FINAL REPORT

MACHINE LEARNING CS-584

TO DETERMINE THE PATTERN OF ONLINE BUYERS' PURCHASING INTENTION

GITHUB: https://github.com/Aditya-Shivakumar0301/ML_Project

TEAM MEMBERS:

RAGHUNATH BABU (A20511598) rbabu@hawk.iit.edu
ADITYA SHIVAKUMAR (A20513527) ashivakumar@hawk.iit.edu

INTRODUCTION:

Currently, many businesses rely on the internet to operate, and one effective way to attract customers to their online stores is by promoting deals. In the past, all visitors to an e-commerce website would receive promotions without discrimination. However, online stores have now realized the importance of targeting their advertising campaigns to the right demographic by evaluating visitor information in real-time. This approach helps to establish connections with the most relevant customers and endorse promotions that are more likely to encourage them to return to the website and make successful purchases. Since machine learning is capable of analyzing enormous quantities of information and identify trends and patterns that are difficult to spot via manual analysis, it can be used for predicting the buying intentions of online consumers. Machine learning algorithms may identify correlations and make precise predictions regarding the probability that a customer will make a purchase by analyzing information such as browsing and purchase history, demographics, and other relevant variables.

There are several papers published to predict the online shoppers purchasing intentions. The paper ' Analysis of Different Predicting Model for Online Shoppers' Purchase Intention from Empirical Data' [1] presents an empirical study on the prediction of online shoppers' purchase intention using different machine learning models. The authors compare the performance of logistic regression, decision tree, random forest, and artificial neural network models in predicting purchase intention and conduct feature selection to identify the most important features for prediction. The results show that the artificial neural network model outperforms the

other models in predicting purchase intention. Another paper [2] proposes an algorithmic solution for forecasting buying intention in direct-to-consumer brands using artificial neural networks. The authors collected data from online surveys and used feature selection techniques to identify the most important features for predicting buying intention. They trained and tested the neural network model using the collected data and evaluated the performance using various metrics. The results show that the proposed model can effectively predict buying intention, highlighting the importance of using machine learning techniques for marketing research in direct-to-consumer brands.

PROBLEM DESCRIPTION:

The problem's objective is to predict the purchasing intentions of online shoppers by developing machine learning models and analyzing the trends in the dataset of online shoppers' purchase intentions, The approach is to perform clustering and classification by grouping similar customers based on their purchasing behavior and predict if a customer could do a purchase based on their history of browsing and purchasing behavior. By implementing these methods, we anticipate learning more about consumer behavior and improve promotional strategies that will increase revenue for the company.

DATA SET DESCRIPTION

The data set that is being used in this project was obtained from the UC Irvine Machine Learning Repository.

DATASET SOURCE

The dataset consists of feature vectors belonging to 12,330 sessions. The dataset consists of 10 numerical and 8 categorical attributes.The 'Revenue' attribute can be used as the class label.

Numerical Attributes:

- Administrative, informational, and Product related attributes reflect the number of different categories of pages that the customer visited throughout that time period.
- Administrative Duration, Informational Duration and Product Duration represents the total time spent in each of these page categories.
- The parameters monitored by "Google Analytics" for every web page on a website's online store are depicted by the "Bounce Rate," "Exit Rate," and "Page Value" features.

- The proportion of visitors who arrive at the website through a specific page and then depart it without sending any further requests to the analytics server during that session is commonly referred to as the "Bounce Rate" for that page.
- The proportion of pageviews that were the last in the session is used to compute the "Exit Rate" value for a particular web page.
- The average value of a web page that a user viewed prior to concluding an e-commerce transaction is provided by the "Page Value" feature.
- The "Special Day" feature shows how near a site visit is to a particular holiday , when transactions are more likely to be completed.

