Final Project Report

ONLINE SHOPPER'S PURCHASING INTENTION

Jyotsna Eltepu

Introduction

In 2021, nearly $1 in every $5 spent on retail purchases came from digital orders. The insights about user behavior can help in targeting the users who are more likely to make a purchase. When you have sufficient data about a user, you can design a product, service, or campaign around the user. And the insights can be used to increase customer loyalty and retention.

The goal of the project is to predict the purchasing intention of an online shopper so that the business can get insights into user behavior for determining what drives the consumers to purchase a product and use targeted marketing to increase the sales and profits.

Data Wrangling

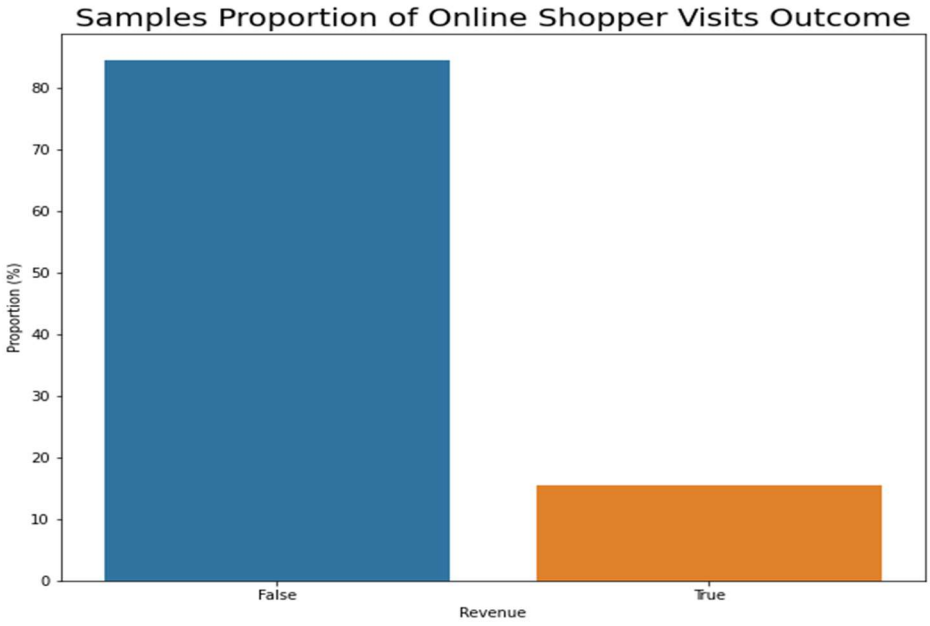
The data used in this analysis is an Online Shoppers Purchasing Intention data set provided on the UC Irvine's Machine Learning Repository. The dataset consists of feature vectors belonging to 12,330 sessions of different users in a 1 year period. The dataset consists of 10 numerical and 8 categorical attributes. The dataset did not contain any null values.

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 12330 entries, 0 to 12329
Data columns (total 18 columns):
 # Column Non-Null Count Dtype
0 Administrative 12330 non-null int64

