Online Shoppers Intention Models Analysis

Hayden Berberich

May 2024

1 Abstract

This study aimed to build a binary classification model using the Online Shoppers Purchasing Intention Dataset from the UC Irvine Machine Learning Repository. This dataset was chosen due to its ability to increase revenue for businesses. It was also selected for the use of a binary classification model. First, histograms are plotted to visualize the distributions of input and output features. Then, the balance of the data is calculated to determine if the data is overly imbalanced. The training process included training models with different amounts of layers and neurons to determine which model works best with this data. The best model was calculated using validation accuracy resulting in an 8-1 model being the most efficient. The 8-1 model was then evaluated using a receiver operating characteristic. Then feature importance was assessed. First, models were trained for each input feature with that input being the only input the model received. Next, a table was created displaying the accuracies of each model trained with its specific input. Then, a bar graph was created to visualize the varying accuracies. This resulted in the input feature "Page Values" being significantly the most important. After assessing the importance of each input feature, different models were created to find the importance of dropping features. Features were iteratively dropped from least important to most important until there was only one feature left. This allowed the analysis of how significant each input feature is for the model as a whole. A graph was then displayed showing the accuracy of the models over the course of dropping input features.

2 Phase 1

In phase 1, I decided to use the Online Shoppers Purchasing Intention Dataset from the UC Irvine Machine Learning Repository. I chose this dataset because it would allow me to detect the actions that were associated with revenue. This information could be used by companies to adjust advertisements and sites to increase revenue. I also chose this dataset because I am new to artificial intelligence and the output column is a boolean that represents whether or not the online shopping session ended in a sale. This means that the model would

be a binary classification problem which is easier than a regression problem. I also chose this dataset because it does not have any missing data which would result in more accurate training.

After loading the data into my notebook, I plotted histograms to visualize the distributions of each input feature. The distributions of month, administrative, and page values can be seen here.

