Predicting Purchases in Online Shops

Sina Alba

February 2022

Colab Notebook Link

Abstract

Purchase prediction has an important role for decision-makers in e-commerce to improve consumer experience, provide personalised recommendations and increase revenue. Therefore, this work aims to evaluate several machine learning algorithm outputs, namely **logistic regression, K-nearest neighbors** (KNN), **support vector machine** (SVM), and **logistic regression with cross validation** for predicting purchases in online shops. The classification algorithms are evaluated based on various measurement criteria, such as accuracy, precision, recall, F1 score and an area under the curve (AUC). Among the algorithms, the KNN algorithm showed better performance. Furthermore, Logistic Regression also showed noticeable metrics after KNN. To put everything into a nutshell, KNN is considered the best algorithm for online shop purchases prediction.

Keywords: Machine learning; Online Shops; Predicting Purchases; Classification; Over-sampling; One Hot Encoding; KNN; Logistic Regression; SVM; Cross Validation

---

1 Introduction

The paper is organized as follows:Section 2 presents the methodology of the research, Section 3 demonstrates the outcomes of different machine learning algorithms, and finally,the paper concluded in Section 4.

2 Methodology

The collected data set included 12330 instances and 18 attributes and one outcome which is the revenue registration. Instances and data set features, handling missing values, exploratory data analysis and visualizations, and application of the machine learning algorithms are explained in this section.

2.1 Instances and Data Set

The data contains the information collected from Department of Computer Engineering, Faculty of Engineering and Natural Sciences, Bahcesehir University, in a 1-year period to avoid any tendency to a specific campaign, special day, user profile, or period. The feature list is shown in **Tab. 1** .

A total 10,422 users (84.5 percentage) were negative class samples that did not end with shopping, and the rest (1908) were positive class samples ending with shopping.

Table 1: Feature List

Serial Features		**Sub Category**	**Data Distributions**	**Data Type**
			Mean +- SD	
1	Administrative	Minimum: 0 Maximum: 27	Mean: 2.31 SD: 3.31	Numerical
2	Administrative Duration	Minimum: 0 Maximum: 3398.75	Mean: 80.74 SD: 176.66	Numerical
3	Informational	Minimum: 0 Maximum: 24	Mean: 0.5 SD: 1.27	Numerical
4	Informational Duration	Minimum: 0 Maximum: 2549.37	Mean: 34.73 SD: 141.36	Numerical
5	Product Related	Minimum: 0 Maximum: 705	Mean: 31.81 SD: 44.6	Numerical
6	Product Related Duration	Minimum: 0 Maximum: 63973.52	Mean: 1192.22 SD: 1905.73	Numerical
7	Bounce Rates	Minimum: 0 Maximum: 0.2	Mean: 0.02 SD: 0.04	Numerical
8	Exit Rates	Minimum: 0 Maximum: 0.2	Mean: 0.04 SD: 0.04	Numerical
9	Page Values	Minimum: 0 Maximum: 361.76	Mean: 5.86 SD: 18.54	Numerical
10	Special Day	Minimum: 0 Maximum: 1	Mean: 0.06 SD: 0.19	Numerical

Table 1 (continued).

Serial	Features Name	Sub Category	Data Distributions Mean +- SD	Data Type
11	Operating Systems	1-8		Categorical/Ordinal
12	Browser	1-13		Categorical/Ordinal
13	Region	1-9		Categorical/Ordinal
14	Traffic Type	1-20		Categorical/Ordinal
15	Visitor Type	Returning Visitor	85.7%	Categorical
		New Visitor	13.6%	
		Other	0.689%	
16	Weekend	False	74.75%	True/False
		True	25.25%	
17	Month	May-Feb		Categorical
18	Revenue	False	84.5%	True/False
		True	15.5%	

2.2 Handling Missing Values

As there is some null values in our data set, we could just get rid of all examples with NA values, but in this case our case we cannot afford that. Hence, we will impute the missing entries with some statistical calculations. Moreover, to handle this problem, we can use random numbers or the mode of data. Hence, in this project, mode is used for both numerical and categorical types.The missing values numbers is shown in **Tab. 2**

Table 2: Missing Values Number

Features	Number of Missing Values
Administrative	115
Administrative Duration	105
Informational	134
Informational Duration	126
Product Related	134
Product Related Duration	139
Bounce Rates	114
Exit Rates	122
Page Values	108
Special Day	129
Month	118
Operating Systems	116
Browser	112

Table 3: Missing Values Number

Table 2 (continued)	
Features	**Number of missing values**
Region	126
Traffic Type	151
Visitor Type	116
Weekend	131

2.3 Exploratory Data Analysis and Visualization

Now we should do some exploratory data analysis and visualizations to understand the data better. Hence, let's look at some instances.
To further determine the path, it is necessary to draw a Correlations Between Variables chart. The relationships between the features can be seen in **Fig. 1**.