Categorical Attribute:

- Operating System attribute specifies the vistor's operating system.
- Browser attribute specifies the visitor's browser.
- Region attribute specifies the geographical location from where the visitor started the session.
- Traffic Type attribute specifies the sources of traffic through which the visitor accessed the web page.
- Visitor Type attribute mentions whether the visitor is a new or an existing (returning) visitor.
- The weekend attribute is set to true if the visitor has performed a purchase during the weekends. It is the to False, if otherwise.
- Month of the Year attribute specifies the month.
- The Revenue attribute is set as True if the user had performed any purchase. It is to False if there is no purchase.

PRE PROCESSING:

1. Chi- Square Test:

The first preprocessing step is implementing the chi-sqaure test. The chi-squared test is a statistical test utilized to determine whether two categorical variables have a significant relationship. It shows that the observed and expected frequencies differ significantly when the chi square test statistic value is large. It also tells if there is a significant relationship between the two categorical variables. In our project, we have found the significance of each predictor variable and the variables that are least significant are removed. (2 categorical variables are removed).

2. Finding the Correlation Matrix:

The next step is to find the correlation matrix. In machine learning, a correlation matrix is a table that displays the pairwise correlations between different variables in a dataset. The correlation

coefficient is a statistical measure which evaluates the strength and direction of a two-variable linear relationship. A correlation matrix can be extremely useful in finding patterns, multicollinearity and can also be helpful for feature selection. In our project, the correlation matrix is found and plotted to have a better understanding about the variables in the dataset.

3. Finding Multi- Collinearity:

The third step that has been performed as preprocessing on the data is to find the multi-Collinearity. When multiple predictor variables in a regression model have a high correlation with one another, something known as multicollinearity takes place. In other words, multicollinearity happens when a model's predictor variable can be estimated linearly from the other variables.

The Variance Inflation Factor (VIF), a measure of multicollinearity in models, is utilized to describe how much the correlation among the predictor variables in the model increases the variance of the estimated coefficient. The output will have a high variance. It is a usual practice to set the VIF threshold at 5, above which multicollinearity is regarded as a challenge. After finding the multicollinearity in the data in our project, the features that have high multicollinearity are eliminated.

$$VIF_i = \frac{1}{1 - R_i^2}$$

4.Dimensionality Reduction:

Once the above steps have been completed, we are finally left with four categorical predictor variables such as Product Related, Product Related Duration ,Bounce Rates, and Exit Rates. The next step is to use Principal Component Analysis (PCA) to minimize the total number of features in the dataset by choosing the principal components that best represent the data's variance. This can enhance the effectiveness of machine learning algorithms by lowering their computational complexity. A variable ‘Product_Related_PCA’ is created to signify the PCA between Product Related variable and Product Related Duration variable. A variable ‘Bounce_Exit_PCA’ is created to signify the PCA between the Bounce Rates variable and Exit Rates variable.

5.One Hot Encoding:

The next step is to implement One Hot Encoding on the data that we have obtained after the several pre processing steps. It is a machine learning method used to encode categorical variables as numerical data. It involves converting each category in a variable into a binary vector, with each vector containing a length equal to the variable's number of unique categories.

CLASSIFICATION:

There are several machine learning algorithms to perform classification. In our project we have chosen some prominent algorithms. They are given below:

1.Logistic Regression

We then started adjusting models. Firstly, we started with a linear model of logistic regression. Logistic Regression operates by fitting a linear decision boundary to the input data and then applying a sigmoid function to translate the linear output into a probability between 0 and 1. After implementing this algorithm, the Mean Cross Validation Roc AUC score of the logistic regression is obtained as 0.916.

$$P = \frac{e^{a+bX}}{1+e^{a+bX}}$$

2.Random Forest

The next step is to implement Random Forest. This popular ensemble learning algorithm makes predictions by combining different decision trees. A random subset of the input features and a subset of the input data are used to train each decision tree in the random forest. We only needed numerical data (categorical variables have to be encoded). Therefore we tried to fit the model with 'almost original' data - from before dimensional reduction. Later, we will also adjust with the same regression data and see where we had better performance. Later, we fitted the model with the transformed dataset and see where we had better performance. Mean Cross Validation Roc AUC score of the Random Forest Algorithm is obtained as 0.90.