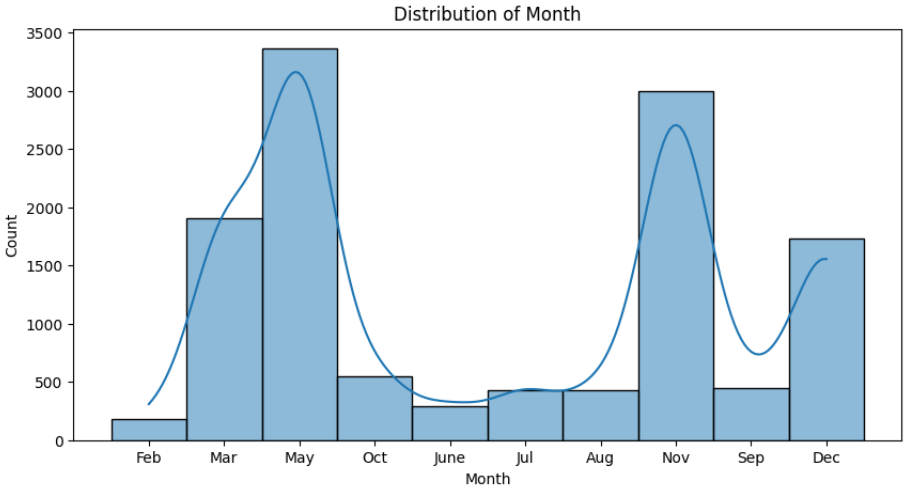

Figure 1: Distribution of Month

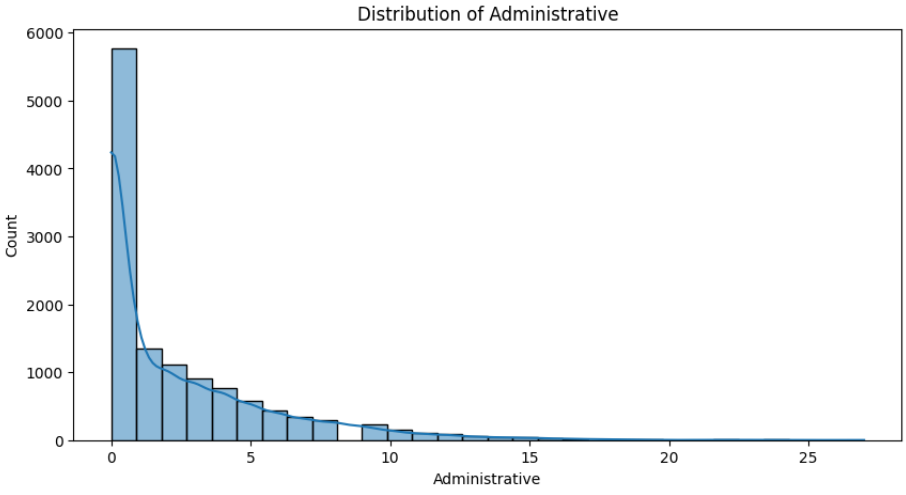

Figure 2: Distribution of Administrative

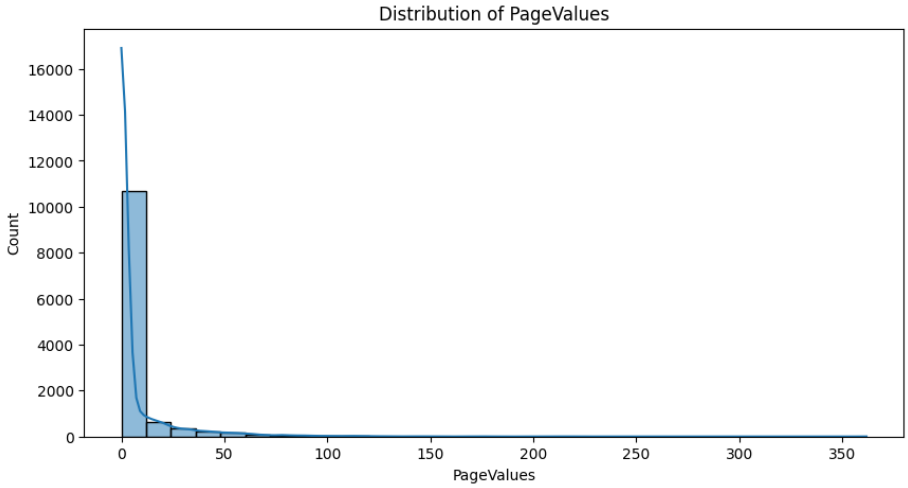

Figure 3: Distribution of Page Values

Then, I plotted the histogram for the distribution of the output label which is revenue.

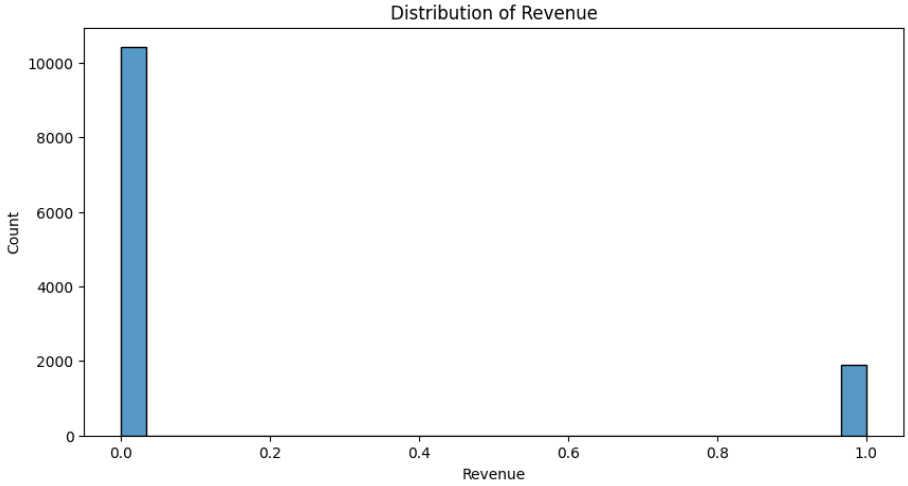

Figure 4: Distribution of Revenue

Then, I calculated the balance of the output data to determine whether or not the data was balanced enough to be used. False is 84.525547 percent and true is 15.474453 percent. This data is imbalanced but not imbalanced enough that it is unusable.

I then checked to see if there are any missing values in the dataset. It returned with nothing missing from any of the inputs or output.

I then normalized the data. I first normalized numerical data and assigned them in a range from zero to one. I then normalized categorical data like month and visitor type. I decided to one-hot encode this data. Finally, I normalized boolean data by assigning 1 to true and 0 to false.

