Predicting Customer Purchases with Machine Learning

**Introduction**

In today’s digital age, e-commerce websites play a significant role in retail and business in general. Predicting whether a customer will make a purchase based on their browsing history is essential and important for businesses to optimize customer engagement and increase revenue. Machine learning models in today’s business world can help with these purchasing intentions, allowing businesses to take actions such as target marketing offering discounts or personalized recommendations.

The online shoppers purchasing history intention dataset provides information on various features like session duration, visiters rates in the wibsile and product-related interactions. Our goal in this assignment is to build a machine learning model that can predict whether an online shopper will complete a purchase or not. By doing so, we can provide the business and the company in hand with important insights to improve website design, customer targeting, and marketing strategies to take actions on.

**Problem Statement**

**Business Problem:**

The rise of e-commerce in todays market has made it crucial for businesses to understand user behavior on their websites. Identifying whether a customer is willing to make a purchase during a their visit can help businesses optimize their website experience, improve customer targeting, and increase selling rates. However, predicting such behaviors based on browsing data can be complex due to many factors that influence purchasing decisions made.

**Objective:**

The objective of this project is to predict whether a customer will make a purchase based on their browsing session and history data. Using the online customers dataset, we will build a machine learning model to classify whether a session will lead to a purchase or not.Solving this problem will allow businesses for understanding things like:

Personalize shopping experiences in real time

Allocate marketing efforts efficiently

Improve customer retention strategies

**Data Collection:**

The dataset for this project icludes:

Administrative: Number of administrative pages visited

ProductRelated: Number of product-related pages visited

BounceRates: The percentage of visitors who leave the site after viewing only one page

PageValues: Page value score assigned by Google Analytics

ExitRates: The percentage of page exits

**Data Exploration**

In data exploration we will explore the characteristics of the dataset in hand to understand the nature of the features and identify any potential issues with the data that can effect the outcome.

Initial Exploration: We'll begin by loading the dataset in jupyter and check its size, feature names, data types, and the first few rows to get an understanding of the dataset.

A screenshot of a computer program

Description automatically generatedA screenshot of a computer

Description automatically generated

Checking for Missing Values: Missing values can be couse problem in later steps, so it’s important to identify any and decide how to handle them.

A screenshot of a computer

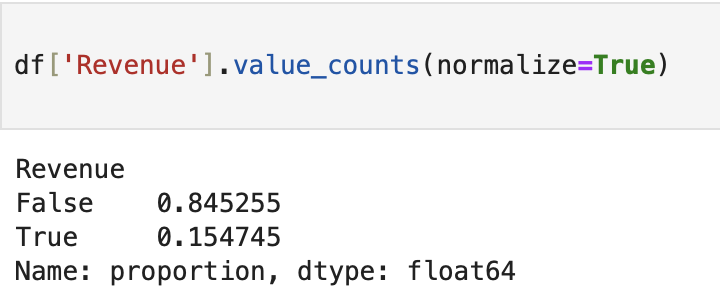
Description automatically generated

Statistical Summary: A statistical summary is very important becouse it gives insights into the distributions of numerical features, allowing to detect things like outliers, skewness, and understand the ranges of different variables in the dataset.

A screenshot of a computer screen

Description automatically generated

Distribution of the Target Variable: The target variable and revenue indicates the success by whether a session led to a purchase. Understanding and knowing the distribution of this information will help the bisiness determine if class imbalance exists, which might require handling during model training.



**Data Preprocessing and Feature Engineering**

In Data processing and feature engineering we are gonna clean the data, handle missing values and create any new features that may improve the model's performance for the company in hand. We'll also ensure the dataset is prepared for training by scaling numerical features if deemed necessary.

Handling Missing Values: If there are missing values, we can either remove them or impute them with the code below in jupyter.A screenshot of a computer

Description automatically generated

Encoding Categorical Variables: Machine learning models require numerical input in order to process the dataset so we are converting categorical variables into numeric vaiables. In this case, we will encode the binary target variable Revenue true and false as 1 and 0, and use one-hot encoding for other categorical variables to get what we want.

A screenshot of a computer

Description automatically generated

Feature Scaling: Some machine learning algorithms improve and get benfits from scaling the features so that they all have similar ranges same as we are gonna do it to this dataset below. We’ll scale the numerical columns using StandardScaler in jupyters notebook.

A screenshot of a computer

Description automatically generated

Feature Selection: we can use techniques like correlation matrix or feature importance for the dataset from tree-based algorithms to find the most impactful and useful features in jupyters notebook to reach the results we want. For example below below:

Correlation matrix to check relationships between features in the dataset

Compute the corelation matrix

Plot the corelation matrix

A screenshot of a computer code

Description automatically generated

A screenshot of a computer screen

Description automatically generated

Splitting Data into Training and Testing Sets: and also to evaluate the model in hand properly, we need to split all the data into training and test sets, define features X and target Y and check the shape of the training and testing sets.

A screenshot of a computer program

Description automatically generated

**Model Training**

In this step, we will train a classification model in jupyter to predict whether a customer will make a purchase that is the Revenue column. We will start with a simple model, such as Logistic Regression, and then move on to more advanced models like Random Forest or XGBoost if necessary until we get the rsult we want for the next steps.

