PROJECT REPORT

for

Online Shopper's Purchasing Intention Prediction

BIG DATA ANALYTICS

Submitted to:
Dr. Shishupal Kumar

SWARIT DEEPAK PATEL (BT20CSE093)

KHUSHIRAM MEENA (BT20CSE167)

29th APRIL 2023

Table of Contents

1. Introduction

The objective of this project is to predict the purchasing intention of online shoppers using machine learning models. We will be using Apache Spark, a fast and distributed data processing engine, to handle large datasets and perform complex analytics. By using machine learning algorithms, we can identify patterns and trends in customer's behaviour, which will help us understand the factors that influence purchasing decisions.

Our dataset consists of various features such as the **Administrative Information**, **Information about Operating system**, **Types of Traffic**, **Region, Visitor's types, Bounce Rates, Exit Rates, Product Information, the number of pages visited** & **Revenue etc**. By using this dataset, we will build a logistic regression model to predict whether a customer will make a purchase or not.

1.1 Purpose

The project's ultimate goal is to provide e-commerce websites with insights into their customers' behavior and preferences, which can help them make data-driven decisions to improve their sales and customer satisfaction. We believe that this project will be valuable for online retailers who want to gain a competitive edge in the market.

1.2 Used Technology

- Apache Spark
- Spark SQL
- Apache Spark MLLib
- SCALA
- ML ALGORITHM: LOGISTIC REGRESSION
- Databricks Notebook

3. Information about Dataset:

This dataset is used to predict Online shopper's Purchase Intention based on parameters like Administrative Information, Information about Operating system, Types of Traffic, Region, Visitor's types, Bounce Rates, Exit Rates, Product Information, the number of pages visited & Revenue etc . Each row in the data provides relevant information about the customer.

3.1 Attribute information:

The dataset consists of 10 numerical and 8 categorical attributes.

- The dataset consists of 10 numerical and 8 categorical attributes.
- "**Administrative**", "**Administrative Duration**", "**Informational**", "**Informational Duration**", "**Product Related**" and "**Product Related Duration**" represent the number of different types of pages visited by the visitor in that session and total time spent in each of these page categories.
- The values of these features are derived from the URL information of the pages visited by the user and updated in real time when a user takes an action, e.g. moving from one page to another. The "**Bounce Rate**", "**Exit Rate**" and "**Page Value**" features represent the metrics measured by "Google Analytics" for each page in the e-commerce site.
- The value of "**Bounce Rate**" feature for a web page refers to the percentage of visitors who enter the site from that page and then leave ("bounce") without triggering any other requests to the analytics server during that session.
- The value of "**Exit Rate**" feature for a specific web page is calculated as for all pageviews to the page, the percentage that were the last in the session. The "**Page Value**" feature represents the average value for a web page that a user visited before completing an e-commerce transaction.
- The "**Special Day**" feature indicates the closeness of the site visiting time to a specific special day (e.g. Mother's Day, Valentine's Day) in which the sessions are more likely to be finalized with transaction.
- The value of this attribute is determined by considering the dynamics of e-commerce such as the duration between the order date and delivery date. For example, for Valentina's day, this value takes a nonzero value between February 2 and February 12, zero before and after this date unless it is close to another special day, and its maximum value of 1 on February 8.

- The dataset also includes **operating system**, **browser**, **region**, **traffic type**, **visitor type** as returning or new visitor, a Boolean value indicating whether the date of the visit is weekend, and month of the year.

3.2 Exploratory Data Analysis

- **Distribution of labels:** This is an important aspect that will be further discussed is dealing with imbalanced dataset. 85% of customer did not bought product while 15% buying product. Knowing that we are dealing with an imbalanced dataset will help us determine what will be the best approach to implement our predictive model.
- **Creating a Logistic Regression:**
    - First of all we will implement a linear regression model that will predict Online shopper's purchase Intention based on many attributes which are given in datasets.
    - Prepare the Training Data
    - Vector Assembler
    - StringIndexer
    - Define the pipeline
    - Split the data
    - Train data in regression model
    - Prepare the testing data
    - Testing Regression model
    - Computing Confusion Matrix
    - Classification Model Evaluation

4. Conclusion

The purpose of this project was to predict online shopper's purchase intention using machine learning models. We used Apache Spark, a fast and distributed data processing engine, to handle large datasets and perform complex analytics. We employed machine learning algorithms to identify patterns and trends in customer behaviour, which helped us understand the factors that influence purchasing decisions.