3. Neural Networks.

In one of the research papers, they predicted the intention of online customers using machine learning algorithms like Multi layer Perceptron and long short-term memory (LSTM) recurrent neural networks. This made us realize the true potential of implementing neural networks. In our

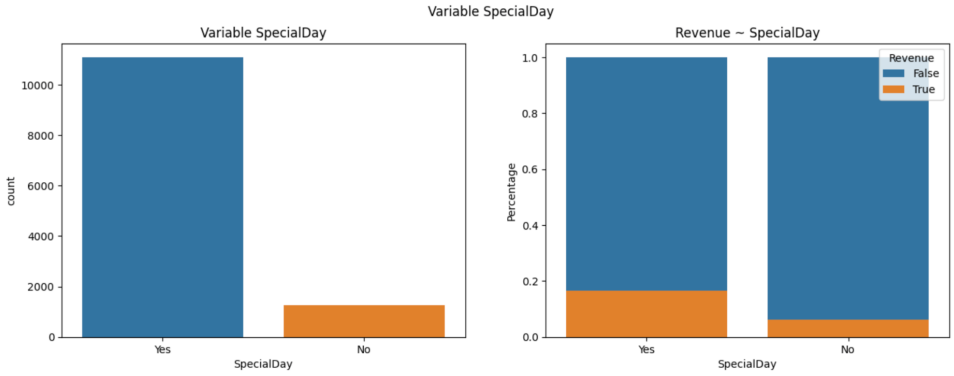
project we have also implemented Neural Networks. Neural Networks also offer a wide range of advantages while solving complex problems. From our implementation, we have obtained an AUC of 0.96.

4. K-means Clustering

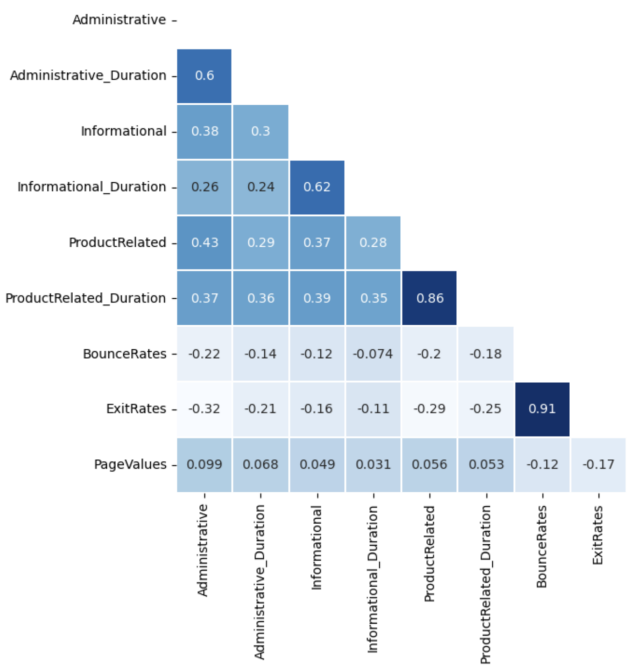
In our project, we have also implemented K-means Clustering. K-means clustering is a well-known unsupervised machine learning approach that separates data into K clusters based on similarity. K-means iteratively minimizes sum of squared distances between data points and centroids until convergence, often determined by elbow analysis. Our code performs the K-means clustering algorithm with the number of clusters ranging from 1 to 10. The within-cluster sum of square is calculated for each number of clusters, which measures the total distance between all points within a cluster. We also ran K-means clustering on a dataset with two features (administrative duration and bounce rates) and three clusters.

RESULTS:

The below graph shows the significance test on 'Special Day' attribute.