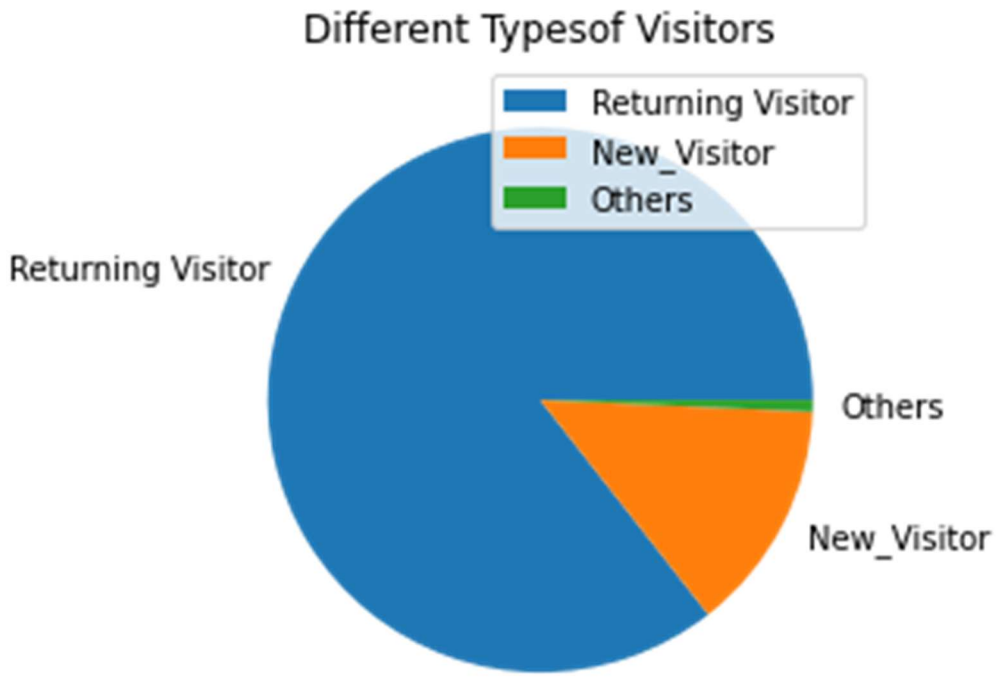
1 Administrative_Duration 12330 non-null float64
2 Informational 12330 non-null int64
3 Informational_Duration 12330 non-null float64
4 ProductRelated 12330 non-null int64
5 ProductRelated_Duration 12330 non-null float64
6 BounceRates 12330 non-null float64
7 ExitRates 12330 non-null float64
8 PageValues 12330 non-null float64
9 SpecialDay 12330 non-null float64
10 Month 12330 non-null object
11 OperatingSystems 12330 non-null int64
12 Browser 12330 non-null int64
13 Region 12330 non-null int64
14 TrafficType 12330 non-null int64
15 VisitorType 12330 non-null object
16 Weekend 12330 non-null bool
17 Revenue 12330 non-null bool
dtypes: bool(2), float64(7), int64(7), object(2)

Exploratory Data Analysis

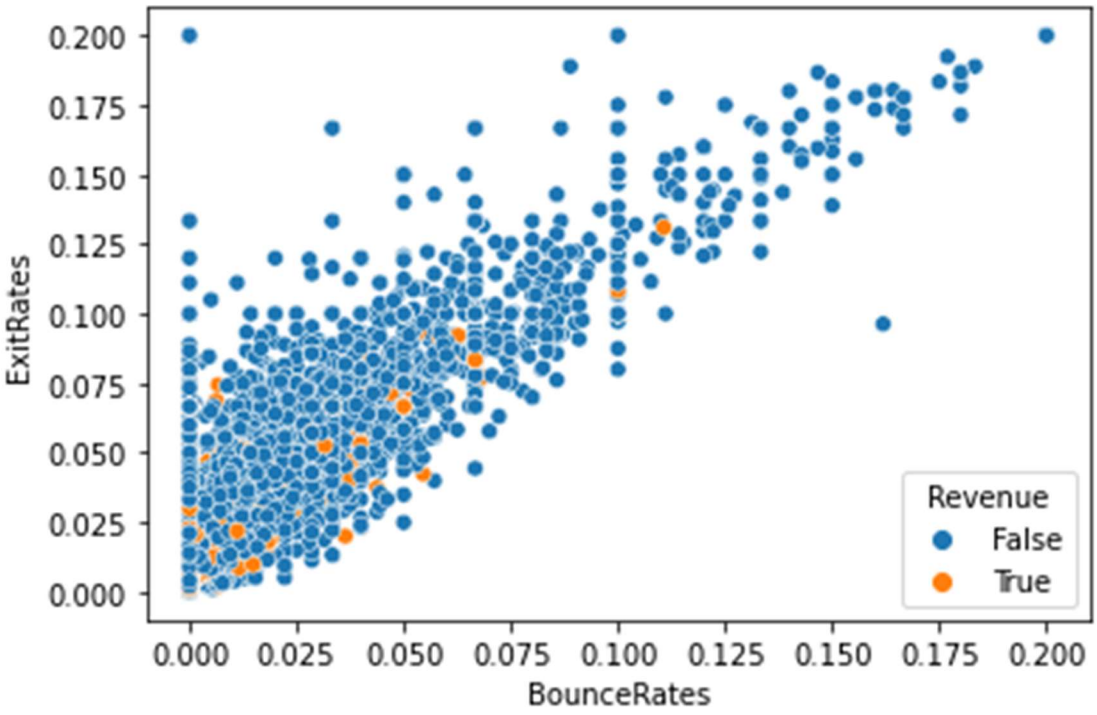
In this dataset, 85% of the users did not end up making any purchase and only 15% users made a final purchase. This is an imbalanced dataset. The following figure shows the sample proportion of the revenue.



