Feature Importance for Oversampled Gradient Boosting Model

4.11 Oversampled XGBoost Model

After the Smote method is applied, the accuracy rate for the XGBoost Model is 86.79 which was 89.97 before. Although the accuracy rate decreased the recall rate was normalized and became 0.87. Below there is the confusion matrix.

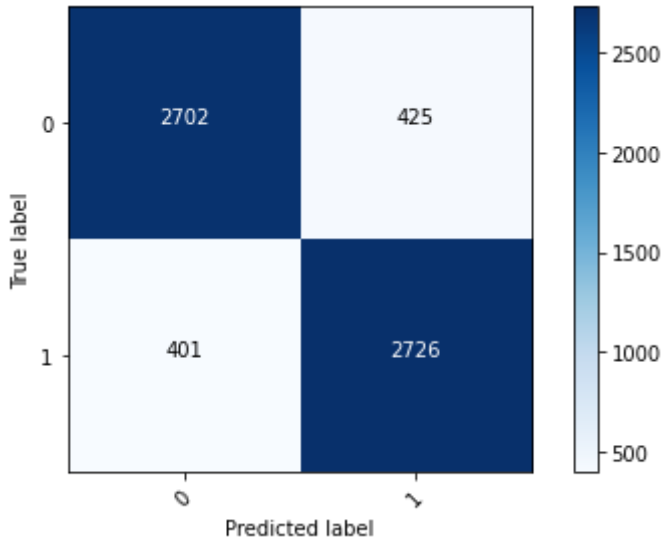


Figure 36: Confusion Matrix for Oversampled XGBoost Model

From the figure, it can be seen that the classification for non-purchaser and purchaser users is quite balanced.

And in the table below there are the recall rates that are normalized.

Table 13: Accuracy Rate for Oversampled XGBoost Model

	precision	recall	f1-score	support
0	0.87	0.86	0.87	3127
1	0.87	0.87	0.87	3127
accuracy			0.87	6254
macro avg	0.87	0.87	0.87	6254
weighted avg	0.87	0.87	0.87	6254

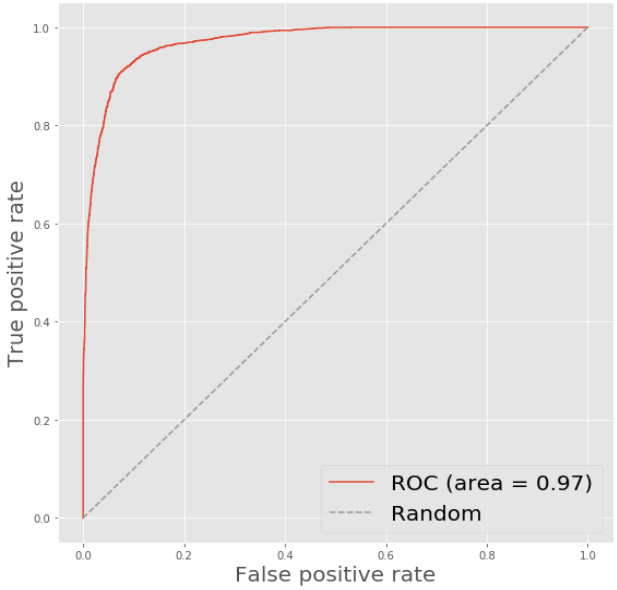


Figure 37: ROC Curve for Oversampled XGBoost Model

The AUC metric is increased to 0.94 from 0.93. And for this model again the page value feature is the most important attribute.