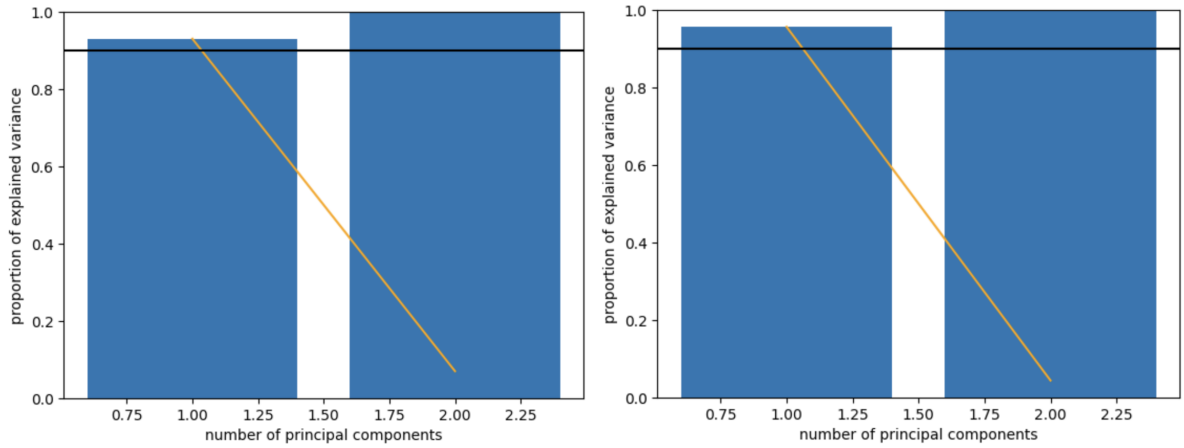
From the correlation matrix plot, high correlation for 4 variables is observed. BounceRates and ExitRates have a correlation of 0.91 and ProductRelated and ProductRelated_Duration have a correlation of 0.86. High correlation values may indicate that one variable is redundant or heavily dependent on the other and should be eliminated to avoid multi collinearity.



VIF for each of the variables is shown in the table below.

	variables	values
0	Administrative	2.650789
1	Administrative_Duration	2.041793
2	Informational	2.113500
3	Informational_Duration	1.777152
4	ProductRelated	6.309248
5	ProductRelated_Duration	6.007085
6	BounceRates	5.483943
7	ExitRates	5.715925
8	PageValues	1.077639

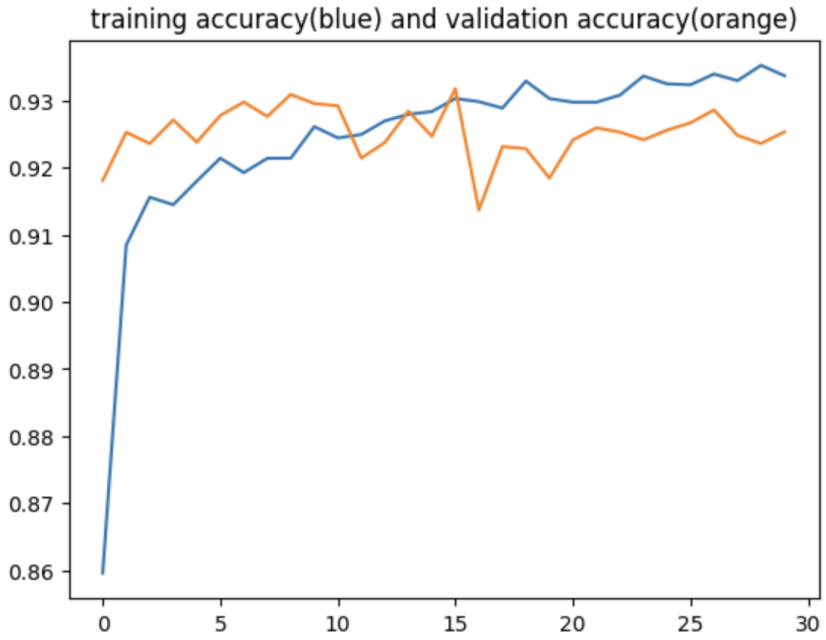
The results of the PCA are then plotted as a line graph and a bar graph to visually show how much variance each principal component explains. The horizontal line at y=0.9 indicates the 90% threshold for the proportion of explained variance.