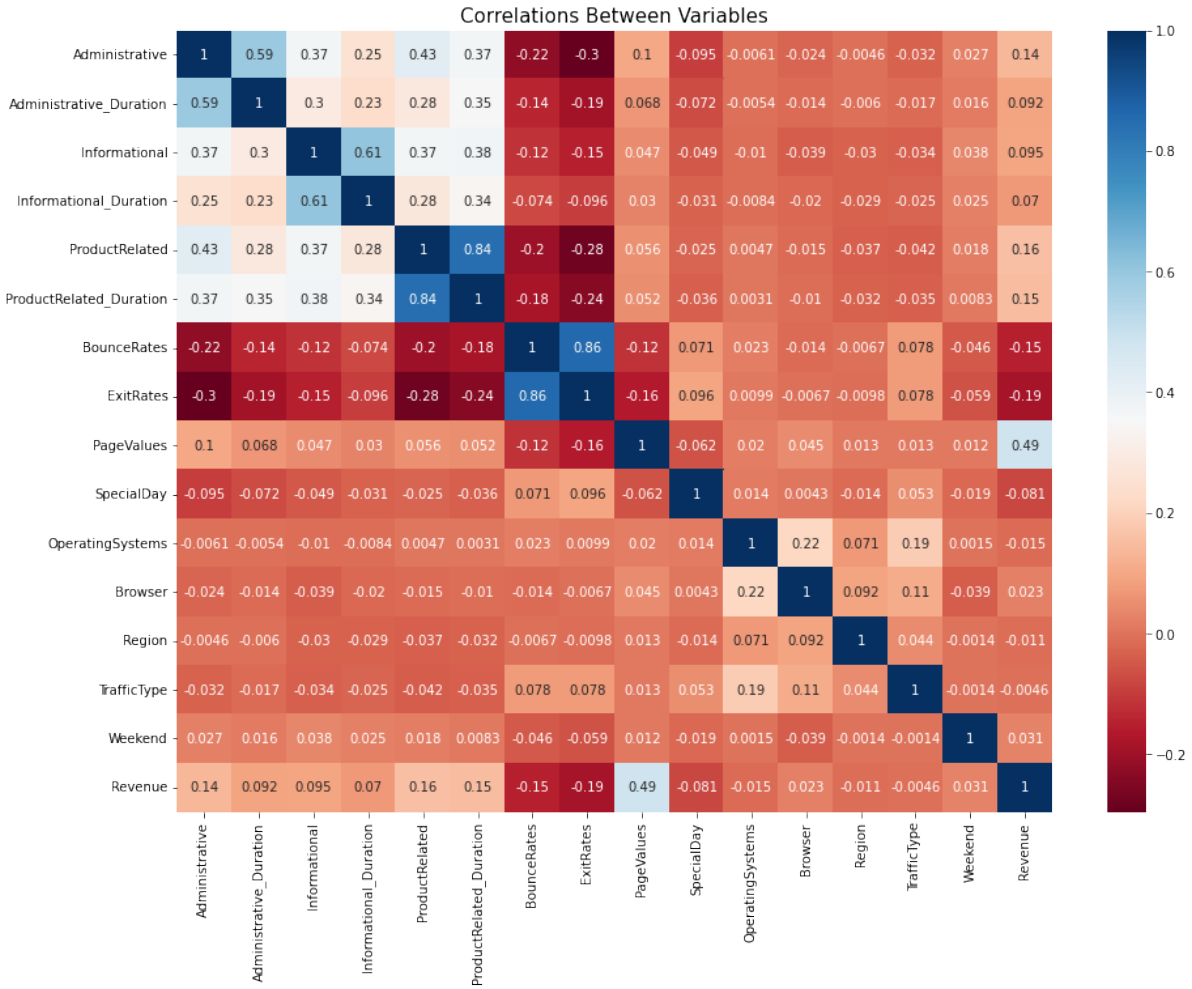

Figure 1: Correlations Between Variables

Also, we can see the correlation of numerical data with target revenue in **Fig. 2**. As can be seen from the diagram, the more colorful the cell, the greater the relationship between them. Also, a lighter color means that the relationship is in the same direction, and a darker color means that it is in the opposite direction. In addition, according to the revenue output tag, the corresponding correlation diagram is drawn in **Fig. 3**.