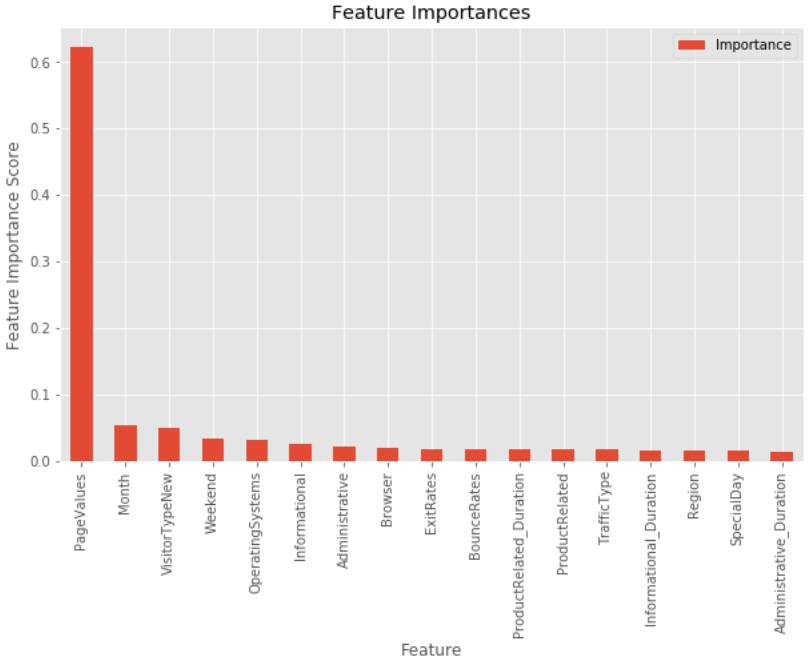


Figure 38: Feature Importance for Oversampled XGBoost Model

4.12 Oversampled LightGBM Model

After the Smote method is applied, the accuracy rate for LightGBM Model is 91.11 which was 89.86 before. The recall rate is normalized and becomes 0.83. Below there is the confusion matrix.

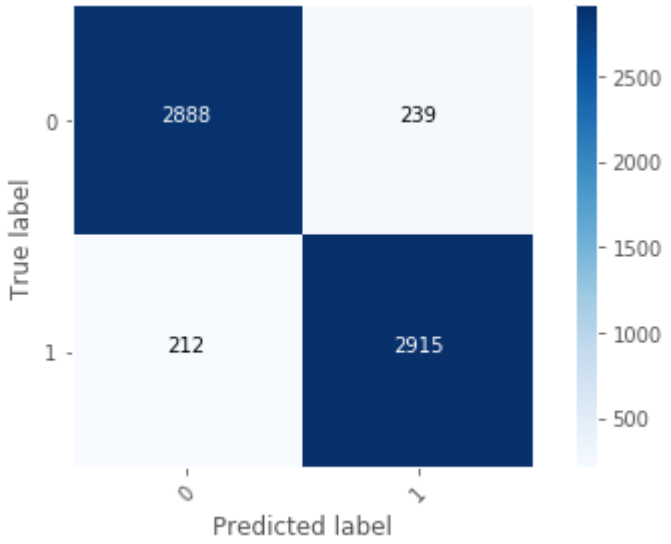


Figure 39: Confusion Matrix for Oversampled LightGBM Model

From the figure, it can be seen that the classification for non-purchaser and purchaser users is quite balanced.

And in the table below there are the recall rates that are normalized.

Table 14: Accuracy Rate for Oversampled LightGBM Model

	precision	recall	f1-score	support
0	0.93	0.92	0.93	3127
1	0.92	0.93	0.93	3127
accuracy			0.93	6254
macro avg	0.93	0.93	0.93	6254
weighted avg	0.93	0.93	0.93	6254

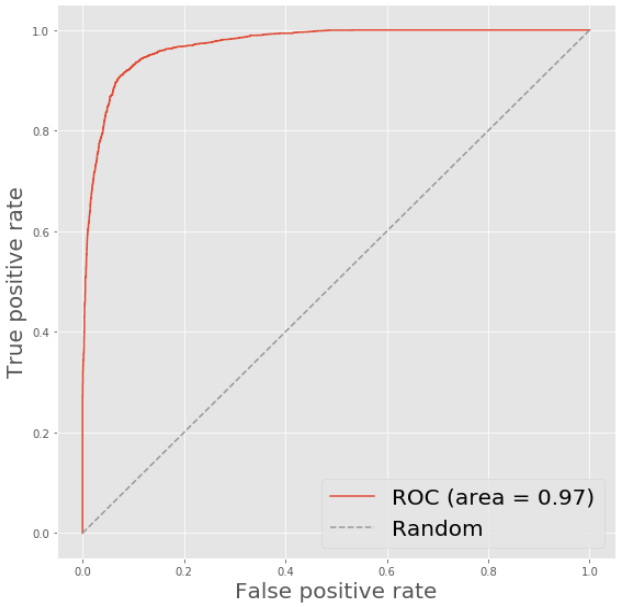


Figure 40: ROC Curve for Oversampled LightGBM Model

The AUC metric is increased to 0.93 from 0.98. And for this model again before the exit rate was the most important feature but after oversampling product-related duration become the most important feature.