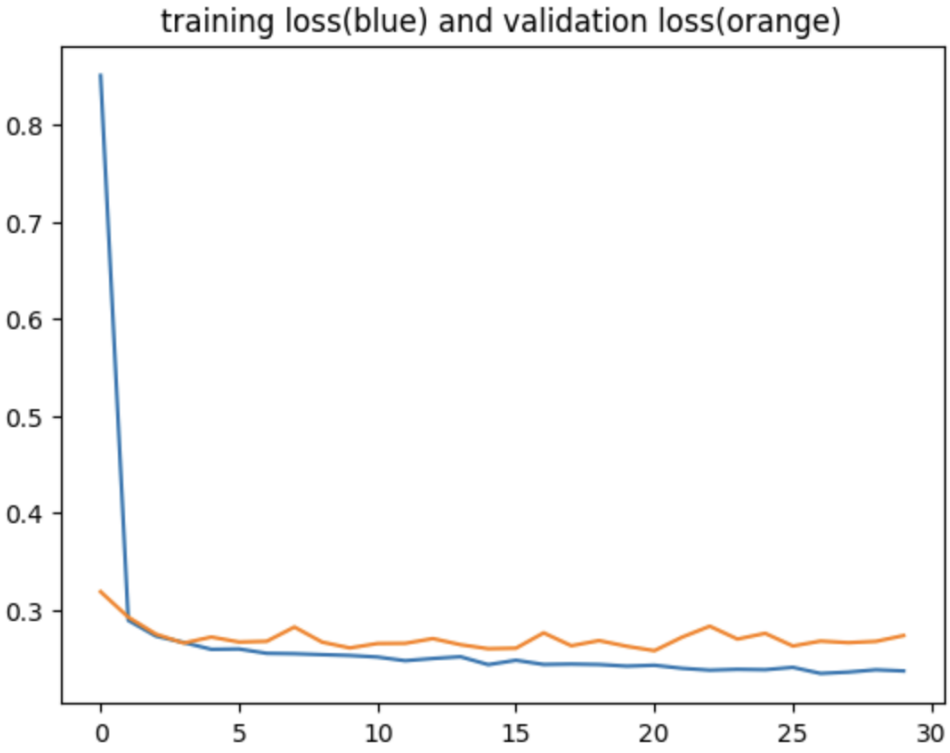
The summary of the neural network model is given below. The mdoel defines and compiles a Sequential model with several Dense layers, a Dropout layer, and binary cross-entropy loss for binary classification.

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense (Dense)               (None, 120)               8040