**Logistic Regression:**

We'll start with logistic regression, a good start for model for binary classification problems in the dataset in hand. These steps are training the model, making prediction on the test set, evaluating the model, displaying the confusion metrics and displaying the classification report.A screenshot of a computer

Description automatically generated

**Random Forest Classifier:**

Next, we'll try a more advanced model like Random Forest, which often performs well for classification tasks and is suitible for this dataset. The steps in here are also similar to the steps above, like evaluating the model and displaying the classification report.

A screenshot of a computer program

Description automatically generated

**Model Comparison:**

Now we can compare the accuracy of the Logistic Regression and Random Forest models to see which one performs better here. In most cases, Random Forest should give better results, but logistic regression provides a good baseline for the data.

A screenshot of a computer

Description automatically generated

**Model Evaluation**

now we evaluate the performance of the models trained in the previous codes we run on jupyter. We'll discuss the results, model limitations, and business implications.

**1. Model Evaluation**

We’ll compare the Logistic Regression and Random Forest models based on accuracy, confusion matrix, and classification report we got to determine the better-performing model for this business problem in hand.

Confusion Matrix: It gives insight into the true positives, false positives, true negatives, and false negatives, which helps in understanding the types of errors the model is making for us.

Classification Report: This includes metrics such as precision, recall, and others which give a more detailed and elaborated view of the model’s performance, especially if the dataset is not balanced.

Here is how the evaluation results might look:

A screenshot of a computer screen

Description automatically generated

2. **Strengths and Limitations of the Model**

**Logistic Regression:**

Strengths: Easy to use, fast, and simple to understand . It works well if the relationship between the data features and the target are mostly linear.

Limitations: Doesn't perform well with complex and heavy datasets that have non-linear patterns. It doesnt capture how features interact with each other.

Random Forest:

Strengths: Can handle complex, non-linear patterns and interactions between features in the dataset. It usually works well across different datasets for us.

Limitations: Harder to interpret and understand compared to logistic regression. It can overfit, but this can be controlled by using more trees in the model.

**3. Business Implications**

For the online customer buying intention dataset predicting if a visitor will make a purchase can bring a lot of value to the business:

Better Marketing: The model helps target marketing efforts toward customers who are most likely to buy the products of the company.

Personalized User Experience: The website can be adjusted to better engage users who are predicted to make a purchase from the website.

Inventory Management: The model can help forecast demand more accurately which can save money by optimizing inventory.

**4. Data-Driven Recommendations**

Based on the results, we suggest:

Use Random Forest: Since it performs better than logistic regression in accuracy and precision, it should be the go-to model for predicting customer purchases.

Further Feature Engineering: Look into adding more features that might improve performance, like customer demographics or browsing acting.

Monitor and Update the Model: Keep an eye on the model’s performance and update it with new data regularly and always to keep it accurate as customer demand and behavior changes.

**Final Discussion and Business Recommendations**

To wrap up, we’ve observed from all the results we got from jupyter on many key findings. Through data analysis, we noticed that customers who spent more time on product pages and visited the site on weekends were more likely to make a purchase. This insight can help shape future marketing strategiesin the company. Our random Forest model over performed Logistic Regression in accuracy, suggesting that non-linear models are better suited for predicting customer purchase intentions from the dataset. Additionally, features like ProductRelated\_Duration and Bouncerates were more influential in predicting purchase behavior than others.

However, there are a few limitations. The dataset is limited to session-level features, and adding more customer-specific data, such as demographics or past purchases, could improve the model's performance and give a better result. The issue of class imbalance where there are significantly more nonpurchase sessions than purchase sessions could be addressed and solved by using techniques like SMOTE. Also external factors like promotions, seasons, or economic conditions were not considered, and including these could lead to better predictions for the business planing of the company.

Based on our findings, several business recommendations in here emerge. Marketing should target high value and rich customers. Personalized offers could help convert these users into paying customers for the company. The website experience can also be optimized by improving the design and flow of product pages on the website particularly for users with high bounce rates. Using the Random Forest model for real-time prediction of purchasing intentions can inform targeted advertising for us. It’s essential to monitor the model’s performance over time, updating it with new data to ensure its accuracy as customer behavior evolves each day or year.

**Conclusion**

The machine learning model developed by the essay writer in this project provides a very good and strong foundation for predicting customer purchase intentionsa and also offering businesses ideas for better marketing and website optimization. By leveraging predictive analytics, companies can make more informed and analysed decisions, increase sales, and improve customer happiness. While the model shows promise, there is room for improvement through the addition of more features and tools. Further tuning and continuous updates are also important to adapt to changing customer behaviors.

**GITHUB LINK:**

[**https://github.com/alijaweddelawari/B104**](https://github.com/alijaweddelawari/B104)



Assessment Submission Form

|  |  |
| --- | --- |
| **Student Number**  (If this is group work, please  include the student numbers of all group participants) | GH1024093 |
| **Assessment Title** | Predicting Customer Purchases with Machine Learning |
| **Module Code** | B104 |
| **Module Title** | Artificial intelligence and machine learning |
| **Module Tutor** | Mohammad Mahdavi |
| **Date Submitted** | 25/09/2024 |

**Declaration of Authorship**

I declare that all material in this assessment is my own work except where there is clear acknowledgement and appropriate reference to the work of others.

I fully understand that the unacknowledged inclusion of another person’s writings or ideas or works in this work may be considered plagiarism and that, should a formal investigation process confirms the allegation, I would be subject to the penalties associated with plagiarism, as per GISMA Business School, University of Applied Sciences’ regulations for academic misconduct.

Signed………………Ali Jawed Delawari……………. Date ………25/09/2024………………