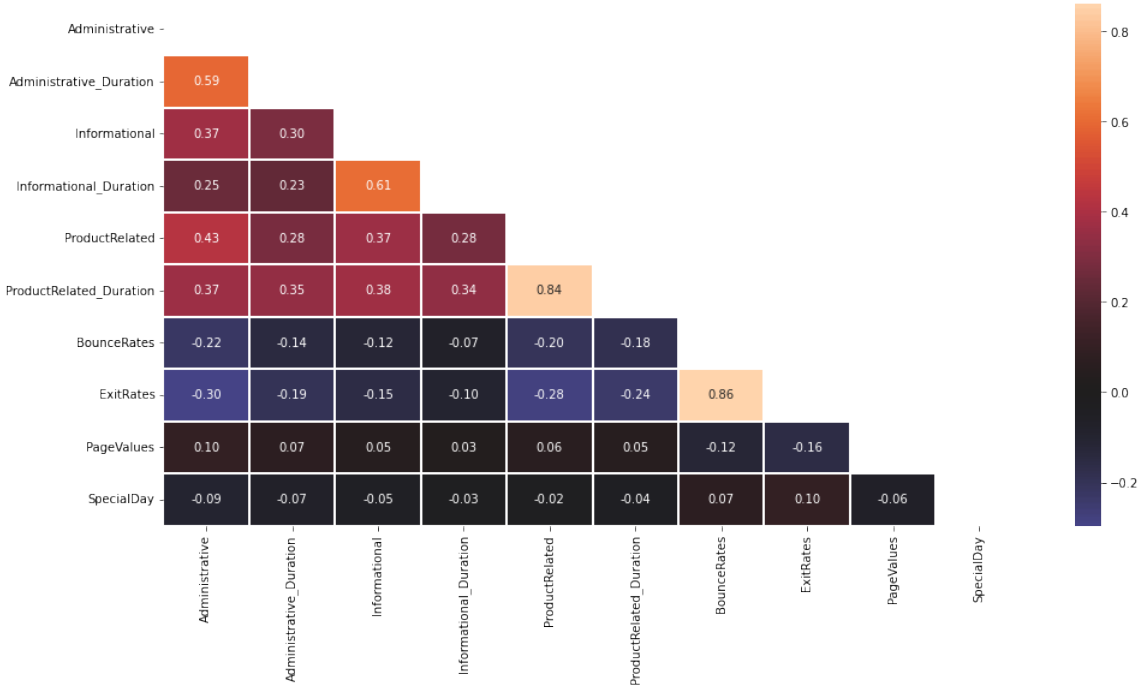

Figure 2: Correlations Between Numerical Variables

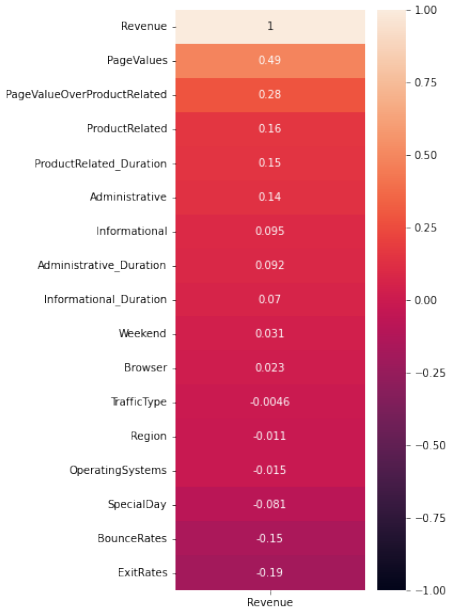

Figure 3: Correlation of Revenue

In addition, a new feature has been created due to its good relationship with revenue called **Page Value Over Product Related**, which divides Page Values by Product Related. Finally, the attributes that have the highest absolute value related to the revenue target are selected. Also, in the meantime, it should be noted that items that have a high correlation with each other should not be selected simultaneously. For example Exit Rate and Bounce Rate due to correlation 0.86, only one should be selected. Furthermore, As a result, the selected features are as follows:
Revenue, Page Values, Exit Rates, Region, Month, Product Related, Weekend, Informational Duration

let's look deeper in data set to get a good vision about chosen features.
At first, it is clear from **Fig. 4** there Were many visitors in the shopping website but very few people have generated the revenue. Also, it is clear that our data is **imbalanced**. As a result, we use the **Over Sampling** method to balance the data.