 dense_1 (Dense)             (None, 256)               30976

 dense_2 (Dense)             (None, 120)               30840

 dense_3 (Dense)             (None, 33)                3993

 dropout (Dropout)           (None, 33)                0

 dense_4 (Dense)             (None, 1)                 34

=================================================================
Total params: 73,883
Trainable params: 73,883
Non-trainable params: 0
_________________________________________________________________
```

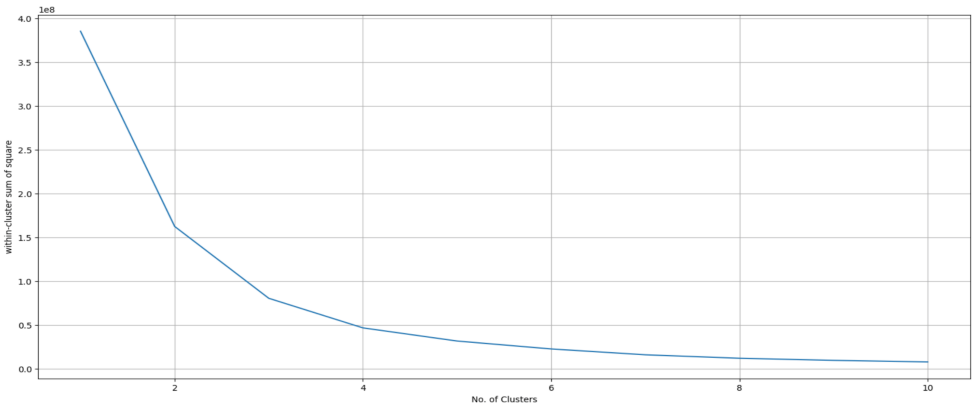
The plot of training and validation accuracy against the number of epochs. From the graph, we can observe that the accuracy increases as the number of epochs of increases. The accuracy stabilizes after a certain point.



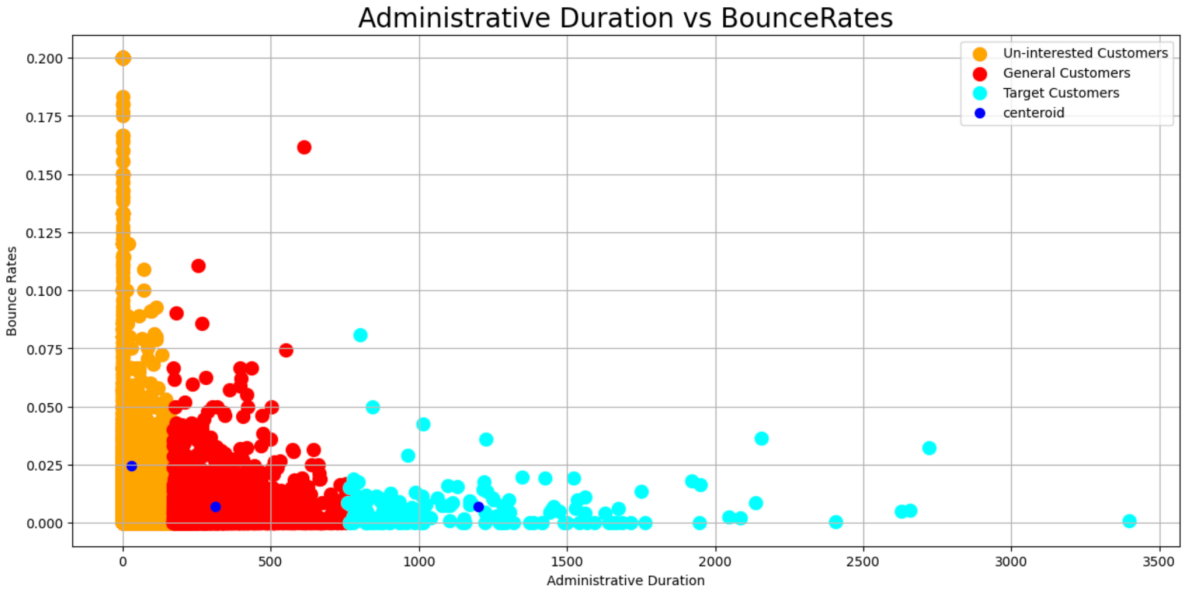
The plot of training and validation loss against the number of epochs. From the graph, we can observe that the loss decreases as the number of epochs of increases. The loss converges after a certain number of epochs.



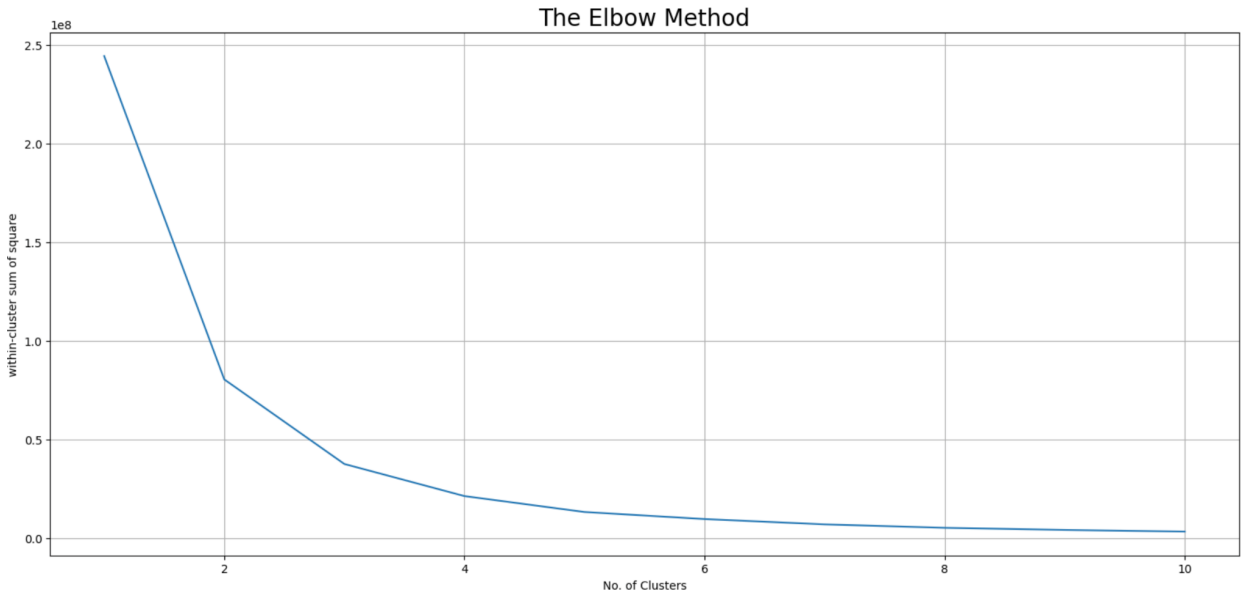
The code generates a line plot of within-cluster sum of squares values against the number of clusters. This plot is used to determine the optimal number of clusters in the dataset based on the "elbow" method, where the elbow point on the plot is used as a guide for selecting the optimal number of clusters. From this cluster plot is we could see the elbow point at k=3 which represent the number of optimal clusters in this particular analysis(Administration_duration vs Revenue).



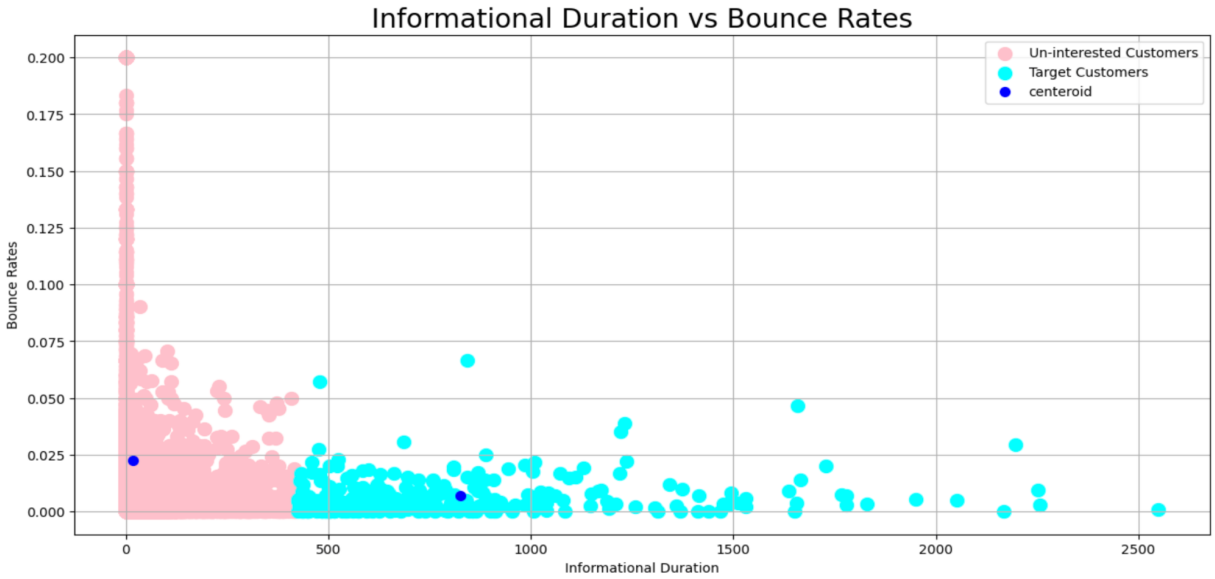
