**Capstone Project**

**This Case Study has 3 (three) checkpoints defined in it.**

|  |  |  |
| --- | --- | --- |
| **Check Point Topics** | **Remarks** | **Max Marks** |
| * Data manipulation and Visualization using Python (30 marks) * Statistical Analysis and Exploratory Data Analysis (50 marks) | **Checkpoint 1** | **80** |
| * Visualization using Power-BI Dashboard (40 marks) * Model Building using ML algorithms (80 marks) | **Checkpoint 2** | **120** |
| Final Presentation and Viva (50 marks) | **Checkpoint 3** | **50** |

**Domain:**

Automobile Industry

**About:**

There is an automobile company XYZ from India which aspires to enter the Indian used car market by setting up their company locally to give competition to their counterparts.

**Challenges:**

They want to understand the factors affecting the pricing of cars in the market, since those may be very different from the new car market. Essentially, the company wants to know:

* Which variables are significant in predicting the price of a used car?
* How well do those variables describe the price of a car

Based on various market surveys, the consulting firm has gathered a large dataset of different types of used cars across the market.

**What is Expected?**

Being a data analyst, you must come up with a first-step document that lists the output of your exploratory analysis, any issues or problems you may see with data that need follow-up, and some basic descriptive analysis that you think highlights important outcomes/findings from the data. Based on your findings, the next level of analysis will be charted out.

Also, you need to build an appropriate predictive model for predicting the price of a used car. You can perform a comparative study of several predictive models with various approaches and give your inferences accordingly.

**Data Dictionary:**

|  |  |
| --- | --- |
| **Column Name** | **Description** |
| Sales\_ID | Sales ID |
| name | Name of the used car |
| year | Year of the car purchase |
| km\_driven | Total km driven |
| Region | Region where it is used |
| State or Province | State or Province where it is used |
| City | City where it is used |
| fuel | Fuel type |
| seller\_type | Who is selling the car |
| transmission | Transmission type of the car |
| owner | Owner type |
| mileage | Mileage of the car |
| engine | engine power |
| max\_power | max power |
| seats | Number of seats |
| sold | used car sold or not |
|  |  |
| **Target Column** | **Description** |
| selling\_price | Current selling price for a used car |

**Check Point 1**

**Task 1.1(Data Manipulation and Visualization using Python)**

Here are some indicative types of analysis you can perform. Please note that this is not an exhaustive list, you may add more.

Come up with appropriate results and visuals for the following:

1. Which variables are significant in predicting the price of a used car?
2. How well do those variables describe the price of a car?
3. Which brands are selling most?
4. Are there specific locations selling more?
5. Which factors are more important in deciding cars' selling prices? Ex. km driven or type of owner or fuel type?
6. Perform relevant hypothesis testing (t, chi-Square, Anova tests)

**TASK 1.2 (Exploratory Data Analysis & Statistical Analysis)**

Data Preparation/Analysis tasks include (but are not limited to) the following.

1. Descriptive statistics for both numerical and categorical and draw a few insights from them. (Univariate Analysis)
2. Bi- Variate Analysis and Multi-Variate Analysis
3. Missing values identification and treatment
4. Outlier analysis and treatment
5. Data scaling using min-max and/or Z-score normalization
6. Data transformation
7. Feature Engineering
8. Perform relevant hypothesis testing (t, chi-Square, Anova tests)

**Checkpoint 2**

**TASK 2.1 (Visualization using Power-BI Dashboard)**

**Connect the data with the Power BI desktop and perform Data Manipulation using Power Query Editor. Perform the below tasks in Power BI Desktop.**

1. Identify the Region where the selling price is high.
2. Which Transmission type is generating the highest selling price?
3. What is the average selling price by region?
4. Identify the top 5 States where the selling price is high.
5. Which transmission type vehicles were sold largely in number?
6. Which fuel type vehicle has the highest KM Driven?
7. Visualize Average Mileage type by Fuel type.
8. Identify the cities where the number of used cars selling price was highest.

**Recommendations:**

* As a data analyst, what are the approaches you suggest to the marketing team to identify the ideal target group to make the campaign successful? Recommended based on your analysis.

**NOTE:** Results and graphs must be backed with appropriate inferences and insights.

**TASK 2.2 (Model building using ML algorithms)**

1. Build an appropriate ML model/s on the data.
2. Compare various ML models with appropriate regularization and/or hyperparameter tuning.
3. Evaluate the performance of the model.
4. Identify the right metric to evaluate the performance of the model.
5. Identify issues and concerns on the given data and suggest the best technique/s to overcome the issues.

**Checkpoint 3**

Prepare a crisp Final presentation including all the Checkpoint achievements and appear for the Q&A session.

**The above three Checkpoints completes the Capstone Project**