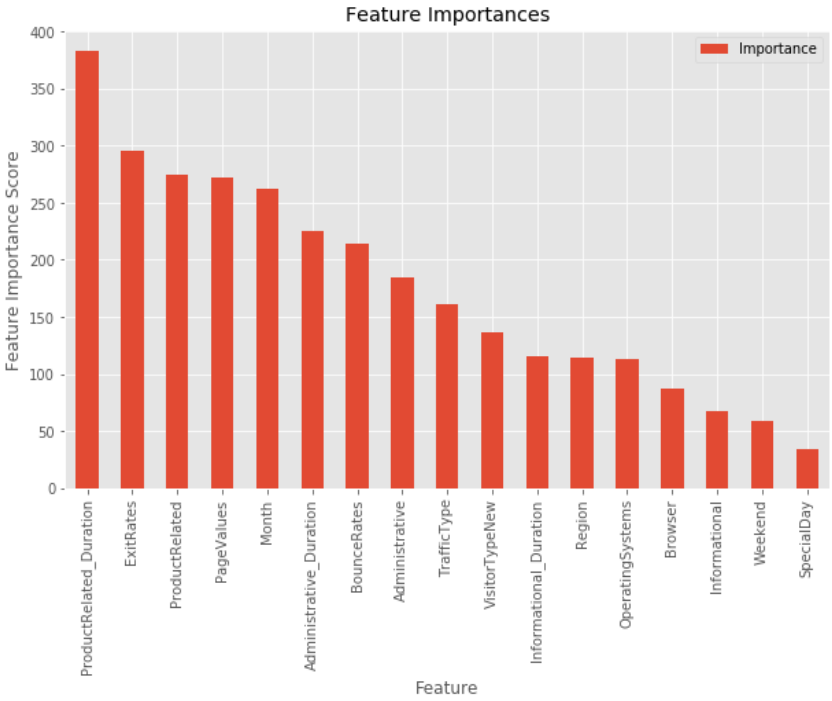


Figure 41: Feature Importance for Oversampled LightGBM Model

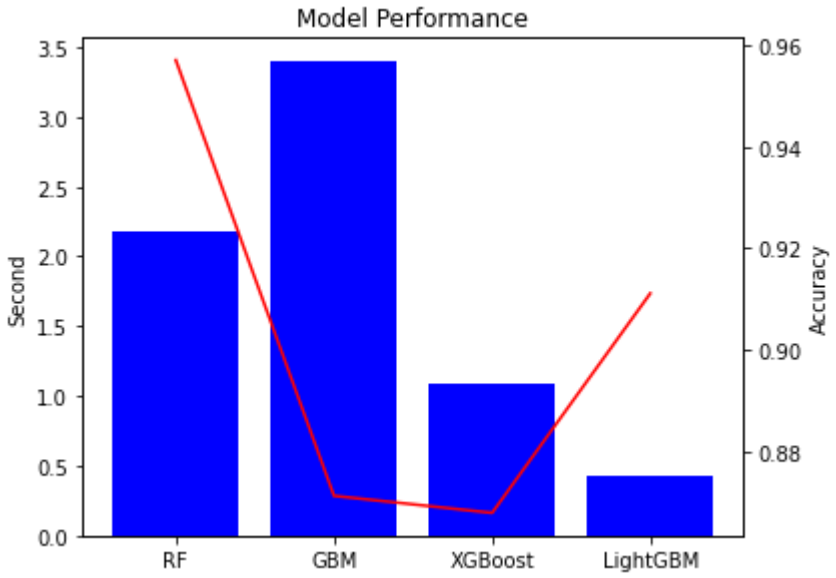


Figure 42: Oversampled Model Performance vs. Accuracy Rates

To compare the models overall the figure above can be examined (Figure 42). As we can see from the graph the highest performance is the LightGBM model again. This time Random Forest has the highest accuracy rate. The XGBoost model has the lowest accuracy and Gradient Boosting has the lowest model performance.

4.13 Oversampled Logistic Regression Model

After the Smote method is applied, the accuracy rate for the Logistic Regression Model is 83.74 which was 87.94 before. The accuracy rate decreased but the recall rate was normalized and became 0.78 which was 0.37 before. Below there is the confusion matrix.