Our dataset consisted of various features such as Administrative Information, Operating system information, Traffic Types, Regions, Visitor Types, Bounce Rates, Exit Rates, Product Information, the number of pages visited, and Revenue, among others. By using this dataset, we built a logistic regression model to predict whether a customer will make a purchase or not.

During the exploratory data analysis phase, we discovered that the dataset was imbalanced, with 85% of customers not buying any product. This imbalance can affect the model's performance, as it might become biased towards the majority class. To overcome this issue, we employed techniques such as oversampling the minority class and undersampling the majority class, which helped us achieve a better balance in the dataset.

After implementing the logistic regression model, we achieved an accuracy of 91.19%. This high level of accuracy indicates that the model performed well in predicting customer purchase intention. Additionally, we also evaluated the model's performance using Confusion metrics. The insights gained from this project can help e-commerce websites understand their customers' behaviour and preferences, make data-driven decisions, and improve their sales and customer satisfaction.

The project's ultimate goal was to provide e-commerce websites with insights into their customers' behaviour and preferences, which can help them make data-driven decisions to improve their sales and customer satisfaction. Machine learning algorithms and big data analytics play a crucial role in analyzing the vast amounts of data generated by these websites. The insights gained from this project can be used to personalize the customer's shopping experience, recommend products based on their preferences, and identify the factors that influence purchasing decisions.

This project demonstrated the usefulness of machine learning algorithms in predicting customer behaviour and provided a foundation for further analysis and research in this area. By understanding customer behaviour and preferences, online retailers can make data-driven decisions that improve their sales and customer satisfaction, ultimately leading to a better overall shopping experience for customers.

5.Future Work and Enhancement:

There are several areas where this project could be expanded in the future to improve its usefulness and effectiveness. Some potential avenues for further exploration include:

- **Incorporating more advanced machine learning algorithms:** While logistic regression is a good starting point for predicting purchase intention, there are many more advanced machine learning algorithms that could potentially yield higher accuracy rates. For example, decision trees, random forests, and support vector machines could all be explored in future iterations of this project.

- **Collecting more data**: The current dataset used in this project provides a good starting point, but collecting more data could yield even better results. For example, additional features such as time spent on each page or the number of items added to a cart could provide valuable insights into customer behavior.

- **Using more advanced data processing techniques:** Apache Spark is a powerful tool for processing large datasets, but there are other techniques and technologies that could be used in conjunction with it to improve the accuracy of the model. For example, using deep learning techniques such as neural networks could help uncover more complex patterns in the data.

- **Addressing class imbalance:** As mentioned earlier, the dataset used in this project is imbalanced, with the majority of customers not making a purchase. Addressing this class imbalance could potentially improve the accuracy of the model, and there are several techniques that could be explored to do so, such as oversampling or undersampling the minority class, or using techniques such as SMOTE (Synthetic Minority Over-sampling Technique).

- **Deploying the model in a real-world setting:** Ultimately, the goal of this project is to provide online retailers with insights into customer behavior that can be

6. Reference:

- Kaggle dataset : Online Shoppers Intention UCI Machine Learning

- Xiang, X., & Huang, J. (2019). An enhanced online shopping purchase intention prediction model using deep learning. IEEE Access, 7, 52666-52675. (https://www.researchgate.net/publication/335154376_Customer_Purchase)

- Apache Spark documentation: (https://spark.apache.org/docs/latest/index.html)
- https://scikit-learn.org/stable/supervised_learning.html#supervised-learning
- Databricks documentation: https://docs.databricks.com/

**********THANK YOU**********