The code then creates a scatter plot with the Administrative Duration on the x-axis and Bounce Rates on the y-axis, where each data point is colored based on its assigned cluster. The code also plots the centroids of each cluster as blue dots. The plot allows us to visualize how the data is clustered based on the two features and how well K-means clustering has separated the data into the intended clusters. From this graph we can infer that as the administrative duration increases the bounce rate also decreases(meaning they will navigate away from the website just after navigating 1 page of the website). These 3 optimal clusters explains about the type of customers who spend shortest administration duration and have high chance of navigating away from a website



The code generates a line plot of within-cluster sum of squares values against the number of clusters. The elbow point(max bend) is 2. The optimal number of clusters for these 2 feature variable comparison is 2. The plot is shown below.



The below graph plots the data points using different colors based on the cluster assignments. The cluster centers are also plotted with a blue color. The x-axis represents the duration of informational pages visited by the customers, and the y-axis represents the bounce rates of those customers.This analysis says that customers who spent a long duration on the website is less likely to bouce to the other website. Group 1(un-interested customers) - short informational duration has high chances of navigating away. Group 2(interested customers) - high informational duration but has very less chance of navigating away.



CONCLUSION:

In the project, the data is preprocessed through chi-square test, correlation matrix, multicollinearity, dimensionality reduction using PCA, and one-hot encoding. Then, prominent classification algorithms are implemented such as logistic regression, random forest, neural networks, and K-means clustering. Mean cross-validation ROC AUC scores are obtained for algorithm to evaluate