**Integrated Capstone Project**

**This Case Study has three checkpoints defined in it.**

|  |  |  |
| --- | --- | --- |
| **Check Point Topics** | **Remarks** | **Max Marks** |
| 1.1 Data manipulation using Python  1.3 Statistical Analysis using Python | **Check point 1** | **25** |
| 2.1 Visualization using Python  2.2 Exploratory Data Analysis  2.3 - Model Building using ML algorithms | **Check Point 2** | **50** |
| 3.1 Data Analysis using Big Data Tools  3.2 Data Analysis on Cloud  3.3 Deployment of ML model using Flask/Django | **Check point 3** | **25** |

**Domain:**

Automotive Industry

**About:**

There is an automobile company XYZ from USA which aspires to enter the US 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 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 used car |

**Check Point 1**

**Task 1.1(Data Manipulation 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 for the following:
  1. Which variables are significant in predicting the price of a used car?
  2. How well 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 price? Ex. kms driven or type of owner or fuel type?

**Task 1.3 (Statistical Analysis using Python)**

* 1. Descriptive statistics for both numerical and categorical and draw few insights from them.
  2. Perform relevant hypothesis testing (t, chi-Square, Anova tests)

**Check point 2 (Visualization using Python, EDA, Model building using ML Algorithms)**

**TASK 2.1 (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 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 price? Ex. kms driven or type of owner or fuel type?

**TASK 2.2 (Exploratory Data Analysis)**

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

1. Univariate, Bi- Variate Analysis and Multi- Variate Analysis
2. Missing values identification and treatment
3. Outlier analysis and treatment
4. Data scaling using min-max and/or Z-score normalisation
5. Data transformation
6. Feature Engineering

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

**Task 2.4(Model building using ML algorithms)**

**Model Building:**

1. Build an appropriate ML model/s on the data.
2. Compare various ML models with appropriate regularization and/or hyper-parameter 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**

**Task 3.3 -Deployment of Models using Flask/Django**

Deploy the Machine Learning Model created in Task 2.4 using the Flask application.