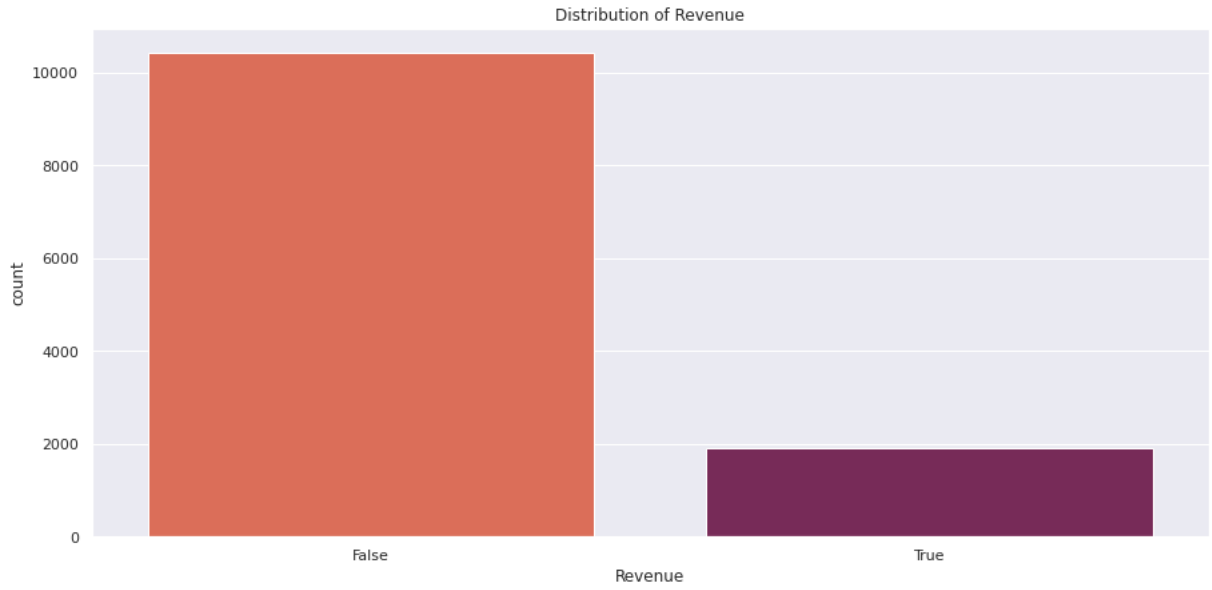

Figure 4: Distribution of Revenue

Moreover, from the **Fig. 5** we can say that very few people have visited the shopping website in the weekend,

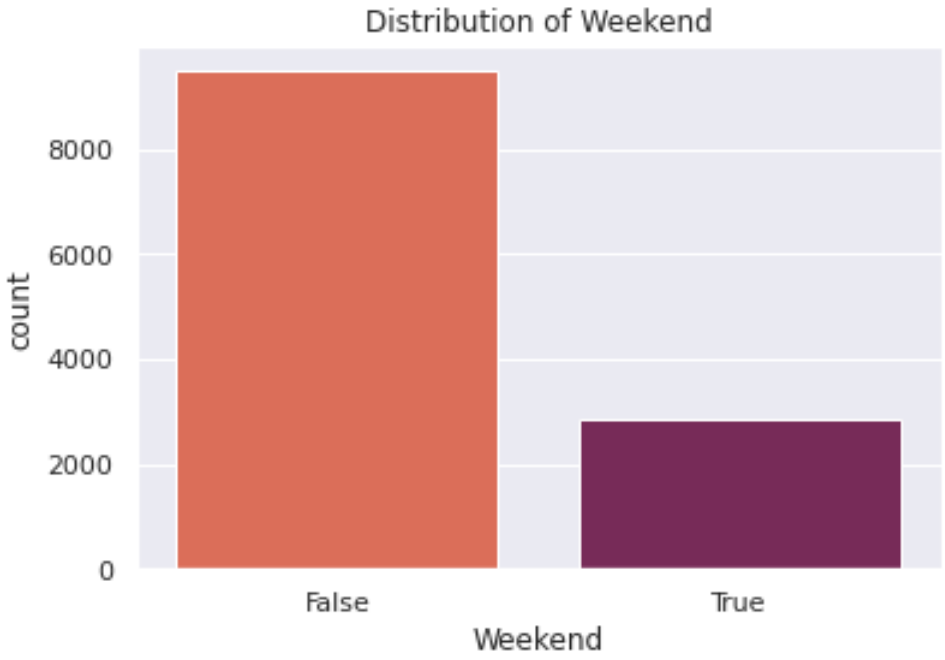

Figure 5: Distribution of Weekend

Also if we look at the distribution of the Region, we can figure out from the **Fig. 6** and **Fig. 7** that most of the visitors and revenue are from region one.