their performance. Overall, these steps and algorithms were utilized to predict the intention of online customers. In the future, other machine learning algorithms such as support vector machines, decision trees, and gradient boosting could be implemented for predicting online shoppers' purchasing intentions. Additionally, the use of deep learning models, such as recurrent neural networks, may improve the accuracy of the predictions. The incorporation of natural language processing techniques could also be explored to extract

valuable information from customer reviews and feedback. Finally, ensemble methods, such as bagging and boosting, may be used to combine multiple models for better predictions.

For contributions. Aditya Shivakumar performed the initial data pre processing and random forest. Mr. Raghunath Babu performed PCA, Neural Network and clustering. The report was done by both the teammates.

REFERENCES:

1. Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and LSTM recurrent neural networks | SpringerLink
2. Predicting Purchase Intentions with Logistic Regression (relataly.com)
3. Real-Time Prediction of Online Shoppers' Purchasing Intention Using Random Forest - PMC (nih.gov)
4. Prasad, B., & Ghosal, I. (2022). Forecasting Buying Intention Through Artificial Neural Network: An Algorithmic Solution on Direct-to-Consumer Brands. FIIB Business Review, 11(4), 405–421. https://doi.org/10.1177/23197145211046126
5. M. R. Kabir, F. B. Ashraf and R. Ajwad, "Analysis of Different Predicting Model for Online Shoppers' Purchase Intention from Empirical Data," 2019 22nd International Conference on Computer and Information Technology (ICCIT), Dhaka, Bangladesh, 2019, pp. 1-6, doi: 10.1109/ICCIT48885.2019.9038521.
6. Dataset : Online Shoppers Intention UCI Machine Learning | Kaggle