3 Phase 3

In phase 3, I trained multiple models with different amounts of layers and neurons to find the model with the best accuracy. First I shuffled the rows to ensure fair training. Then, I split the data into training and validation sets. I started with a logistic regression model for a baseline classifier that resulted in 88 percent precision, 89 percent recall, and 86 percent F1-score. This model slowly improved over each epoch. Next, I trained a 2-1 neural network model resulting in 90 percent precision, 90 percent recall, and 90 percent F1-score. This model almost improved over every epoch, but slightly slowed down towards the end. Then, I trained a 4-1 neural network model resulting in 89 percent precision, 90 percent recall, and 89 percent F1-score. This model gradually improved but had some epochs towards the end that did not improve. Then, I trained an 8-1 neural network model resulting in 90 percent precision, 90 percent recall, and 90 percent F1-score. This model consistently improved at the beginning but towards the end it was not improving more than it was improving. Then, I trained a 16-8-1 neural network model resulting in 90 percent precision, 90 percent recall, and 90 percent F1-score. This model continuously improved towards the beginning but quickly dwindled down to rarely improving. Then, I trained a 32-16-8-1 neural network model resulting in 89 percent precision, 90 percent recall, and 89 percent F1-score. This model quickly significantly improved but then quickly plateaued. Finally, I trained a 64-32-16-8-1 neural network model resulting in 90 percent precision, 90 percent recall, and 90 percent F1-score. This model very quickly and significantly improved then stopped improving. The table below represents the training and validation accuracies of each model.

Model	Acc. on Training Set	Acc. on Validation Set
Random baseline classifier	87.7%	88.4%
Logistic regression model	87.8%	88.4%
Neural network model (64-32-16-8-1)	90.6%	90.2%
Neural network model (32-16-8-1)	90.7%	90.1%
Neural network model (16-8-1)	91.0%	89.9%
Neural network model (8-1)	90.5%	90.3%
Neural network model (4-1)	89.5%	89.9%
Neural network model (2-1)	89.8%	90.2%

Table 1: Model Accuracies

The 8-1 neural network model has the best validation accuracy so it can be called the "best" model. A reason that the 8-1 model performs the best on

the validation accuracy is due to larger models overfitting the data and smaller models underfitting the data.

The 8-1 neural network model is evaluated using ROC and AUC below.

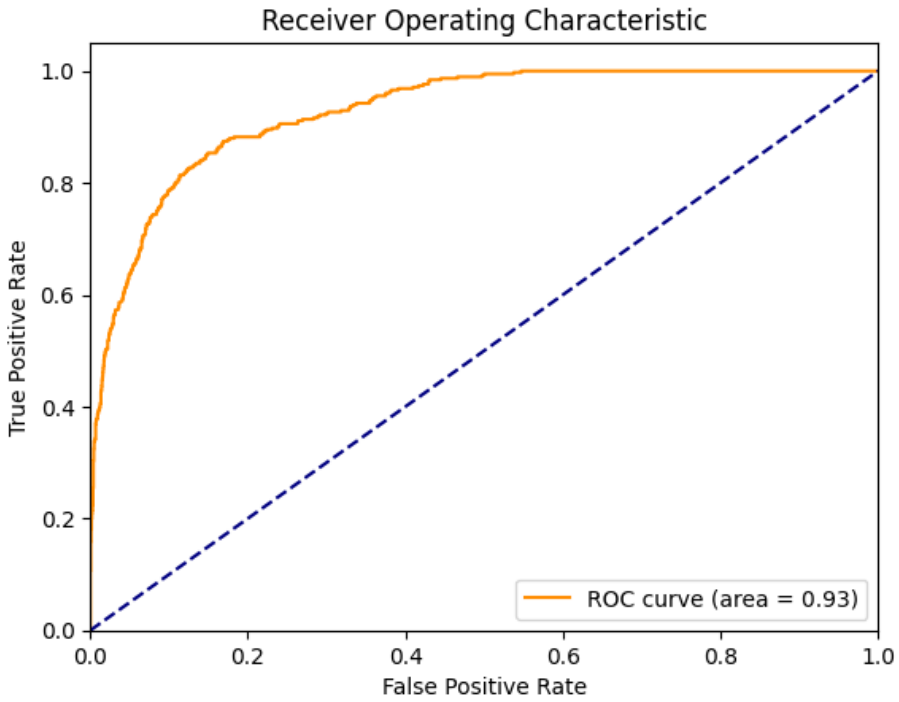

Figure 5: Receiver Operating Characteristic

4 Phase 4

Phase 4 consisted of determining feature importance. First, I trained a model for each input with that input being the only input the model received. This allowed me to check the accuracies to determine feature importance. The results are as follows.

