**CSCE 5310: METHODS OF EMPERICAL ANALYSIS**

**Project Increment 1**

**Project Description:**

1. Project Title: **Used Cars Price Analysis for Fair Pricing and Decision Making**

Project Team:

1. Vikram Kumar Reddy Ayaluri - 11711384
2. Lekhitanand Bandi - 11727419
3. Akshadha Reddy Itikela - 11700379
4. Tharun Kuravadi Sathish Babu - 11659825
5. Sai Vignesh Rayal Rolla - 11569007
6. Goals and Objectives

* Motivation:

The used car market is a significant industry, and pricing plays a crucial role in attracting buyers and ensuring sellers receive fair value for their vehicles. Determining fair prices for used cars is essential for both buyers and sellers, as it affects financial decisions and market stability.

In this digital world, humans are mostly spent time using internet and technologies. Where humans are going through many different online platforms to sell their products. To earn good value for the product , the consumer visits the most well-known platforms such as social media marketing, advertisements, broadcasting or by consumer recommendations. So, there are many types of websites and social media platforms such as OLX, telegram groups, Instagram product promotions and many more. But the products must meet the customers trust with assurance that they are spending their allowance on the right product for the right price.

* Significance:

Fathalla et al. [1]. proposed the biggest concern is price prediction for used vehicles. Nowadays the automotive industries are rapidly increasing and most of these industries are releasing their products with new features incorporating and tries to attract the customers. So, In this way the previous models are left behind. Also, consumers are mostly focused on the models of the product that even in the future upcoming models are released the price drop for the current model should at least meet their minimum round price.

* Objectives:

Considering only the United States region, there is a rapid growth in second hand cars buy and sell marketing. [2] From this website, consider the visualization of time series graph which shows the average price increases for last twelve months. From interpreting the time series charts, the price range is rapidly decreasing from the October and November months in 2023 compared to year 2022. This has really a huge impact on the market and every owner tries to keep their existing car until the best market price is going to be presented for the model. So, the second-hand cars are not available in stock for the interested users to buy. But the market price changes continuously over time. For the next few months, the market price may increase. New models are going to be increased and existing models are going to be in the same level of average selling price for their model. So, the owners of the used cars may even sell their cars.

* Features:

The current idea focuses to use the different car brands names and based on the car brand’s average market’s selling price and year of the model launch, the model building can be done and predict the price of the reasonable value to be car sold using the current year details. The existing model takes the car age as input and predicts the price of the car which holds the price to sell in the market. So, our model helps the sellers and consumers, based on the car’s parts and engine condition. They can bid or buy the car. From the consumers side this model helps to take the crucial decision to buy the particular car, even it is second hand product based on the age.

**Increment 1**

1. **Related Work [Background]**

Price Prediction and Valuation:

* Regression Models: Machine learning algorithms such as regression to forecast used car prices, considering factors like mileage, age, brand, and more are often used by researchers.
* Market Analysis: Price trends, regional variations and seasonal fluctuations can be understood by analysis of historical data.

Consumer Behavior and Preferences:

* Feature Importance: Research aimed to identify influential factors (e.g., mileage, brand, color) in buyer decision-making and their relative significance.
* Regional Disparities: Comparative analysis across regions highlighted divergent preferences for specific car types or brands.

Market Dynamics and Economics:

* Supply-Demand Evaluation: Utilizing used car datasets contributed to understanding market dynamics, demand and supply and fluctuations.
* Feature Impact: Research explored how various features or conditions impact the resale value of cars.

Sustainability and Environmental Impact:

* Vehicle Emissions and Age: Studies leveraged datasets to examine the relationship between a car's age and its emissions, to make environmental policy considerations.

Business Strategies and Marketing:

* Management of Inventory: Businesses used these datasets to for optimization of inventory management by identifying fast-selling or high-demand car types.
* Tailored Marketing: Understanding consumer preferences assisted in developing targeted marketing strategies.