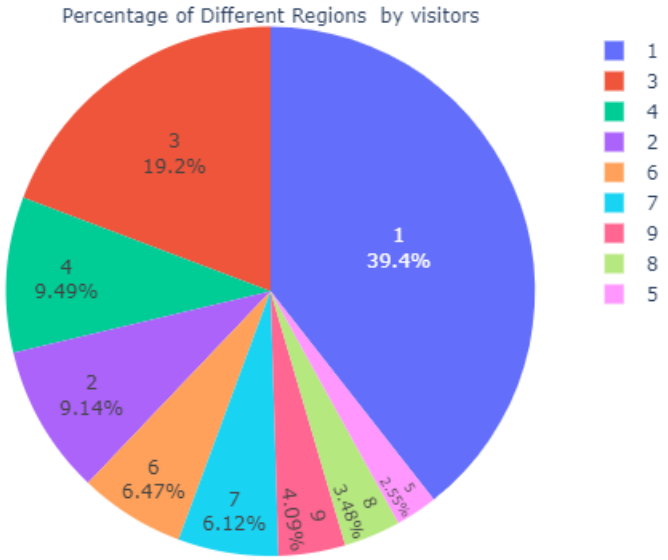

Figure 6: Percentage of Different Regions by visitors

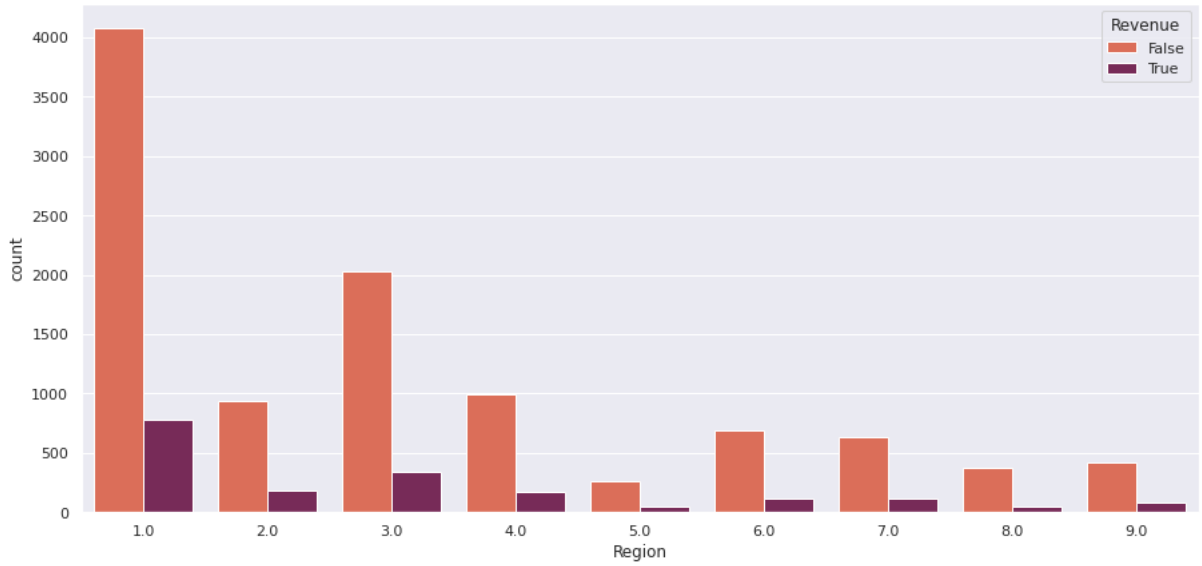

Figure 7: Distribution of people visited by Region with respect to Revenue

As it can be seen from **Fig. 8**, The May had the most visitors and in next is November with 24.1 percentage. But what is the cause? As it can be seen from **Fig. 9**, the number of Special Days in May is higher than in other months, which could be a reason for the site being so popular in this month.