Feature	Accuracy
Administrative	0.8438767194747925
Administrative_Duration	0.8438767194747925
Informational	0.8450932502746582
Informational_Duration	0.8450932502746582
ProductRelated	0.8426601886749268
ProductRelated_Duration	0.8442822098731995
BounceRates	0.8450932502746582
ExitRates	0.8450932502746582
PageValues	0.8751013875007629
SpecialDay	0.8450932502746582
OperatingSystems	0.8450932502746582
Browser	0.8450932502746582
Region	0.8450932502746582
TrafficType	0.8450932502746582
Weekend	0.8450932502746582
Month_Feb	0.8450932502746582
Month_Mar	0.8450932502746582
Month_May	0.8450932502746582
Month_Oct	0.8450932502746582
Month_June	0.8450932502746582
Month_Jul	0.8450932502746582
Month_Aug	0.8450932502746582
Month_Nov	0.8450932502746582
Month_Sep	0.8450932502746582
Month_Dec	0.8450932502746582
VisitorType_Returning_Visitor	0.8450932502746582
VisitorType_New_Visitor	0.8450932502746582
VisitorType_Other	0.8450932502746582

Table 2: Feature Accuracies

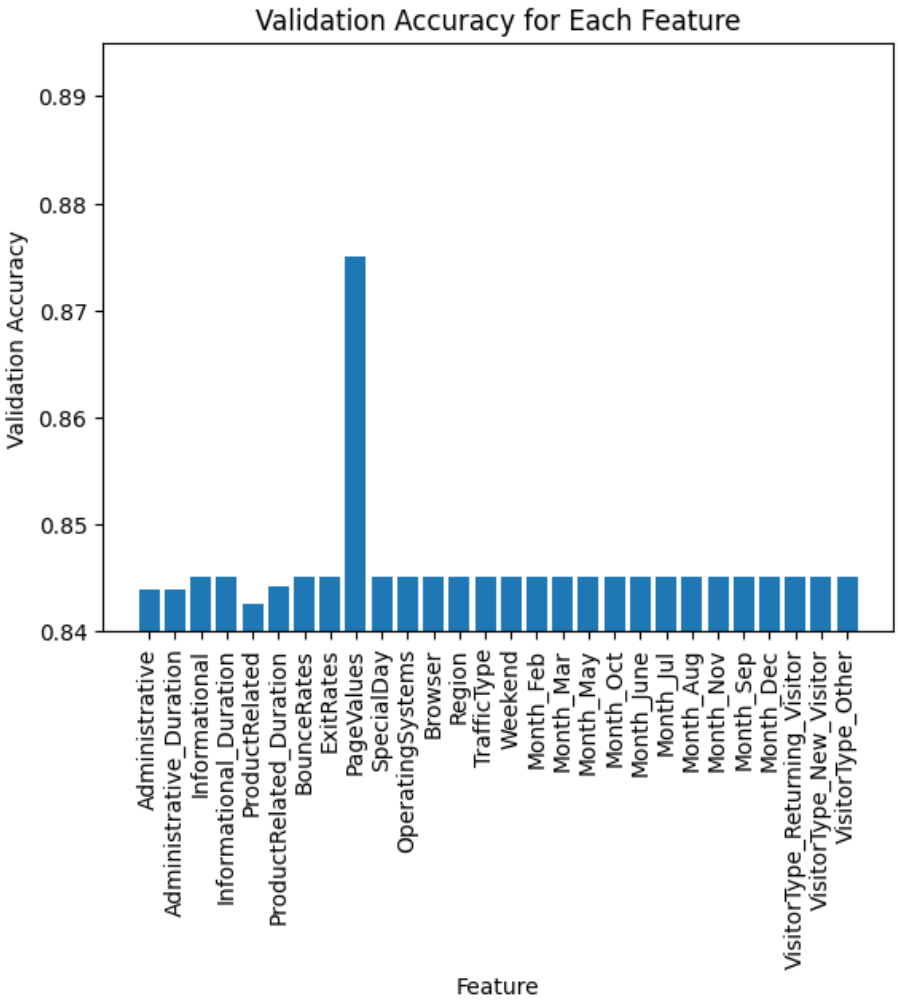

Figure 6: Receiver Operating Characteristic

I then removed features from least important to most important to analyze the effect each input has on the accuracy of the model. I then iteratively removed features and increased the amount of features removed after each iteration. The results are as follows.

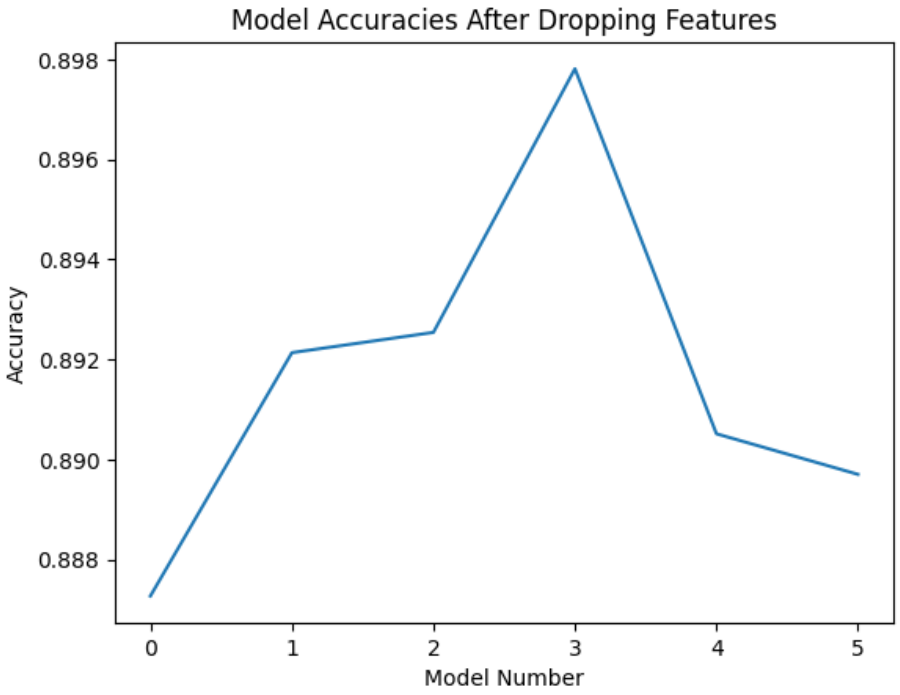

Figure 7: Receiver Operating Characteristic

The results are scattered when you would expect the accuracies to drop when more features are removed. This could be due to the importance of the most important feature, page values. The page values feature is significantly the most important. The "luck" that the model has with page values could determine the outcome of model because the other inputs are trivial compared to it.

5 Conclusion

The study successfully built a binary classification model using the Online Shoppers Intention Dataset from the UC Irvine Machine Learning Repository. After training multiple models to determine the best architecture for this dataset, a model of two layers (one layer with 8 neurons and one layer with 1 neuron) was proven to be the most efficient. It received a training accuracy of 90.5 percent and a validation accuracy of 90.3 percent. The model also received a weighted average precision of 90 percent, a recall of 90 percent, and an F1-score of 90 percent. It was determined that smaller models would underfit the data because they had too few parameters to allow for a large enough capacity to learn the data. Similarly, larger models were seen to overfit the data and learn the datasets noise and outliers so it would perform poorly on unseen data. In addition to model architecture, the study also determined feature importance.

It did so by training a model for each of the input features. This resulted in the "Page Values" input feature being the most important at 87.51 percent. This input feature was significantly more important that all the other features with them falling in the 84 percent range. The study then removed features from least important to most important to analyze the effect each input has on the accuracy of the model. This resulted in scattered results. Rather than the accuracies continuously going down as input features were dropped, the accuracies went up and down. This could be due to the page values feature being significantly more important than the other features. The model was influenced more by how well it could understand the page values input feature rather than the amount of other features.