Case Study 1 – Rare Event Classification

Overview:

A real-world dataset is provided from a pulp-and-paper manufacturing industry. The dataset comes from a multivariate time series process. The data contains a rare event of paper break that commonly occurs in the industry. The data contains sensor readings at regular time-intervals (x's) and the event label (y). The task is to achieve acceptable levels of precision and recall metrics on a classification model.

Data:

The csv file “data.csv” has the following columns:

time – timestamp when the readings were taken

x1 → x61 - anonymized features

y – target

Problem statement:

Build a classification model, which can predict failures in the test data set with good precision and recall. The training cutoff date is ‘21-05-1999’

Expectations:

1. Process flow/ Approach
2. Data Science Technique used (Can use any technique)
3. Programming Tool used (Can use any tool)
4. Model performance
5. What are the main drivers and how they are impacting? State how business can use this information

Case Study 2– Product Recommender

Overview:

TechGiant is a B2B technology based company. It has about 17 Thousand customers all across the world. These customers are categorized in Large, Medium, Low and other variants of similar kind for the sales force to plan and execute an optimal customer targeting strategy. The company's product portfolio consists a vast 72 different types of products. However, the issue the sales force is facing, is the fact that they don't have accurate customer intelligence on which customer is going to purchase which product in the immediate future. They are seeking for a product recommender solution from the Analytics wing of TechGiant Company.

Data:

The excel sheet "Prod Recommendation Data to share.xlsx" has the following columns

Account ID – Customer unique identifier

Product Type - 72 different product types from P1 to P72

Account Size – Customers are categorized into different sizes based on their IT purchase behavior

FISC_QTR_VAL – Captures which fiscal quarter a particular customer has purchased a particular product in(Data has purchases from Financial Year (FY) 2015 Q1 to FY2020 Q1 – Total 21 Quarters)

Rev – Captures how much a customer spent in a given quarter on a particular product (in Dollars)

Problem statement:

Build a product recommender system, which can predict a customer's most likely list of top 3 product purchases in FY20 Q1. (Candidate can treat this as a supervised or an unsupervised machine learning problem)

Expectations:

1. Process flow/ Approach
2. Data Science Technique used (Can use any technique)
3. Programming Tool used (Can use any tool)
4. Model accuracy
5. How business can use this model? What is the benefit to them?
6. For large customers, which are the most popular products to recommend?
7. What are the best products the customer can transition to that are different from their current portfolio?

Case Study 3– Forecasting

Overview:

Demand forecasting is key business problem for most of the product-based organizations. It is very important from the perspective of achieving sales targets at the same time inventory management. Real-life data from technology industry is provided. It is time series data about the number of units purchased. Objective of this case study is to build a robust forecasting model to forecast demand for next 12 months with high accuracy.

Data:

The csv file “Forecasting_Case_Study_Data.csv” has monthly unit sales data with following columns:

Time – Fiscal month for which units data is recorded

Actuals → Units sold in respective fiscal month

Problem statement:

Build a forecasting model, which can forecast units for next 12 months with high accuracy

Given file has timeseries data for 42 months

Expectations:

1. Process flow/ Approach
2. Data Science Technique used (Can use any technique)
3. Programming Tool used (Can use any tool)
4. Model performance
5. What is the ideal time frame for which we are confident about the expected demand forecast?
6. What are the other drivers you can think of? (Make some assumptions)
7. What are the other approaches you would experiment with?

DELLTechnologies

Case Study 4

Building a Retrieval-Augmented Generation(RAG) Based Chatbot using arXiv papers as a Knowledge Base

Overview:

The explosive growth of Generative Artificial Intelligence (Generative AI) and Large Language Models (LLMs) represents a remarkable technological advancement that has reshaped various fields. The exponential growth of research papers and scientific literature in the field of Natural Language Processing demands efficient methods for retrieving and summarizing information. A chatbot that uses information downloaded from **arXiv** dataset as a knowledge base can help researchers quickly access and understand relevant papers .

Data:

Collection of metadata for 10000 papers related to Natural Language Processing domain downloaded from arxiv.

Fields: **Title Abstract Authors**

Problem statement:

The objective of this case study is to build a chatbot capable of understanding user queries and providing relevant answers from the arXiv dataset provided.

Expectations:

1. For exploring the data Topic Model or any suitable technique can be used
2. Large Language Models to be used (Any Suitable LLM)
3. Data Science Technique to be used for the chatbot (Retrieval Augmented Generation)
4. Programming Tool used (Python or any suitable language)
5. Chatbot performance
   - Relevance of the paper to the query
   - Quality of the answer as compared to the paper
6. Bonus Marks for checking Hallucinations by chatbot.