Predictive Maintenance and Risk Assessment:

* Maintenance Predictions: Analysis of historical data facilitated predictive maintenance models, forecasting maintenance needs and associated costs for different car models at different ages.
* Risk Analysis: Assessing the probability of car issues or repair needs based on previous patterns and data trends.

1. **Dataset**

Dataset Title: “Used-cars-catalog”

The used cars catalog dataset consists of car advertisements containing numerous categorical and numerical data attributes. The main aim of this dataset is to explore the market of used cars which can later be used to predict the price of the car.

The information in the dataset is collected through web scraping in Belarus, in western Europe, on December 2, 2019.

Data Characteristics:

* Feature Variety: This dataset encompasses both categorical and numerical attributes, providing comprehensive details about various aspects of cars in advertisements.
* Categorical Attributes: Details like Manufacturer\_name, Model\_name, Transmission, Color, Engine\_fuel, Engine\_type, Body\_type, State, Drivetrain, Location\_region, among others.
* Numeric Attributes: Include Odometer\_value, Year\_produced, Engine\_capacity, Price\_usd, Number\_of\_photos, Up\_counter, duration\_listed, along with multiple "Feature" attributes.

Dataset Purpose:

* Analytical & Predictive Potential: The dataset is well-suited for predictive modeling, aiming to comprehend and forecast factors influencing used car prices or the probability of an exchange.
* Market Analysis: It also serves as a valuable resource for analyzing market trends, exploring regional preferences, and examining aspects influencing the duration cars remain listed for sale.

Data Granularity:

* Individual Advertisement Level: The dataset provides specific details about each car advertisement, offering a detailed perspective on numerous attributes.
* Diverse Feature Set: Encompassing a broad spectrum of features, it offers in-depth insights into manufacturer specifics, technical details, pricing, and more.
* Temporal Dimension: Features such as Year\_produced and duration\_listed potentially unveil the age of cars and their average market listing duration, providing temporal insights.

1. **Detail design of Features:**

In the analysis of the used cars dataset, we focus on analytical methods and designing the features.

**Feature Engineering:**

* 1. Car Age Calculation:

To know the impact of the car’s age on its price, we extracted a new feature called ‘car\_age’. We usually calculate the car’s age by subtracting the ‘year\_produced’ from the current year(2023).  When it comes to determining a car's value, its age is crucial. Our goal was to make it easier to compare vehicles by normalizing this variable.

* 1. Normalization of Continuous Variables:

We have continuous variables such as mileage and price. To normalize these variables, we used min-max scaling. The data is transformed into a fixed range - typically [0, 1] - that helps compare variables with different scales. Data normalization neutralizes scale differences in data analysis and prevents variables with larger scales from dominating the model.

**Analytical Methods:**

* 1. Descriptive Statistical Analysis:
* The first thing we did was do a descriptive analysis to understand the basic characteristics of the data, including measures of central tendency (mean, median) and dispersion (standard deviation, range).
  1. Data Visualization Techniques:

Box plots helped us identify outliers and examine the distribution of prices across different transmission types.

* You can use line and bar plots to look at trends in car prices based on variables like color, fuel, and age.
* To create these visual representations, we used Matplotlib and Seaborn libraries.
* Visualizing these plots helped us understand the relationships between variables and prices.
  1. Comparative Analysis:
* We compared how different factors such as model, age impact the car price.
* We divided the data into these categories and performed a comparative analysis to determine how much the prices differ.

1. **Analysis**

EDA(Explanatory Data analysis):