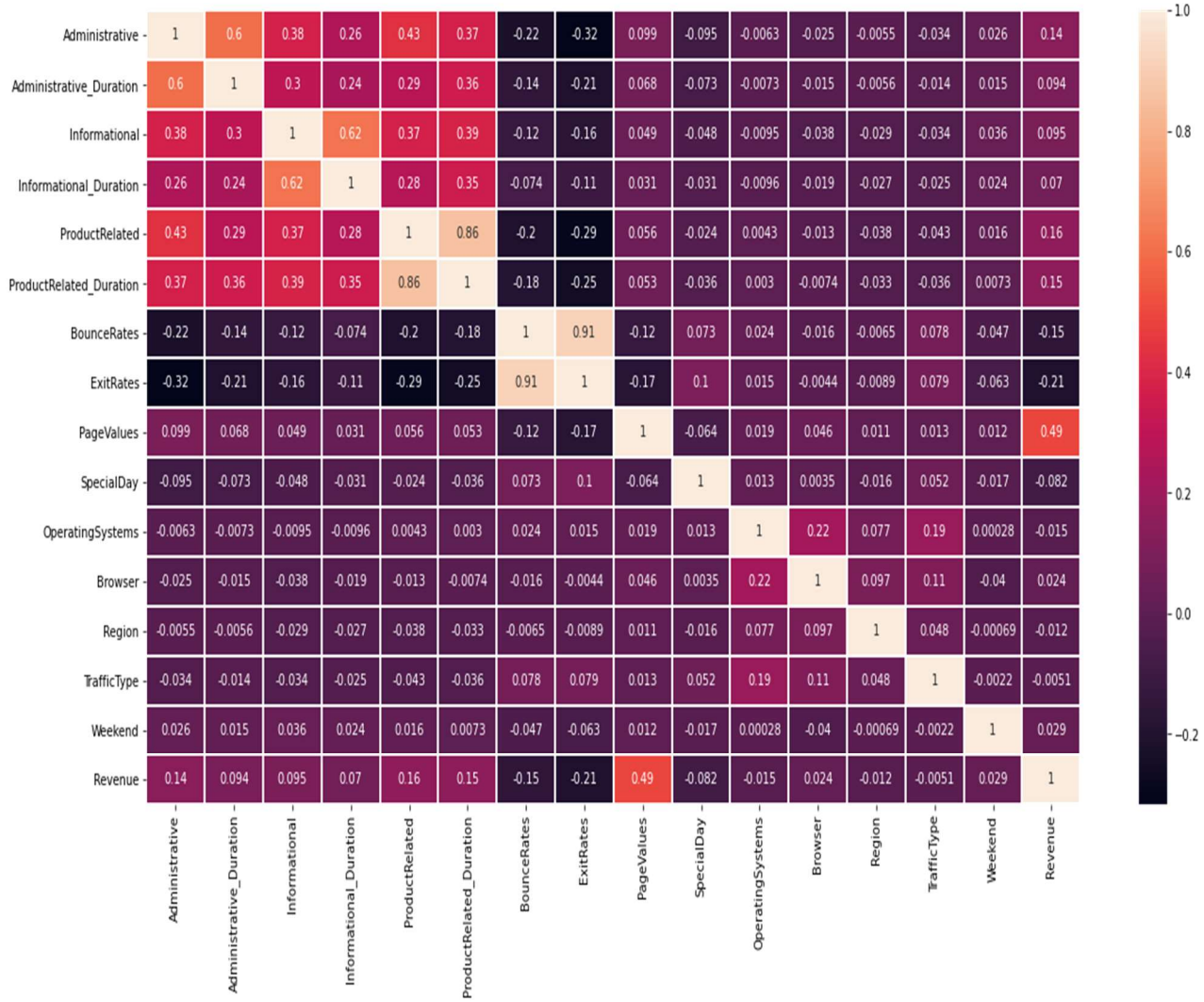
The target feature is the Revenue column. I found that the majority of the visitors were returning visitors.