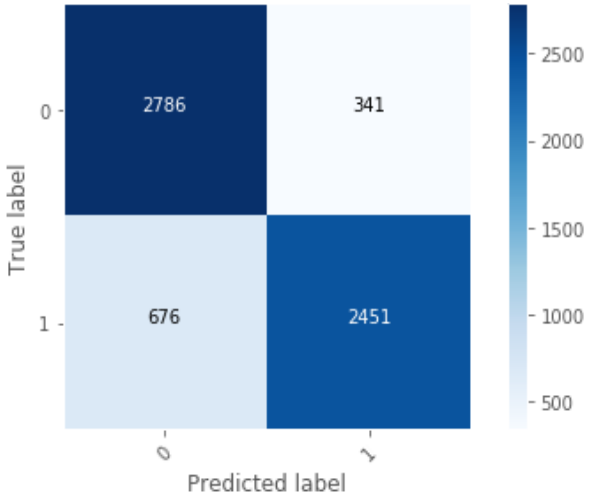


Figure 43: Confusion Matrix for Oversampled Logistic Regression

From the figure, it can be seen that the classification for non-purchaser and purchaser users is quite balanced.

And in the table below there are the recall rates that are normalized.

Table 15: Oversampled Logistic Regression Accuracy Rates

	precision	recall	f1-score	support
0	0.80	0.89	0.85	3127
1	0.88	0.78	0.83	3127
accuracy			0.84	6254
macro avg	0.84	0.84	0.84	6254
weighted avg	0.84	0.84	0.84	6254

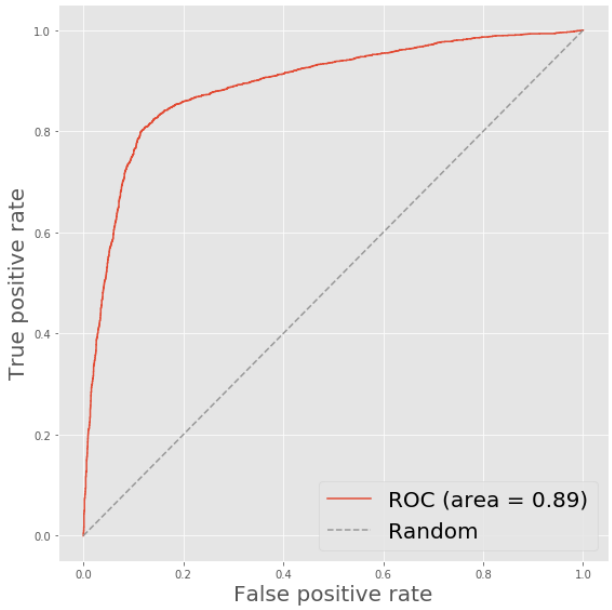


Figure 44: ROC Curve for Oversampled Logistic Regression

If we check the ROC curve the AUC is 0.89. Which was 0.79 before. There is an increase in the classification rate (Figure 44).

4.14 Oversampled Support Vector Machine Model

After the Smote method is applied, the accuracy rate for the Support Vector Machine Model is 72.11 which was 84.64 before. The accuracy rate decreased but the recall rate was normalized and became 0.75 which was 0.01 before. Below there is the confusion matrix.

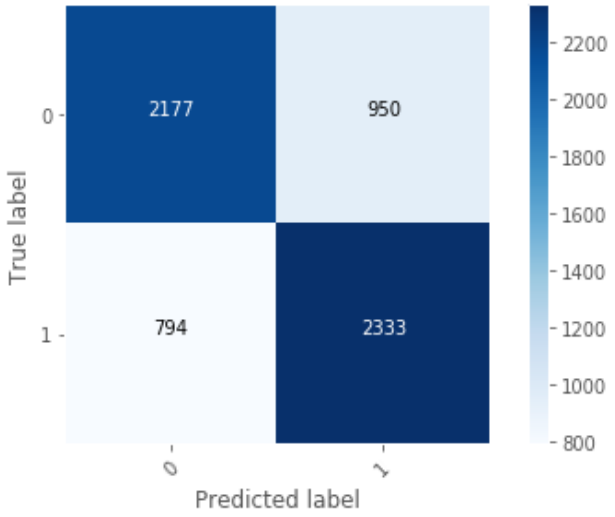


Figure 45: Confusion Matrix for Oversampled Support Vector Machine

From the figure, it can be seen that the classification for non-purchaser and purchaser users is quite balanced.

