**Integrated Capstone Project**

**This Case Study has four check points defined in it.**

|  |  |  |
| --- | --- | --- |
| **Check Point Topics** | **Remarks** | **Max Marks** |
| 1.1 Data manipulation using Python ( 50 marks)  1.2 Analysis using SQL Queries (50 Marks) | **Checkpoint 1** | **100** |
| 2.1Visualization using Power-BI (50 marks)  2.2 Data Analysis using Big Data Tools(50 marks) | **Check Point 2** | **100** |
| 3.1 Data Analysis and ML Model Training and Deployment on Cloud (100 marks) | **Checkpoint 3** | **100** |
| 4.1Final Presentation and Viva( 50 marks) | **Check point 4** | **50** |

**Domain:**

Automobile Industry

**About:**

The XYZ automobile company in 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)**

**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 seating arrangement should be minimum 4.

**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
4. The code file(html file for Task 1.1) 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).

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
3. Upload all the deliverables in the UNext LMS

**Check point 2 ( Visualization using Power-BI)**

**TASK 2.1(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 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 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** **Data Analysis using Big Data Tools**

**What is Expected?**

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.

● Develop a PySpark application to load data Spark DataFrames and save it into Hive tables on a Hadoop cluster in an optimized format.

● Perform profiling of the data through PySpark and ensure that it is migrated correctly whereever the source is an RDBMS

● 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

● If the predictive model identified in Spark MLlib then develop a PySpark application to implement and evaluate the ML model identified with appropriate metrics\

● Ensure that the best practices are followed and the design & code use the features of Spark and take advantage thereof.

**Deliverables/Submission guidelines of Checkpoint 2**

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. Task 2.1
   * 1. PowerBI .pbix file to be submitted.
     2. Have all comments written properly in the .pbix file.
     3. The .pbix file should contain the Batch Name and the Group Number, Group member names at the top.

Task 2.2

* + Submit Jupyter code file in html format. The code file(html file for Task 2.2) should contain the Batch Name and the group name, group members (One of the group member) at the top (in Jupyter Notebook).
    1. 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 ‘’’.
    2. Submit the code file as HTML file format (you have an option in Jupyter Notebook to save the file as HTML).
    3. Name of the file must be in the form of:
    4. *BatchName\_\_GroupNumber\_FirstName\_SecondName.html*
* Put all Tasks 2.1 & 2.2 as zip file (Mentioning batch name, Group number and your name) and upload it on the LMS.

**CheckPoint 3**

**Task 3.1 - Data Analysis + ML Model Training and Deployment on Cloud**

**AWS**

1. Redshift to PowerBI Connectivity
2. Move the Datasets to AWS s3
3. Create Redshift Instance
4. Ensure you create required tables in Redshift
5. Create a data pipeline/copy command to move the data from storage to data warehouse(Redshift). You are allowed to use other copy commands as well to move the data from storage to data warehouse.
6. Connect the Redshift data to PowerBI
7. Perform the tasks mentioned in Task 2.1(Only 4-5 core reports)
8. Dynamodb to s3 bucket confgiure SNS notifications for any new records added in the Dynamodb
9. Transfer the AWS s3 data to AWS Quicksight perform the same analysis doe using powerBI(Any 5 core reports)
10. 50% of storage exceeds then cloud watch has to trigger the alaram
11. Write a Lambda function which logs in cloud trail about S3 file type and size.
12. Build appropriate ML model/s on the data using AWS Sagemaker , Identify the right metric to evaluate the performance of the model **and Deploy on AWS Sagemaker.**

**AZURE**

1. Azure Synapse to PowerBI Connectivity
   1. Move the DataSet to Azure Synapse Storage Gen2
   2. Create a serverless SQL pool to query the data from Storage gen2
   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. Perform the tasks mentioned in Task 2.1(Only 4-5 core reports)
2. Enable Azure blob storage monitoring by adding sample data and upon processing if storage receives more than 20 bytes of data
3. Azure blob to azure data bricks using notebook options databricks to powerbi connectivity
4. Azure blob to Azure SQL copy option using datafactory and connect Azure SQL to Databricks
5. Write Azure functions to trigger to trigger when blob storage exceeds 20 bytes of data.
6. Build appropriate ML model/s on the data using Azure Machine Learning , Identify the right metric to evaluate the performance of the model **and Deploy on Azure Machine Learning**

**GCP**

1. BigQuery to PowerBI Connectivity
   1. Move the Datasets to Google Storage (Bucket)
   2. Create Bigquery Instance
   3. Ensure you create required tables in Bigquery
   4. Create a data pipeline/copy command to move the data from storage to data warehouse. You are allowed to use other copy commands as well to move the data from storage to data warehouse.
   5. Connect the BigQuery to PowerBI
2. Write Cloud Function by adding sample data in the cloud storage and upon processing if storage storage receives more than 20 bytes of data as inbound or outbound
3. Transfer the data from bucket to Looker and perform any 5 reports performed in the step 2.1
4. Configure GCP monitoring services when storage exceeds 20 bytes of data, notify using pub/sub.
5. Configure Google Big Query and enable monitoring services (Cloud Logging) for every record insertion or deletion.
6. Build appropriate ML model/s on the data using Google Big Query Models/Vertex AI , Identify the right metric to evaluate the performance of the model **and Deploy the model on GCP Machine Learning.**

**Deliverables/Submission guidelines of Checkpoint 3**

Task 3.1

Complete all the above tasks on your respective Cloud Platform allotted and for submission take screenshots of each task specified with step by step flow in a word document with proper caption mentioned along with your Batch/Group/Team member names convert as a PDF file and submit the PDF document on the LMS

**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**

* You have to prepare a power point presentation with screenshots of outputs (10 -15 slides)
* Submit the ppt.

The above four Checkpoints completes UNext Capstone Project