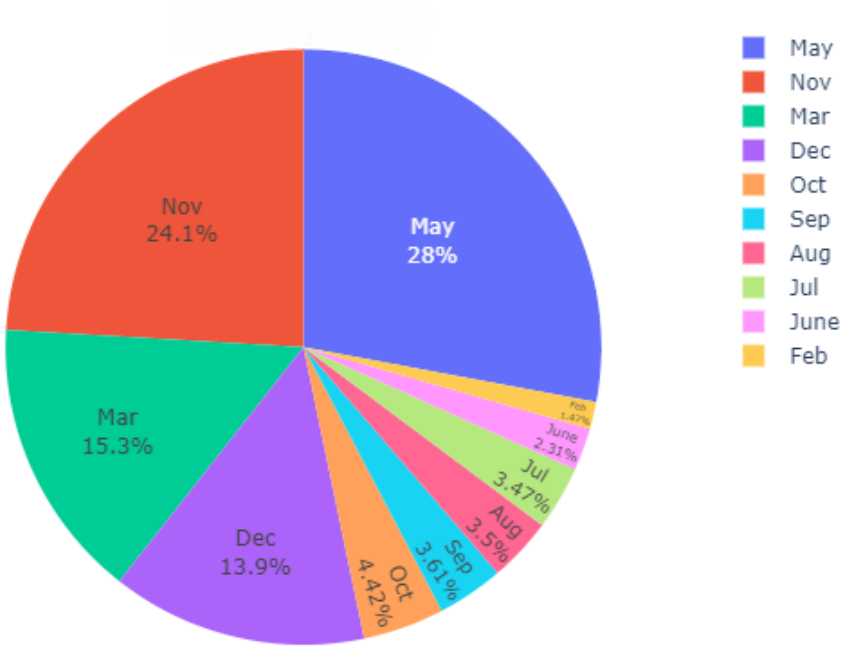

Figure 8: Percentage of Month by visitors

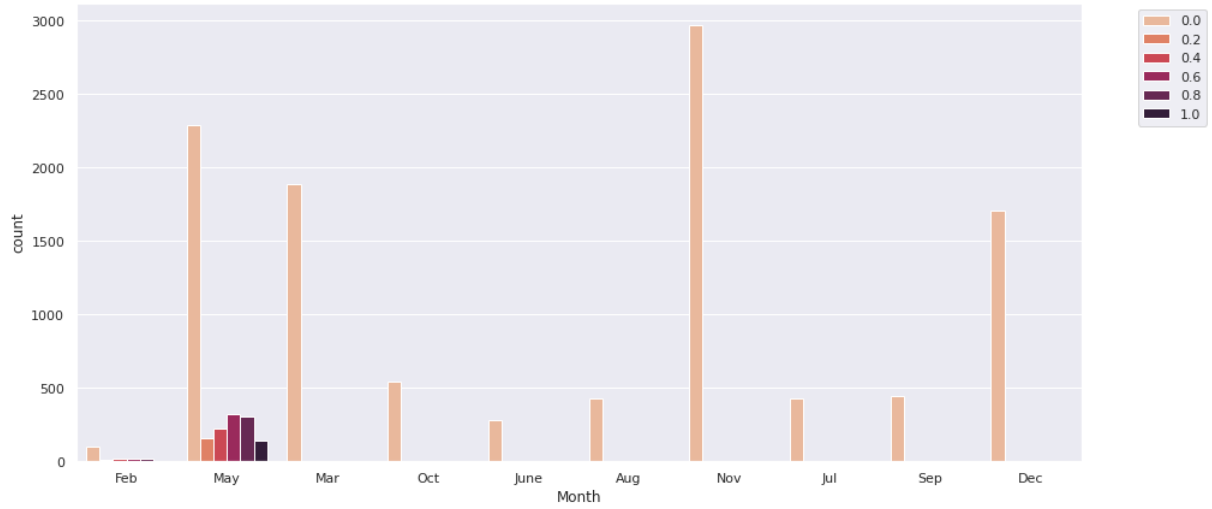

Figure 9: Special Days

For more intuition, we can compare Administrative Duration and Month in **Fig. 10**. As can be deduced most of the revenue is generated in the month of February and In this month the duration of administrative searches are huge.