And in the table below there are the recall rates that are normalized.

Table 16: Oversampled Support Vector Machine Accuracy Rates

	precision	recall	f1-score	support
0	0.73	0.70	0.71	3127
1	0.71	0.75	0.73	3127
accuracy			0.72	6254
macro avg	0.72	0.72	0.72	6254
weighted avg	0.72	0.72	0.72	6254

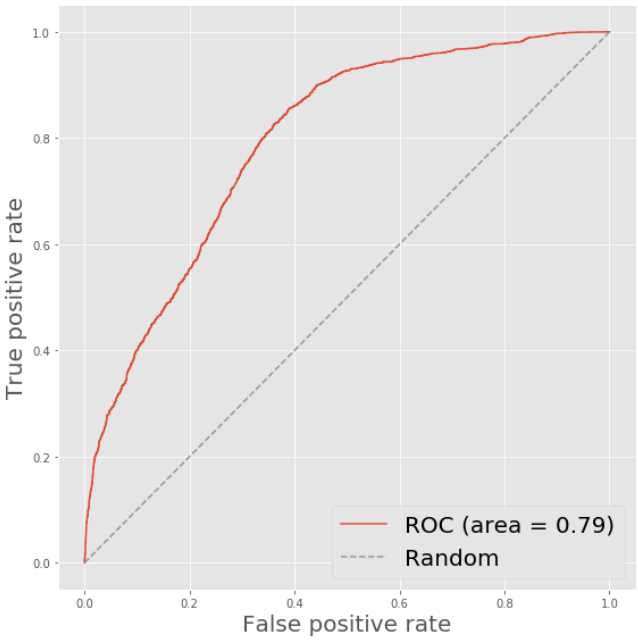


Figure 46: ROC Curve for Oversampled Support Vector Machine

If we check the ROC curve the AUC is 0.79. And we can see that the curve shape is slightly corrupted. (Figure 46).

4.15. Model Tuning

For every model the hyperparameter fine tuning applied. In the below table the best parameters and beast accuracies can be seen. After the model tuning process we can say that again the best model for oversampled dataset is XGBoost model with the rate of 0.9354.

Table 17: Oversampled Model Tuning

Oversampled Model	**Parameters**	**Best Parameters**	**Best Accuracy**
Random Forest	params = { 'max_depth': [3, 6, 10, 20, None], 'max_features': ['auto', 'sqrt'], 'n_estimators': [100, 500, 1000]}	{'max_depth': None, 'max_features': 'auto', 'n_estimators': 1000}	0,9303
LightGBM	params = { 'max_depth': [-1, 1, 5, 10], 'num_leaves': [20, 30, 40]}	{'max_depth': -1, 'num_leaves': 40}	0,9298
XGBoost	params = { 'max_depth': [3, 6, 10, 20, None], 'learning_rate': [0.01, 0.1, 0.2], 'n_estimators': [100, 500, 1000]}	{'learning_rate': 0.1, 'max_depth': 10, 'n_estimators': 500}	0,9354
Gradient Boosting	params = { 'max_depth': [3, 6, 10], 'learning_rate': [0.01, 0.1, 0.2], 'n_estimators': [10, 100, 500, 1000]}	{'learning_rate': 0.2, 'max_depth': 10, 'n_estimators': 1000}	0,9353
Logistic Regression	params = {'penalty': ['l1','l2'], 'C': [0.01,0.1,1,10,100]}	{'C': 0.1, 'penalty': 'l2'}	0,8410
Support Vector Machine	params = {'C': [0.1,1, 10, 100], 'kernel': ['rbf', 'poly', 'sigmoid']}	{'C': 100, 'kernel': 'rbf'}	0,8272

5.CONCLUSION

In conclusion, for the online purchasing prediction the dataset used is imbalanced. This imbalance can cause some problems in the modeling part. There are six different models applied to this dataset.

These are Random Forest, Gradient Boosting, XGBoost, LightGBM, Logistic Regression, and Support Vector Machine. At a glance, most of the models' accuracy rates are normal, in fact, high for the prediction.

But when we check the recall rates there is a problem. The models are performing well for predicting users with no purchase although they are not able to classify users with purchase. They classify them wrong almost 50%. The imbalance of data can cause wrong classifications like this. When checking the model's accuracy, looking at one metric might be misleading. To make sure the other metrics such as recall, precision, etc. should be considered as well.