I looked at the bounce rates and exit rates of different user sessions and plotted a scatterplot to see the spread of revenue around them. Most of the revenue is generated when we have lower exit and bounce rates.



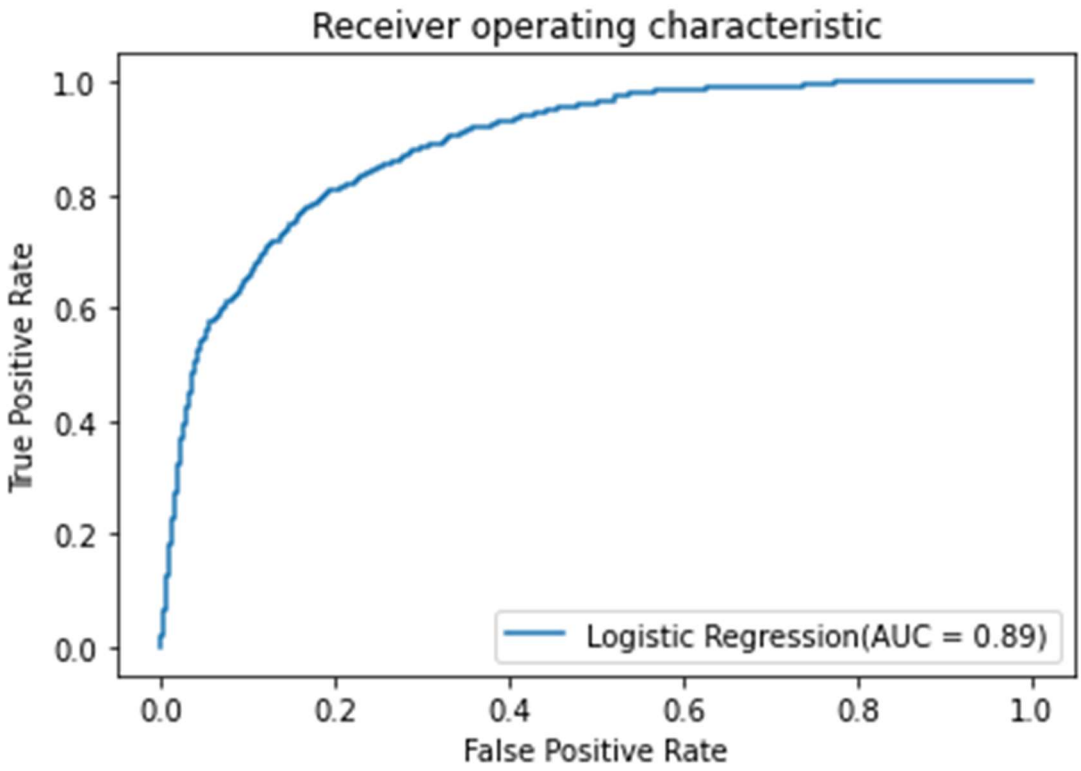
Then I looked at the data correlation by plotting a heatmap and found that there is a high correlation between the page values and revenue. The following figure shows the correlation between the different features.



Model Selection

I used three different machine learning classification models, Logistic Regression, Random Forest Classifier and Support Vector Classifier. The metric I focused on was F1 score as we have an imbalanced dataset.

