**Integrated Capstone Project**

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

| **Check Point Topics** | **Remarks** | **Max Marks** |
| --- | --- | --- |
| 1.1 Data manipulation using Python ( 25 marks)  1.2 Analysis using SQL Queries (25 Marks)  1.3 Statistical Analysis using Python (25 Marks) | **Check point 1** | **75** |
| 2.1 Visualization using Python(20 marks)  2.2 Exploratory Data Analysis(40 marks)  2.3Visualization using Power-BI (25 marks)  2.4 - Model Building using ML algorithms (40 marks) | **Check Point 2** | **125** |
| 3.1 Data Analysis using Big Data Tools(35 marks)  3.2 Data Analysis on Cloud (35 marks)  3.3 Deployment of ML model using Flask (30 marks) | **Check point 3** | **100** |

**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 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?
  6. Perform relevant hypothesis testing (t, chi-Square, Anova tests)

**Task 1.2 (SQL-Oracle)**

**Stage 1:**

1. Construct and ER-Diagram for the above-mentioned Requirement
2. Construct Tables as per the ER-Diagram.
3. Identify the relationships between tables and use appropriate standards for the same where applicable
4. Insert the appropriate data into the identified tables from the sample dataset provided.

**Stage 2:**

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 and seating arrangement should be minimum 4.
6. Prepare the data by handling missing values, outlier analysis, data transformation and normalization.

**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, Visualization using Power-BI, 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

**TASK 2.3(Visualization using Power-BI)**

**Connect the data with 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 or Provinces 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 used cars where the selling price was highest.

**Recommendations:**

* As a data analyst, what are the approaches do you suggest 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 2.4(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 3.1 - Data Analysis using Big Data Tools**

1. Big Data technologies like HDFS, Hive and PySpark need to be used as the historical data increases in size. As part of this task the following activities need to be done.
2. ● Develop a PySpark application to load data Spark DataFrames and save it into Hive tables on a Hadoop cluster in an optimized format.
3. ● Perform profiling of the data through PySpark and ensure that it is migrated correctly whereever the source is an RDBMS
4. ● Write PySpark routines to cleanse the data, prepare the data to handle missing values, and the data transformations identified in task 1.1 again making sure that the data is written into Hive tables in an efficient format
5. ● If the predictive model identified in task 2.4 available in Spark MLlib then develop a PySpark application to implement and evaluate the ML model identified with appropriate metrics\
6. ● Ensure that the best practices are followed and the design & code use the features of Spark and take advantage thereof.

**Task 3.2 - Data Analysis on Cloud**

**AWS**

1. Move the Automobile Dataset CSV used in TASK 2.1 to AWS s3
2. Create a data pipeline to move the data from storage to data warehouse(Redshift). You are allowed to use another copy command as well to move the data from storage to data warehouse.
3. In AWS load the data to Athena load the above CSV file to Athena
4. Configure Data Lake and Athena for your data
5. Ensure you have the required privileges in Data Lake to access your table.
6. Create the dashboard in PowerBI
7. Create the dashboard in PowerBI for the Automotive Industry like the steps you performed in Task 2.3

**AZURE**

1. Move the Marketing Data to Azure Synapse Storage Gen1.
2. Create a serverless SQL pool to query the data from Storage gen1
3. Create a Linked service to PowerBI
4. Ensure you have sufficient privileges on Synapse to access the serverless sql pool.
5. Perform various analytics on PowerBI
6. Create the dashboard in PowerBI like Task 2.3

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

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