To prevent the wrong classification situation Smote or Random Over Sampler method can be used to randomly reproduce data and transform the dataset into balanced. For this project Smote method is used. The dataset is reproduced with Smote and all models are applied again. For Random Forest Model also Random Over Sampler method is applied. After model tuning process among all the models we can say that the XGBoost is the best-performed model.

For further studies with a suitable dataset, we can predict online purchasing intentions in real-time. With that, it will be possible to suggest users' real-time campaigns. And also user's reactions can be measured to these real-time campaigns. With real-time insights, e-commerce sites can make users shop without even knowing their purchasing tendencies.

REFERENCES

[1] Online Shoppers Purchasing Intention Dataset Data Set. (n.d.). doi:10.31289/jite.v3i1.2599.s170

[2] Sakar, C. O., Polat, S. O., Katircioglu, M., & Kastro, Y. (2018). Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and LSTM recurrent neural networks. Neural Computing and Applications, 31(10), 6893-6908. doi:10.1007/s00521-018-3523-0

[3] Baati, K., & Mohsil, M. (2020). Real-Time Prediction of Online Shoppers' Purchasing Intention Using Random Forest. IFIP Advances in Information and Communication Technology Artificial Intelligence Applications and Innovations, 43-51. doi:10.1007/978-3-030-49161-1_4

[4] Poel, D. V., & Buckinx, W. (2005). Predicting online-purchasing behaviour. European Journal of Operational Research, 166(2), 557-575. doi:10.1016/j.ejor.2004.04.022

[5] Joshi, R., Gupte, R., & Saravanan, P. (2018). A Random Forest Approach for Predicting Online Buying Behavior of Indian Customers. Theoretical Economics Letters, 08(03), 448-475. doi:10.4236/tel.2018.83032

[6] Beck, M. (2021, June 11). Can You Predict If a Customer Will Make a Purchase on a Website? Retrieved from https://towardsdatascience.com/can-you-predict-if-a-customer-will-make-a-purchase-on-a-website-e6843ec264ae

[7] Xiao, E. (2020, December 18). Browsing or Purchasing: Real-Time Prediction of Online Shopper's Purchasing Intention (I) . Retrieved from https://medium.com/swlh/browsing-or-purchasing-real-time-prediction-of-online-shoppers-purchasing-intention-i-4f13e0447b7c

[8] Deliçay, M. (2021, January 1). T.C. Cumhurbaşkanlığı Strateji ve Bütçe Başkanlığı - SBB. Retrieved from https://www.sbb.gov.tr/wp-content/uploads/2021/01/Perakende_E-Ticaretin_Yukselisi.pdf

[9] Ulukan, G. (2021, March 01). RTB House'un online alışveriş araştırmasına göre yoğun hafta sonu alışverişi bir efsane. Retrieved from https://webrazzi.com/2020/12/21/rtb-house-un-online-alisveris-arastirmasina-gore-yogun-hafta-sonu-alisverisi-bir-efsane/

[10] Published by Statista Research Department, & 5, J. (2021, July 05). U.S. e-commerce market size 2016-2023. Retrieved from https://www.statista.com/statistics/272391/us-retail-e-commerce-sales-forecast/

[11] Dataman, D. (2021, July 11). Using Over-Sampling Techniques for Extremely Imbalanced Data. Retrieved from https://medium.com/dataman-in-ai/sampling-techniques-for-extremely-imbalanced-data-part-ii-over-sampling-d61b43bc4879

[12] A. Amin et al., "Comparing Oversampling Techniques to Handle the Class Imbalance Problem: A Customer Churn Prediction Case Study," in IEEE Access, vol. 4, pp. 7940-7957, 2016, doi: 10.1109/ACCESS.2016.2619719.

[13] Daoud, E. A. (2019, January 15). Comparison between XGBoost, LightGBM and CatBoost Using a Home Credit Dataset. Retrieved from https://publications.waset.org/10009954/comparison-between-xgboost-lightgbm-and-catboost-using-a-home-credit-dataset

[14] Shung, K. P. (2020, April 10). Accuracy, Precision, Recall or F1? Retrieved from https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9

[15] Last, F., Douzas, G., & Bacao, F. (2017, December 12). Oversampling for Imbalanced Learning Based on K-Means and SMOTE. Retrieved from https://arxiv.org/abs/1711.00837v2