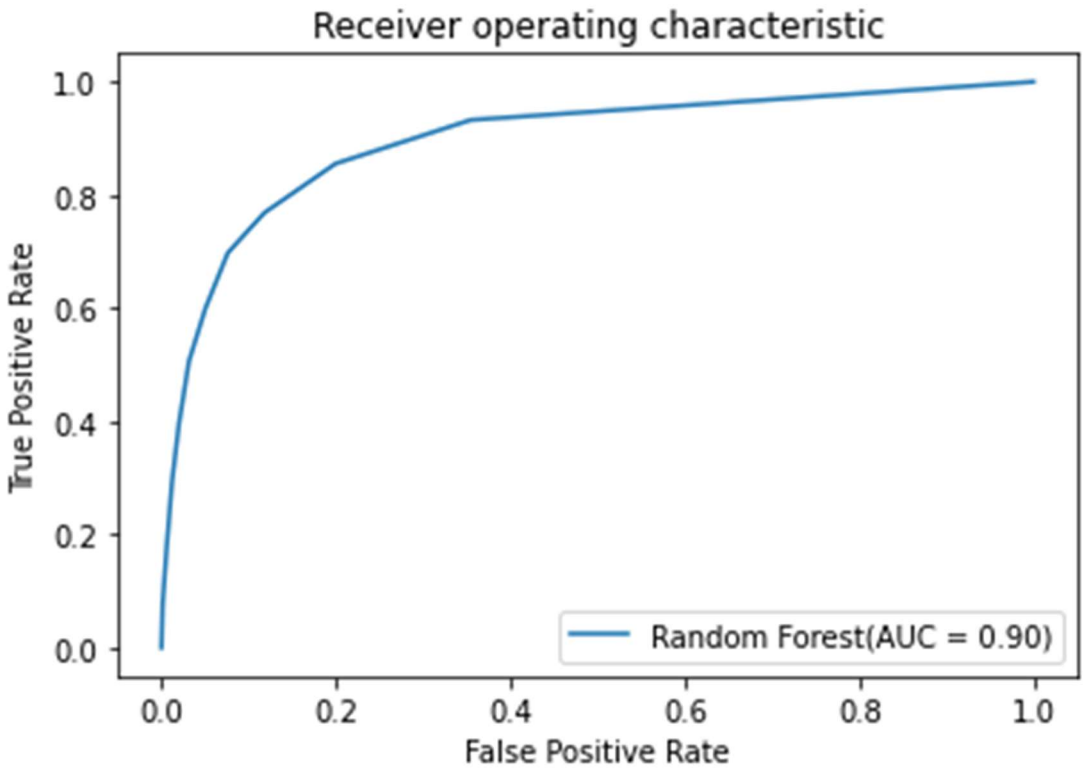
Logistic Regression



Classification Report:

	precision	recall	f1-score	support
False	0.88	0.98	0.93	3077
True	0.76	0.37	0.50	622
accuracy			0.87	3699
macro avg	0.82	0.67	0.71	3699
weighted avg	0.86	0.87	0.86	3699