* In this Analysis, The crucial step is importing the dataset and searching for best preprocessing techniques to implement on the data.
* Data preprocessing techniques are mostly implemented on the high dimensional datasets. Since, the analysis uses more than 38000 rows. Preprocessing technique is implemented to neglect any null or redundant values present in the data points to maintain the better accuracy of the model and performance.
* Using the visualization modules provided by the python the relationship between the independent and dependent variables are represented. The Matplotlib and seaborn are the visualization tools used to comprehend the attributes from the dataset.
* The boxplot is one of the visualization technique used to represent any uncertainty in data and gives the detailed understanding of outliers. Also, boxplot was distributed in percentile level. Which is robust to outliers.
* Heat map is used to display the relationship between the attributes of the dataset. Which gives the behavior of the data.
* Using the histograms, the majority of votes that consumers mostly interested to buy based on the body shape of the vehicle can be described. Similarly, the consumers drags attention to the most selling engine type can be compared due to the computation of feature which gives the highest frequency.
* Using the barplots, which are well known for comparing the relationship between the two attributes. In our case barplot used to describe cost difference for every brand.
* **Model selection:**
* Selecting the model depends on the behaviour of the data and characteristics of the data. The data depends on the engine type, shape, brand and more implicit features to be considered. Since, taking the Car age as a consideration to predict the price, linear regression model which is best suitable to perform statistical analysis on the data.
* **Linear regression model**: which is a predictive model, used to predict the future outcome based on the historical data values.
* To perform linear regression model, the data is splitted into train and test samples with the 80 and 20 percentage into consideration.
* Now fit the model using the training samples and obtained prediction samples are compared with the test samples using the metrices like MAE, RMSE, R2 score and mean square error.
* Next, Age of the car is computed based on the key attribute from the dataset. Where age of the car is derived as the difference between the current year and each car’s launched model’s year.
* Using visualization tools like bar-graphs to interpret the comparison between predicted and actual values.
* **Minmaxscaler:** which is a preprocessing metric used to allot a scalar value between 0 and 1. So that, which tries deepens the interpretation between the predicted and test variables. In this case It is mostly used for comparative analysis.
* **K-fold cross validation**: To validate the model, the training set score are being evaluated based on the number of iterable mentioned in the parameter for the linear regression. By default, If not specify any technique to validate which used **kfold** cross validation technique is used since the current model is predictive model.
* Based on the given age of the vehicle, the model retrieves the best optimal price to buy the vehicle in second-hand.
* **One-way Anova:** To understand if there is any significant difference among the group or if all the categories in the group are actually the same, we performed one-way anova on prices of each few types of engine capacity. We figured that at least one among the group is different from the rest of the group.

1. **Implementation**

* Data Preprocessing

Steps include handling missing values, splitting the dataset into training and testing sets, and scaling numeric features using StandardScaler.

* Model Training

Models are trained on the preprocessed data, and hyperparameter tuning is performed for Random Forest using GridSearchCV.

* Simulation

The linear regression algorithm is used to predict the estimated price based on the age of the car.

1. **Preliminary Results**

Visualization:

* Exploratory data analysis through visualization further enhanced our understanding of the dataset. Several important insights are drawn from the visualizations created with the data.
* As the age increases, the price of a car decreases. But, after certain years, high age also increases the price. Rarity of the cars might be the reason.
* From the box-plot of price vs transmission type of the cars, it can be drawn that automatic cars are relatively high priced when compared to manual cars.
* Brown coloured cars have higher average prices and green coloured cars have a lowest average price.
* Similarly, from the line graph engine\_fuel vs price of the car, we can say that hybrid-petrol cars are having the highest average.
* From a few bar graphs we have concluded that most of the cars are not exchangeable. Also, the majority of them run on gasoline. From the bar graph of frequency of type of the cars, we observed that sedan is the most common type of car.

Covariance and Correlation:

* Provided insights into the relationship between the features of the used cars catalog data, i.e., when one variable changes the other is also affected.
* Car age and price of the car have the highest correlation value among all the features.
* Also, duration listed and up-counter have the same level of correlation.

One-way Anova:

* Compared the means of two or more groups to see whether there's any statistical support for a significant difference in the corresponding population means.
* Demonstrated that certain types of engine capacities are not equal.
* From the analysis using python and SPSS the resulted p-value is ~0.001, which is less than alpha(=0.05). Therefore, we are rejecting the null hypothesis. So, There is at least one group that is significantly different, when it comes to pricing.

