A logo for a university

Description automatically generated **East West University**

**Project Report**

**Semester:** Fall-2024

**Course Title:** **Statistics For Data Science** **Course Code:** **CSE303**

**Sec:** 07

**Group No: 07**

**Group Members:**

|  |  |
| --- | --- |
| **Student Name** | **Student Id** |
| **Farhan Ibtesham Joy** | **2022-3-60-150** |
| **Rijia Parveen Raya** | **2022-3-60-192** |
| **Sheikh Sarafat Hossain** | **2022-3-60-109** |

**GitHub Link:** [**https://github.com/shiekhsarafathossain/CSE303\_GroupProject**](https://github.com/shiekhsarafathossain/CSE303_GroupProject)

**Submitted to-**

**Puja Chakroborty**

Lecturer

Department of Computer Science & Engineering

East West University

**Date of Submission: 19-January-2025**

**A Dynamic analysis on ‘bike purchase’ dataset**

**Introduction:**

**Background**

With the age of modern data science, being able to predict bike sales is very important in order to reach your target audience. They help businesses grow sales by giving them a clue into what customers have done, like who bought the bikes and what regions contribute the most to sales. Furthermore, businesses are able to optimize their inventory, pricing, and marketing plans and staying informed on the market trends all based on study of the elements affecting sales.

**Motivation**

It’s necessary for any business for any of its reasons to understand what type of customers it has and so does visualize the data of the past customers to take an analysis that if the customer is going to buy bike or not.

Businesses can predict accurately the demand and resource allocation, they can elevate the customer satisfaction and also make important improvements. The objectives of this project are to implement predictive model with a help of machine learning techniques.

**Problem Statement**

We want to analyze a bike sales dataset, preprocess the data, if required, by handling missing values, identify patterns, and creating a strong predictive model for sales, based on the features.

**Dataset:**

**Description**

The dataset “Bikesales Dashboard.csv”. The bike sales portion contains detailed sales information. There are several key fields such as Martial Status, Income, customer demographics, Occupation, Education and Customer Age.

**Analysis**

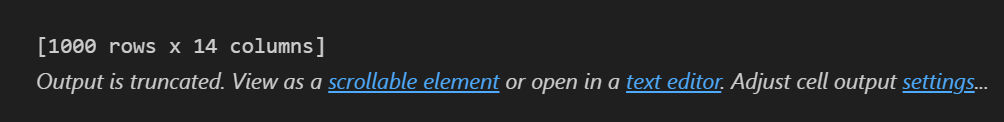
Initial exploration of the dataset reveals:

• Number of rows: 1000

• Number of columns: 14

• Data types: We found numeric and categorical fields with no missing values in our dataset.

A screenshot of a computer screen

Description automatically generated

**Train-Test Split**

The dataset is split into training and testing subsets in a standard 80:20 ratio. By doing it we ensure the effective model validation

**Methodology (In Steps):**

1. Data Preprocessing:

We take care of missing values, (if any), Encode categorical variables also Normalize numerical fields.

2. Exploratory Data Analysis (EDA):

We then visualize correlations among features and find important predictors.

3. Model Development:

We used Linear Regression and Logistic Regression

4. Model Evaluation:

We then match performance metrics of various models to one another.

**Performance Metric:**

In our dataset, we did mainly linear regression and logistic regression those are mainly thought as one of the key models for dataset analysis and machine learning. Also, we added some key points, like – heatmap generation, age and income comparison for bike purchasing and so on. as one of the key models for dataset analysis and machine learning. Also, we added some key points, like – heatmap generation, age and income comparison for bike purchasing and so on. Linear regression mainly gave datasets a pure and optimized learning process that eventually predicted our sales how it should go in the long run. Additionally, our dataset also concludes F1 score, confusion matrix and so on.

A graph of a bike

Description automatically generated with medium confidenceA yellow and blue circle with text

Description automatically generated

A graph of age vs. bike purchase

Description automatically generatedA graph of income and bike purchase

Description automatically generated

A screenshot of a graph

Description automatically generatedA green squares with white text

Description automatically generated

**Best Model Indication:**

Seeing the target variable is binary (Purchased Bike Yes/No) in this case, we can pick Logistic Regression as a better model. Logistic regression is an algorithm that is oriented towards classification, including probabilities, and nice boundaries, whereas linear regression is intended for continuous results and is bad at binary classification.

A diagram of a graph

Description automatically generated with medium confidence

**Conclusion and Future Directions:**

**Conclusion**

This project was successful at creating a predictive model of bike sales based on a dataset of 1,000 columns. We ran linear regression algorithms and went through on things like, F1 score, precision, recall, confusion matrix, heatmap and so on.

**Future Direction**

In future, we can collect more data (seasonal trends, product promotion) to be able to improve the prediction precision. Additionally, we can use other advanced algorithms to get more accuracy by using other advanced models. We may use all 4 models to get numerous variations in future.