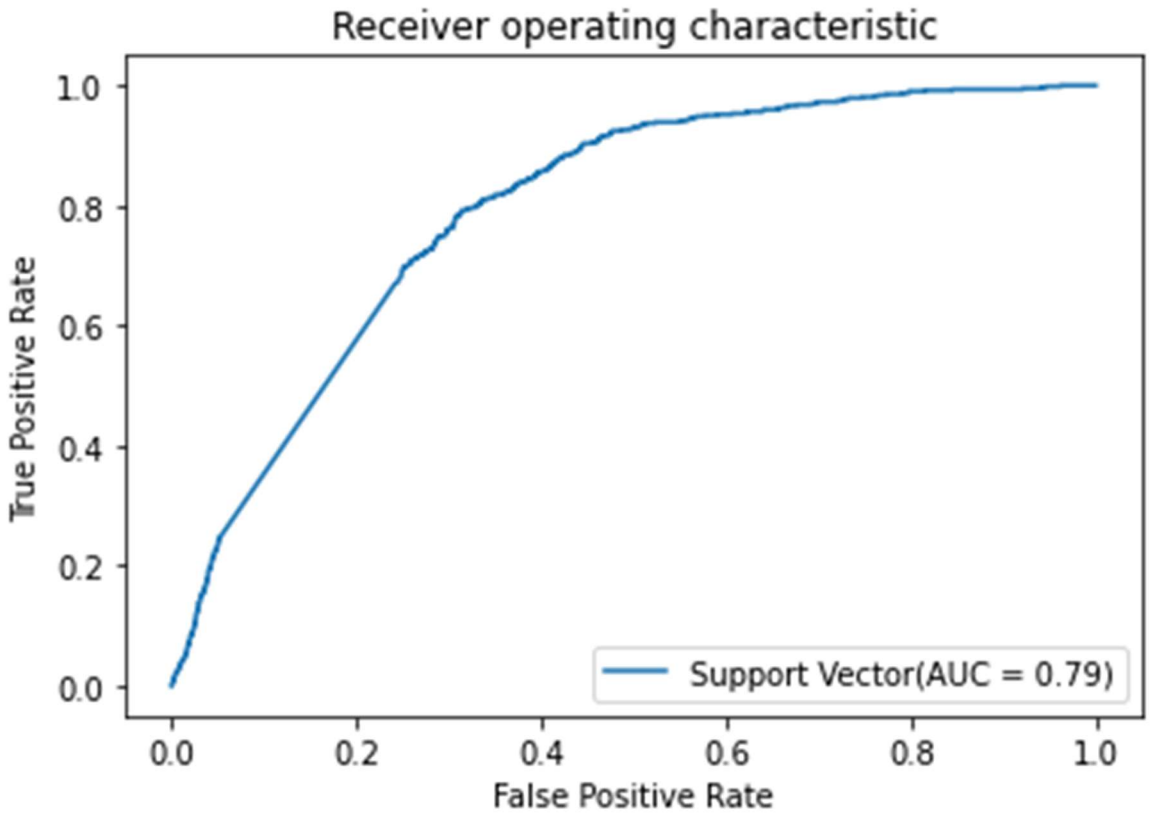
Random Forest Classifier



Classification Report:

	precision	recall	f1-score	support
False	0.90	0.97	0.93	3077
True	0.74	0.49	0.59	622
accuracy			0.89	3699
macro avg	0.82	0.73	0.76	3699
weighted avg	0.88	0.89	0.88	3699

Support Vector Classifier



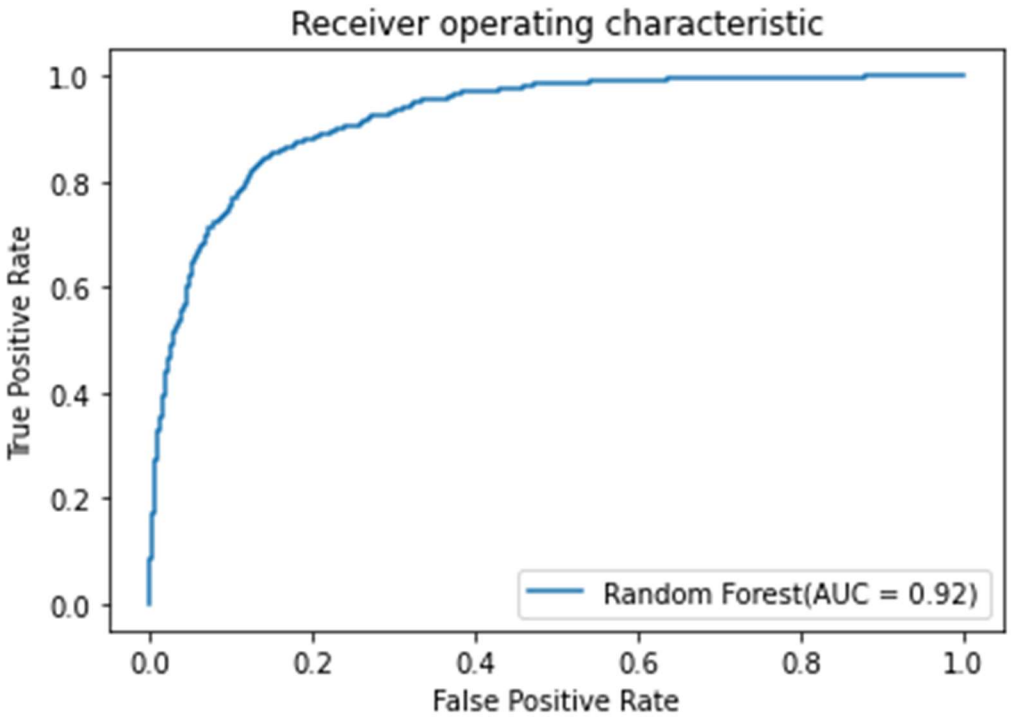
Classification Report:

	precision	recall	f1-score	support
False	0.83	1.00	0.91	3077
True	0.00	0.00	0.00	622
accuracy			0.83	3699
macro avg	0.42	0.50	0.45	3699
weighted avg	0.69	0.83	0.76	3699

We can see from the above ROC curves for the three models that the Random Forest model has the maximum area under curve and it has the maximum F1 score. Hence, this the best performing model out of all the models.

I further applied hyperparameter tuning on the Random Forest model using the GridSearchCV and added the best parameters to the existing model. The accuracy of the model further increased from 88.32 % to 89.18% and also the F1 score increased to 94%. Following is the classification report after hyperparameter tuning.

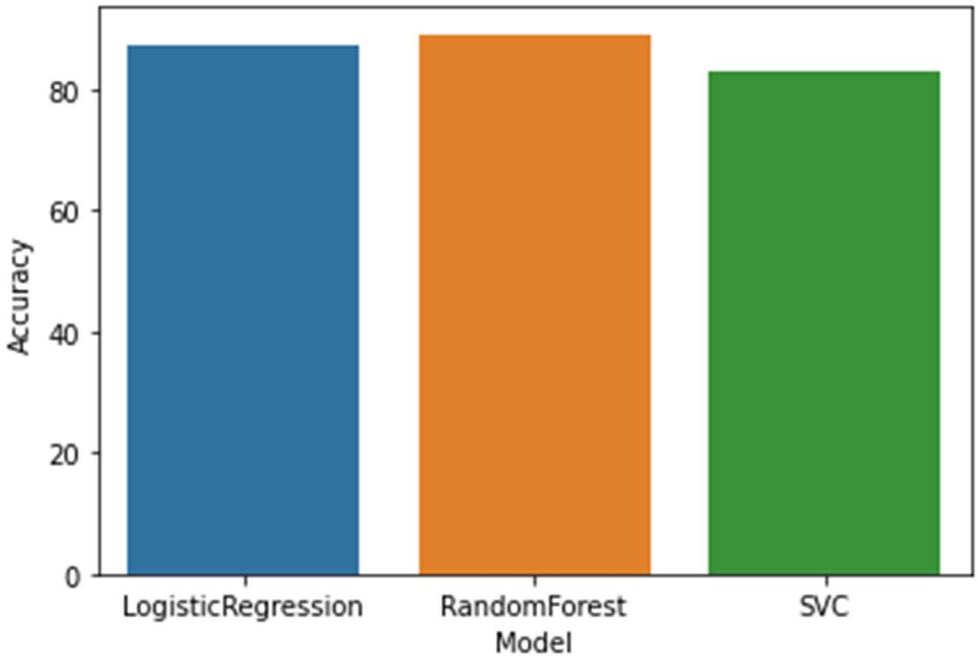
	precision	recall	f1-score	support
False	0.91	0.97	0.94	3077
True	0.79	0.51	0.62	622
accuracy			0.89	3699
macro avg	0.85	0.74	0.78	3699
weighted avg	0.89	0.89	0.88	3699



The following figure shows the comparison of the accuracies of the three models.

