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

**Domain:**

HR Analytics

**About:**

HRWorks Pvt Ltd is a Bangalore based start-up that commenced its operations in the summer of 2010. HRWorks was conceived by a team of HR practitioners.

HRWorks sees itself as the first true end-to-end Talent Acquisition Solutions organization which has the passion to bring together decades of experience in Technology Consulting and Talent Acquisition areas to usher in a paradigm shift in the way Talent Acquisition is practiced in today’s ultra-demanding business environment. HRWorks not only advises its customers on where their Talent Acquisition practices are, but also recommends and implements individually tailored, viable solutions using analytics.

Business process re-engineering with its three tenets − People Capability, Process Maturity and Technology Adoption − form the core ability of the company to provide customers with an enterprise-class customized solution to address their Talent Acquisition challenges. They bring in deep domain knowledge of how Talent Acquisition happens in corporates and provide viable recommendations to their customers.

**Challenges:**

Client service is all about the quality of the people involved in delivering business. However, one of the major challenges for HRWorks and its clients revolved around managing a quality workforce. Organizations spend tremendous amount of time and energy to create a homogenous environment where people thrive and succeed. Despite all the effort to keep an environment that is conducive, people leave organizations in search of better opportunities. In order to fill the vacuum, HR is bound to recruit new talent, thus forming a vicious circle in between attrition and recruitment; and to mitigate this, organizations keep trying to bridge the gap by strengthening their recruitment processes and creating a culture of inclusivity.

HRWorks wanted to find a unique solution which goes beyond the process aspect of human resource management. At first, HRWorks identified and prioritized the renege problem and put forward in a subtle way:

“A significant proportion of the candidates do not join the company that has made an offer. If we can identify them in advance, then companies don’t have to waste their resources.”

**What is Expected?**

HRWorks supports several information technology (IT) companies in India with their talent acquisition. One of the challenges they face is that about 30% of the candidates who accept the job offer do not join the company. This leads to huge loss of revenue and time as the companies initiate the recruitment process again to fill the workforce demand. HRWorks wants to find if a model can be built to predict the likelihood of a candidate joining the company.

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 for classifying joined and not joined for the offers released. You can perform comparative study of several predictive models with various approaches and give your inferences accordingly.

.

**Data Dictionary:**

* Sno: Sl number auto increment
* Candidate\_Ref: Candidate reference number
* DOJ\_Extended: Date of Joining of extended
* Duration\_to\_accept offer: Duration to accept the offer by candidate
* Notice\_Period: Notice period of previous employer
* Offered\_Band: E1 < E2 < E3 and so on
* Percent\_hike\_expected\_in\_CTC: expected hike by candidate
* Percent\_hike\_offered\_in\_CTC: hike offered by joining organisation
* Percent\_difference\_CTC: difference between expected and offered
* Joining\_Bonus: any joining bonus offered.
* Candidate\_relocate\_actual: relocating required or not
* Gender: Gender of candidate
* Candidate\_Source: How candidate applied or reached
* Rex\_in\_Yrs: years of exp
* Location\_ID: Code for current location of organisation
* Postal\_Code : Postal Code of the area of organisation
* Domicile\_Id: Code for home state of the candidate
* Region\_Code: Code representing unique Region name
* Age: Age of candidate
* Date\_of\_Interview : Date on which Interview was conducted.
* LOB\_id: Unique id for Line of Business(LOB)
* LOB: Line of business

**Target variable:**

* Status: joined 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:
  + Analysis of percentage joined of offer released.
  + What are the key drivers that influence the candidate joining/not joining a company?
  + Are there specific locations where candidates are not joining?
  + Joining status depends on the duration to accept an offer?
  + Hike offered has an impact on joining status.?

**Task 1.2 (SQL-Oracle)**

**Stage 1:**

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

**Stage 2:**

* Generate Info those candidates who have accepted offer and joining time is less then 30 days and candidates who are ready to re-locate.
* Generate Info those candidates who have accepted offer and also display list the candidates who have been offered and yet to accept the offer within 10 days;
* Generate Info those candidates who are willing to join and the ECTC is 25% hike from their CTC.
* Generate Info those candidates who are willing to join and the ECTC is 25% hike from their CTC joining time is less then 30 days and candidates who are ready to re-locate and joining bonus is offered.
* Generate the count of the candidates who are hired through what source and also who have joined and declined the offer.

**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.**

* Which gender is having the highest number of experience in the dataset?
* Which location has witnessed the highest number of joinees? How are the joinees compared across different locations?
* Identify the Gender taking the longest time to accept the offer from the company.
* Indicate the Gender earning the highest Average of percent hike in CTC.

**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