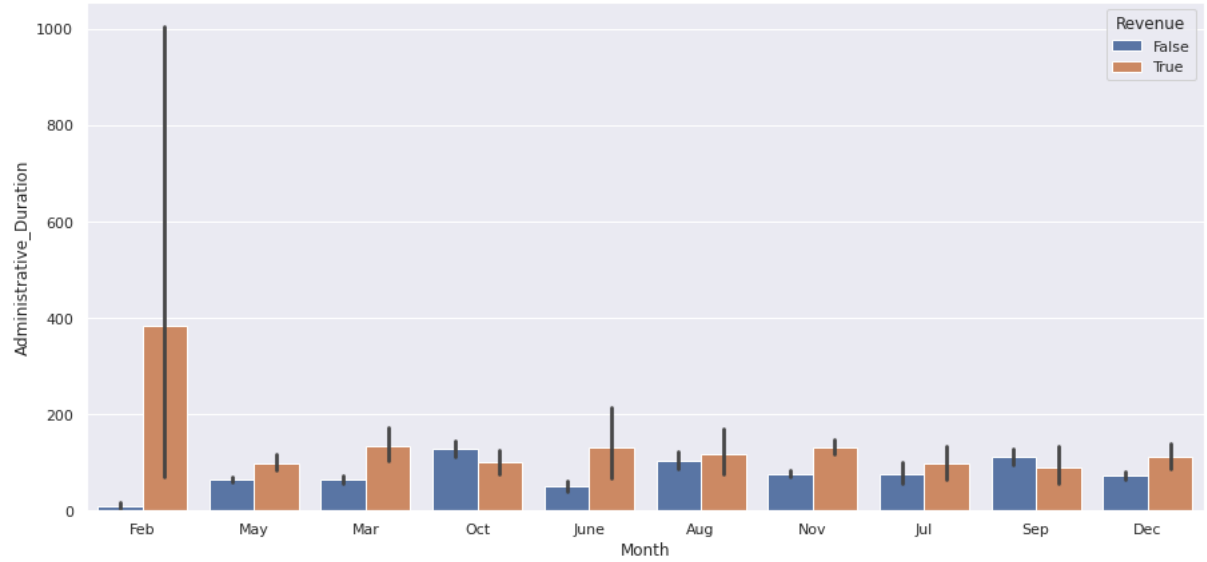

Figure 10: Administrative Duration respect to Month

Moreover, from **Fig. 11** we can figure that revenue is high for informational duration especially in the month of March and July.

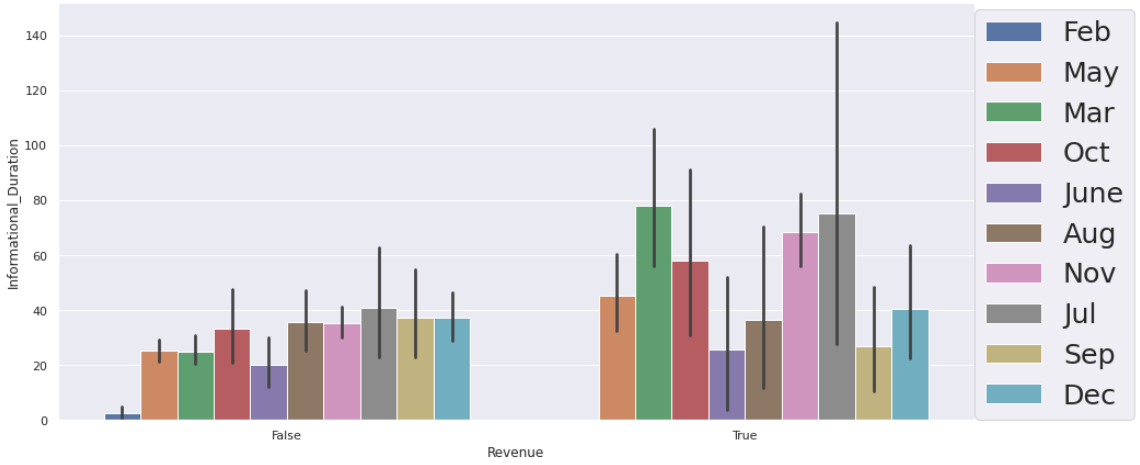

Figure 11: Informational Duration respect to Month

2.4 Application of the machine learning algorithms

This study applied logistic regression, K-nearest neighbors (KNN), support vector machine (SVM), and logistic regression with cross validation algorithms.The algorithms are briefly described as follows.It should also be noted that in the Preprocessing data stage, firstly, the categorical data should be replaced with integers (zero, one) by one hot encoding method. It can also be seen that our data set classes are imbalanced. Therefore, it is necessary to use the Over Sampling method to make our models more accurate.The result of using Over Sampling can be seen in **Fig. 12**.

