**Integrated Capstone Project**

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

|  |  |  |
| --- | --- | --- |
| **Check Point Topics** | **Remarks** | **Max Marks** |
| 1.1 Data manipulation using Python ( 30 marks)  1.2 Analysis using SQL Queries (30 Marks)  1.3 Statistical Analysis using Python (40 Marks) | **Check point 1** | **100** |
| 2.1 Visualization using Python (20 marks)  2.2 Exploratory Data Analysis (40 marks)  2.3 Model Building using ML algorithms (40 marks) | **Check Point 2** | **100** |
| 3.1 Deployment of ML model using Flask (50 marks) | **Check point 3** | **50** |
| 4.1Final Presentation and Viva ( 50 marks) | **Check point 4** | **50** |

**Domain:**

Automobile Industry

**About:**

The XYZ automobile company of India wants to enter the business of selling used cars. They would like to analyse the business challenges before the venture.

**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\_Code | Code representing unique Region name |
| City\_Code | Code representing City |
| State\_Code | Code representing unique state name |
| Postal\_Code | Postal code of the area |
| 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 1.2 (SQL-Oracle) Preferred is Oracle Incase if you have account to SNOWFlake you can also perform these tasks in SNOWFlake**

**Stage 1:**

1. Generate Info of the cars which is of the type first owner and the year of car purchase is between 2016-2020 and the number of kms driven is less than 80,000
2. Generate Info of all the cars whose average mileage is around 25 kmpl and year of car purchase is between 2018-2020 which has minimum seating of 4-5 and fuel type is diesel.
3. Generate Info of all the cars which are not sold, and seller-type is individual or dealer and also which has been used for less than 60000 kms and year of car purchase is 2014-2020.
4. Generate Info of all the cars which are manual and automatic whose mileage ranges between 20-25kmpl approximately and also which is within these cities(Washington, New York City,Chicago,Los Angeles)
5. Generate Info of all the cars which belong to honda category could be either first owner and second owner and also fuel type is petrol and average mileage should be 25kmpl and which are not sold and seating arrangement should be minimum 4.

**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)

**Deliverables/Submission guidelines of Checkpoint 1**

1. You have to prepare a power point presentation with screenshots of outputs (10 -15 slides) for each check point
2. Mention Problem Statement and Your approach to the problems
3. You need to submit all the code files - Task 1.1 & 1.3
4. The code file(html file for Task 1.1 & 1.3) should contain the Batch Name and the group name, group members (One of the group member) at the top (in Jupyter Notebook).
5. All comments/inferences/insights/reasons for doing a particular tasks etc should be written as a ‘markdown text’, but **NOT** using a comment lines with # or ‘’’.
6. Submit the code file as HTML file format (you have an option in Jupyter Notebook to save the file as HTML).
7. Name of the file must be in the form of:

*BatchName\_FirstName\_SecondName.html*

1. Task 1.2 SQL code to be copied in the word doc
2. The presentation file should have the Batch name, group name, Project name, Group members, their responsibilities

**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 like (wherever applicable):

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

(Prepare the data by handling missing values, outlier analysis, data transformation and normalization.)

**Task 2.3(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.

**Deliverables/Submission guidelines of Checkpoint 2**

* You have to prepare a power point presentation with screenshots of outputs (10 -15 slides) for each check point
* Mention Problem Statement and Your approach to the problems
* You need to submit all the code files - Task 2.1 , 2.2 & 2.3
* Also, submit Jupyter code file in html format. The code file(html file for Task 2.1 , 2.2 & 2.3) should contain the Batch Name and the group name, group members (One of the group member) at the top (in Jupyter Notebook).
* All comments/inferences/insights/reasons for doing a particular tasks etc should be written as a ‘markdown text’, but NOT using a comment lines with # or ‘’’.
* Name of the file must be in the form of:

BatchName\_\_GroupNumber\_FirstName\_SecondName.html

**CheckPoint 3**

**Task 3 -Deployment of Models using Flask**

* Deploy the Machine Learning Model created in Task 2.3 using the Flask application.

**Deliverables/Submission guidelines of Checkpoint 3**

* You need to submit all the code files of Task 3. And share the screenshots ( in the image format) of UI which you developed on web in the ipynb file.

**CheckPoint 4**

**Task 4**

Prepare crisp Final presentation including all three Checkpoint achievements and appear for Q&A session

**Deliverables/Submission guidelines of Checkpoint 4**

1. You have to prepare a power point presentation with screenshots of outputs (10 -15 slides)

2. Submit the ppt.