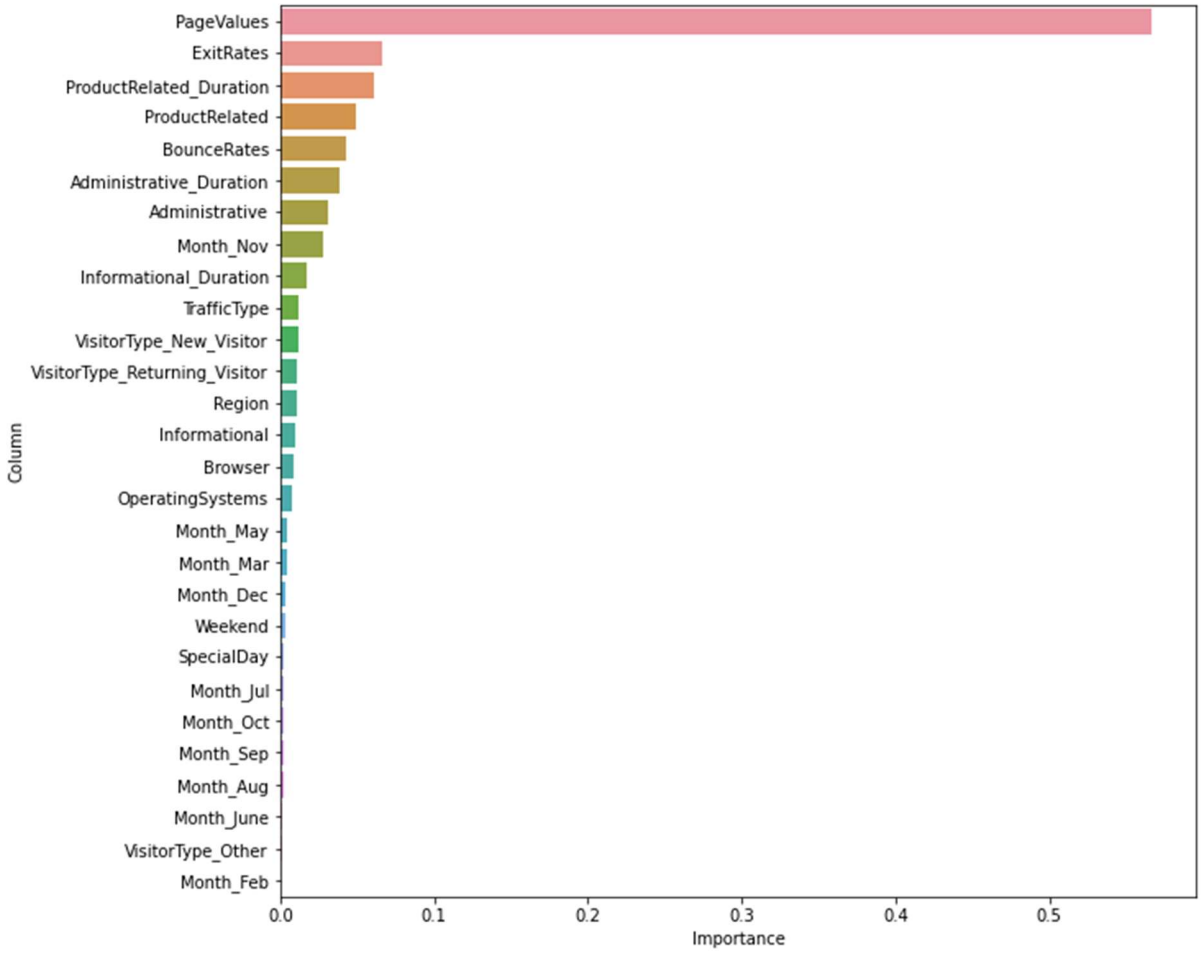
Model	*Accuracy*
LogisticRegression	*87.40*
RandomForest	*89.13*

```
SVC                    83.18
```



Conclusion

The Random Forest model is the best model for predicting the online shopper's purchasing intention. It has a precision of 91% and recall of 97% with an F1 score of 94%. The following figure shows the importance of features in our dataset. The page values feature is the most important feature followed by Exit rates, product related page values and the bounce rates.



The project gave me a good insight into the web analytics space, and I learned about different parameters that are used to measure the performance of the web page and how using this information for each user session we can provide valuable insights about the users which can be further used by the business for growth.

We have limitations with our prediction model since it a small dataset and an imbalanced one. However, the important features we saw in this dataset like page values, bounce rates and exit rates can be used on any larger dataset to predict the purchasing intention.

We can further try running different models, like Naïve Byes, Neural Networks etc and see if it further enhances the model performance.