Linear Regression using Python and SPSS:

* Predicted target variable using an independent variable based on the linear relationship between the variables.
* The linear regression model, employed for predicting car prices based on the age of the car, produced varying results when implemented in Python and SPSS.
* In the Python model, the intercept and coefficient were determined to be -432.20 and 14554.80, respectively. Conversely, in the SPSS model, these values differed, with an intercept of -560.66 and a coefficient of 17912.035.
* It prompts further investigation into potential differences in algorithm implementations, variable scaling, or underlying assumptions.
* The linear regression model's performance metrics reveal a Mean Absolute Error (MAE) of approximately 2124.41, Mean Squared Error (MSE) of 8000299.84, Root Mean Squared Error (RMSE) of about 2828.48, and an R² Score of 0.5911. These metrics combine and indicate the model's predictive power, with room for potential refinement to enhance the accuracy.

1. Project Management

**Implementation status report:**

**Work Completed:**

* Data gathering, cleaning and preprocessing
  + Description: Finding a dataset that can be used to perform Empirical Analysis. Then cleaning the data set and preparing it for further analysis using preprocessing techniques.
  + Responsibility: Lekhitanand Bandi
  + Contribution: Data gathering, cleaning and preprocessing
* Exploratory Data Analysis
  + Description: Performing EDA operations to understand the data better.
  + Responsibility: Akshadha Reddy Itikela
  + Contribution: Exploratory Data Analysis
* Visualization
  + Description: Used appropriate graphs to represent the data, find patterns, draw insights and provide easy understanding.
  + Responsibility: Tharun Kuravadi Sathish Babu
  + Contribution: Performing visualization techniques.
* Linear Regression using Python and SPSS
  + Description: Build a linear regression model that can predict the price of a used car based on the age.
  + Responsibility: Sai Vignesh Rayal Rolla
  + Contribution: Building and validating the model using python and SPSS
* One-way Anova using Python
  + Description: Using python in built libraries to perform one-way anova
  + Responsibility: Vikram Kumar Reddy Ayaluri
  + Contribution: Implementing one-way Anova using Python

**Work to be Completed:**

* Implementation of independent t-test
  + Description: Performing Independent t-test on several categorical features to find out if there is any significant difference between the data.
  + Responsibility: Vikram Kumar Reddy Ayaluri
* Implementation of One-way Anova in SPSS
  + Description: Perform One-way Anova on certain types of engines in SPSS to check if there are any Significant differences among the engine types.
  + Responsibility: Lekhitanand Bandi, Akshadha Reddy Itikela
* Bootstrapping using SPSS
  + Description: Performing bootstrapping using SPSS
  + Responsibility: Tharun Kuravadi Sathish Babu, Sai Vignesh Rayal Rolla

Issues:

* There is a noticeable difference between the coefficient and intercept obtained using the SPSS and Python model. Further investigation will be done and potential differences in algorithm implementation will be done.

1. References

* Fathalla, A.; Salah, A.; Li, K.; Li, K.; Francesco, P. Deep end-to-end learning for price prediction of second-hand items. Knowl. Inf. Syst., vol. 62, pp. 4541–4568, 2020.Available [[CrossRef](http://www.cs.newpaltz.edu/~lik/publications/Ahmed-Fathalla-KIS-2020.pdf)]
* Available <https://www.kbb.com/car-advice/is-now-the-time-to-buy-sell-or-trade-in-a-used-car/>
* Lucija Bukvić et al. “Price Prediction and Classification of Used-Vehicles Using Supervised Machine Learning”. Faculty of Transport and Traffic Science, University of Zagreb, 10000 Zagreb, Croatia,vol 14,issue 24,pp. 17034, Dec-2022. Available <https://www.mdpi.com/2071-1050/14/24/17034> .
* Kanwal Noor.; Sadaqat Jan, “Vehicle Price Prediction System using ML Techniques”,CSE dept, Vol. 167,pp .9, June 2017. available [[ref](https://www.ijcaonline.org/archives/volume167/number9/noor-2017-ijca-914373.pdf)]