**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 + ML (Machine Learning) Model Training and Deployment on Cloud (100 Marks) | **Checkpoint 3** | **100** |
| 4.1 Final Presentation and Viva (50 marks) | **Check point 4** | **50** |

**Domain:**

Insurance Industry

**About:**

XYZ is a startup that provides insurance for cars. It is one of the best car insurance brands known for the highest claim settlement ratio which has acquired its initial policyholders by providing a hassle-free claim process, instant policy issuance, and claim settlements at minimum coverages. It is important for insurance companies to estimate the risk involved while covering a future customer. Predicting the chances of insurance claim for a future customer allows the company to avoid adding potentially loss making customers.

As it's a fast growing startup, the company would like to optimize the cost of the insurance by identifying the policyholders who are more likely to claim in the next 6 months.

**Challenges:**

It is a challenge for the company to identify the policyholders whose chances of filing a claim are high in the next 6 months.

Hence, given the dataset, the first step of this analysis is to assess what data is available and perform some exploratory and descriptive analytics to identify interesting and useful patterns, trends, and insights.

And the next step is to build a predictive model on the given data.

**What is Expected?**

Being a data analyst, you must come up with a first step document that lists 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 to predict if the policyholder will file a claim in the next 6 months or not based on the set of car and policy features. You can perform comparative study of several predictive models with various approaches and give your inferences accordingly.

.

**Data Dictionary:**

**Policy Features:**

* **policy\_id**: Unique identifier of the policyholder
* **policy\_tenure**: Time period of the policy
* **age\_of\_car**: Normalized age of the car in years
* **age\_of\_policyholder**: Normalized age of policyholder in years
* **area\_cluster**: Area cluster of the policyholder
* **population density**:Population density of the city (Policyholder City)

**Car features:**

* **make**: Encoded Manufacturer/company of the car
* **segment**: Segment of the car (A/ B1/ B2/ C1/ C2)
* **model**: Encoded name of the car
* **fuel\_type**: Type of fuel used by the car
* **max\_torque**: Maximum Torque generated by the car (Nm@rpm)
* **max\_power**: Maximum Power generated by the car (bhp@rpm)
* **engine\_type**: Type of engine used in the car
* **airbags**: Number of airbags installed in the car
* **is\_esc**: Boolean flag indicating whether Electronic Stability Control (ESC) is present in the car or not.
* **is\_adjustable\_steering**: Boolean flag indicating whether the steering wheel of the car is adjustable or not.
* **is\_tpms**: Boolean flag indicating whether Tyre Pressure Monitoring System (TPMS) is present in the car or not.
* **is\_parking\_sensors**: Boolean flag indicating whether parking sensors are present in the car or not.
* **is\_parking\_camera**: Boolean flag indicating whether the parking camera is present in the car or not.
* **rear\_brakes\_type**: Type of brakes used in the rear of the car
* **displacement**: Engine displacement of the car (cc)
* **cylinder**: Number of cylinders present in the engine of the car
* **transmission\_type**: Transmission type of the car
* **gear\_box**: Number of gears in the car
* **steering\_type**: Type of the power steering present in the car
* **turning\_radius**: The space a vehicle needs to make a certain turn (Meters)
* **length**: Length of the car (Millimetre)
* **width**: Width of the car (Millimetre)
* **height**: Height of the car (Millimetre)
* **gross\_weight**: The maximum allowable weight of the fully-loaded car, including passengers, cargo and equipment (Kg)
* **is\_front\_fog\_lights**: Boolean flag indicating whether front fog lights are available in the car or not.
* **is\_rear\_window\_wiper**: Boolean flag indicating whether the rear window wiper is available in the car or not.
* **is\_rear\_window\_washer**: Boolean flag indicating whether the rear window washer is available in the car or not.
* **is\_rear\_window\_defogger**: Boolean flag indicating whether rear window defogger is available in the car or not.
* **is\_brake\_assist**: Boolean flag indicating whether the brake assistance feature is available in the car or not.
* **is\_power\_door\_lock**: Boolean flag indicating whether a power door lock is available in the car or not.
* **is\_central\_locking**: Boolean flag indicating whether the central locking feature is available in the car or not.
* **is\_power\_steering**: Boolean flag indicating whether power steering is available in the car or not.
* **is\_driver\_seat\_height\_adjustable**: Boolean flag indicating whether the height of the driver seat is adjustable or not.
* **is\_day\_night\_rear\_view\_mirror**: Boolean flag indicating whether day & night rearview mirror is present in the car or not.
* **is\_ecw**: Boolean flag indicating whether Engine Check Warning (ECW) is available in the car or not.
* **is\_speed\_alert**: Boolean flag indicating whether the speed alert system is available in the car or not.
* **ncap\_rating**: Safety rating given by NCAP (out of 5)

**Target variable:**

* **is\_claim**: Outcome: Boolean flag indicating whether the policyholder file a claim in the next 6 months or not.

**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:
  + Analyse the policy features. Identify the pattern and figure out if there is any relation with the insurance claim.
  + Is the dataset balanced? Are there duplicates in the data?
  + Investigate the various car features and find out the important features that affects the insurance claim.

**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. Retrieve the average population density for policies with and without claims in different area clusters.
2. Retrieve the cars with a safety rating above 4, a gross weight less than 2000 kg, and a policy tenure of more than 5 years.
3. Retrieve car segments along with the count of policies for each segment, only considering segments with more than 100 policies.
4. Retrieve the policyholders with the highest policy tenure and their corresponding car make and model.
5. Retrieve the policy details for policies with the highest policy tenure in each cars age category.
6. Retrieve the top 3 car models with the highest average policy tenure among policies with a claim.

**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, Data Analysis using Big Data Tools)**

**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.**

**Recommendations:**

* As a data analyst, what are the approaches do you suggest the marketing team to identify ideal target group to make the campaign successful? Recommend 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 (Hadoop Distributed File System), 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 PySpark routines to load data from Spark DataFrames and save it into Hive tables in an optimized format on a Hadoop cluster, de-normalizing the data if required.
* Write PySpark applications to cleanse the data, prepare the data to handle missing values, and the data transformations identified in task 1.1, making sure that the data is written into Hive tables in an efficient format as well.
* Ensure that the best practices are followed, and the design & code use the features of Spark and take advantage of them.

**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 members) 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 line 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 datawarehouse(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. Dynamodb to s3 bucket confgiure SNS notifications for any new records added in the Dynamodb
8. Transfer the AWS s3 data to AWS Quicksight perform the same analysis doe using powerBI(Any 5 core reports)
9. 10 mb of storage of S3 exceeds then cloud watch has to trigger the alaram
10. Write a Lambda function which logs in cloud trail about S3 file type and size.
11. 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
2. Enable Azure blob storage monitoring by adding sample data and upon processing if storage receives more than 20 bytes of data
3. Move Raw data from Azure Storage to Azure Databricks, perform 2-3 visualization using PySpark command, transfer the databack to the same storage account in contaziner/blob.from new container to move the data to Azure SQL using copy command in data factory. Connect Azure SQL to PowerBI to perform the visualization.
4. Write Azure functions to trigger to trigger when blob storage exceeds 20 bytes of data.
5. 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.

1. 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
2. Transfer the data from bucket to Looker and perform any 5 reports.
3. Configure GCP monitoring services when storage exceeds 20 bytes of data, notify using pub/sub.
4. Configure Google Big Query and enable monitoring services (Cloud Logging) for every record insertion or deletion.
5. 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 4.1**

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**

**....................................................................................................**