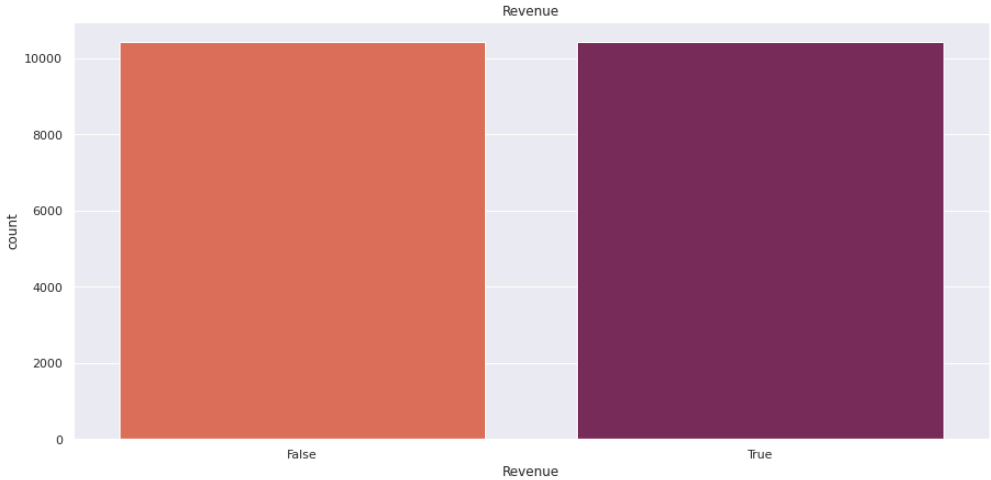

Figure 12: Over Sampling Method

2.4.1 Logistic Regression (LR)

Logistic regression (LR) is generally a linear model that is used for predicting binary variables. LR technique is used for classifying a new observation of an unknown group. LR is a unique system for gathering data into two random and exhaustive data collections.

2.4.2 K-Nearest Neighbors (KNN)

KNN is an elementary classification algorithm of machine learning. It gives an example of the most preferred class among the neighboring K. K is a constraint for changing the classification algorithms.

2.4.3 Support Vector Machine (SVM)

SVM is a supervised measure of learning that can be used for classification and relapse problems, but it is commonly used for characterization problems. SVM functions work admirably and can grasp linear and nonlinear problems.

2.4.4 logistic regression with cross validation

Cross-validation is largely used in settings where the target is prediction and it is necessary to estimate the accuracy of the performance of a predictive model. The prime reason for the use of cross-validation rather than conventional validation is that there is not enough data available for partitioning them into separate training and test sets. This results in a loss of testing and modeling capability.

3 Outcome

3.1 Accuracy

Accuracy is the measurement used for classification evaluation. Accuracy is usually the portion of forecasting authentic prediction.

3.2 Precision

Precision is the dimension of positive predictions that define good predictions.

3.3 Recall

The recall is the component of the accumulated number of actual examples that have already been retrieved.

3.4 F1 Score

The F1 score is an acceptable measure to take advantage of the event. An analogy between precision and recall is found for observing the F1 score, and there remains a jagged class propagation.

3.5 AUC-ROC

The greater the ROC of the AUC, the stronger the portrayal of the model for separating the positive from the negative classes.

3.6 Experimented Analysis

From **Tab. 4**, we obtain the accuracy, precision, recall, F1 score, and AUC for those proposed algorithms. The accuracy, F1 score, precision, recall, and AUC were 79.65%, 79.54%, 80.47%, 79.65%, and 79.74% respectively, in the Logistic Regression algorithm. KNN showed an accuracy, F1 score, precision, recall, and AUC of 88.34%, 88.28%, 88.93%, 88.34%, and 88.27% respectively. SVM had an accuracy, F1 score, precision, recall, and AUC of 73.90%, 73.80%, 74.18%, 73.90%, and 73.83%, respectively. Logistic Regression with Cross Validation shows an accuracy, F1 score, precision, recall, and AUC of 86.53%, 83.06%, 85.11%, 86.53%, 79.77%, respectively.

Table 4: Statistical results (80% train size) of this study

Algorithms	Accuracy	F1_score	Precision	Recall	AUC
LogReg	79.65%	79.54%	80.47%	79.65%	79.74%
KNN	**88.34%**	**88.28%**	**88.93%**	**88.34%**	**88.27%**
SVM	73.90%	73.80%	74.18%	73.90%	73.83%
LogReg with CV	**86.53%**	**83.06%**	**85.11%**	**86.53%**	**79.77%**

Comparing the statistical results, it can be seen that **KNN** has the best criteria among the algorithms. However, the **Logistic Regression with Cross Validation** algorithm has significant outputs. Also, the use of the Over Sampling method has caused the F1 score, precision, recall, and AUC criteria to give very close results to each other.

4 Conclusion

This report demonstrated multiple prediction algorithms to predict online shops purchases. The data set showed different input parameters gathered, and we verified and trained the models for the input parameters given. The prediction of revenue was tested with greater precision by evaluating the algorithms with an attribute collection and data set training. LogReg, KNN, SVM, and LogReg with CV algorithms are constructed to predict. These findings showed that the KNN model accurately predicted the revenue of visitors. Hence, **KNN** is considered the best and most promising algorithm.