Department of Computer Science and Engineering (Data Science)

Subject: Applied Data Science (DJ19DSL703)

Experiment: 1

(Data Science Problem)

Name : Jhanvi Parekh
Sap:60009210033

Aim: Convert Business Problem into a Data Science Problem.

Theory:

In the industry customers from different domain have problem in terms of growth of company, increase the revenue, manage the limited resource, launch a new product etc. To develop a suitable technical solution for their business problem requires a thorough understanding of the system, breaking down the problems and mapping it into technical problems. The manager of the company who interacts with the customers need to understand
the business problem and convert it into a data science problem so that the data scientists can build appropriate model for the customers. An example is shown below:

Business Problem: Customer Churn Prediction

Data Science Problem: Customer Churn Prediction using Machine Learning

To convert this business problem into a data science problem, we need to frame it in terms of data and a specific objective:0

Objective: Develop a machine learning model that predicts customer churn to help the
Example: A telecommunications company is facing high customer churn rates, and they want to reduce the number of customers leaving their service. The company has collected data on customer demographics, usage patterns, customer service interactions, and churn history. They want to understand the factors that contribute to customer churn and build a predictive model to identify customers at risk of churning. By identifying these high-risk customers, they aim to implement targeted retention strategies and reduce churn.

telecommunications company identify high-risk customers and implement retention strategies.

Requirements:

1. Customer data: Demographic information (e.g., age, gender, location), contract details (e.g., contract type, duration), and account information (e.g., account age, payment method).
2. Usage data: Usage patterns (e.g., call duration, data usage, SMS usage) over a specific period.
3. Customer service data: Number of customer service calls, complaints, and resolutions.
4. Churn data: Whether each customer churned or not (target variable).

Data Science Steps:
Data Collection:

1. Gather the relevant data from the telecommunications company's database or data sources.
2. Data Preprocessing: Clean and prepare the data for analysis, handle missing values, and encode categorical variables.
3. Exploratory Data Analysis (EDA): Analyse the data to gain insights and understand relationships between features and churn.
4. Feature Engineering: Create new features or extract meaningful patterns from existing data that might improve the model's predictive power.
5. Model Selection: Choose appropriate machine learning algorithms for churn prediction (e.g., logistic regression, decision trees, random forests, gradient boosting).
6. Model Training: Split the data into training and testing sets and train the chosen machine learning model on the training data.
7. Model Evaluation: Evaluate the model's performance using metrics such as accuracy, precision, recall, F1-score, and ROC-AUC.
8. Hyperparameter Tuning: Optimize the model's hyperparameters to improve its performance.
9. Model Deployment: Deploy the trained model to make real-time churn predictions on new customer data.
10. Interpretability: Interpret the model's predictions and feature importance to understand the factors influencing churn.

Example Output: The data science solution will provide a predictive model that can determine the likelihood of each customer churning. The telecommunications company can then use this model to prioritize customer retention efforts and implement personalized strategies to retain high-risk customers, ultimately reducing overall churn rates and improving customer satisfaction.

Remember that this example is just a general outline, and the specifics of the data science process may vary based on the dataset and the business requirements.

Lab Assignment:

1. Choose any 5 industry problems. Describe in detail the Business Problem.
2. Write the objective of these business problems as data science problem.
3. List the requirements.
4. Mention the Data Science Steps.

1. Industry Problem: Smart Urban Traffic Management

Business Problem:

City traffic congestion is a growing issue, causing delays, increasing pollution, and reducing the quality of life for residents. Traditional traffic management systems are often reactive rather than proactive.

Data Science Problem:

Objective: Develop an intelligent traffic management system that uses real-time traffic data and predictive analytics to optimize traffic flow and reduce congestion in urban areas.

Requirements:

1. **Traffic Data:** Real-time data from traffic cameras, sensors, and GPS devices on traffic flow, vehicle counts, and congestion levels.
2. **Weather Data:** Weather conditions that might impact traffic patterns (e.g., rain, snow, fog).
3. **Event Data:** Information on scheduled events or construction that could affect traffic.
4. **Historical Traffic Data:** Historical records of traffic patterns and congestion levels.

Data Science Steps:

1. **Data Collection:** Gather real-time and historical traffic data from sensors and cameras, along with weather and event data.
2. **Data Preprocessing:** Clean data, handle missing values, and integrate multiple data sources.
3. **Exploratory Data Analysis (EDA):** Analyze traffic patterns, congestion hotspots, and the impact of weather and events.
4. **Feature Engineering:** Create features related to time of day, weather conditions, and event schedules.
5. **Model Selection:** Use machine learning models (e.g., time-series forecasting, reinforcement learning) for traffic prediction and control.
6. **Model Training:** Train models using historical traffic and weather data.
7. **Model Evaluation:** Evaluate the model's performance in terms of prediction accuracy and traffic optimization metrics.
8. **Hyperparameter Tuning:** Optimize model parameters to enhance performance.
9. **Model Deployment:** Implement the model in real-time traffic management systems to adjust traffic signals and manage congestion.
10. **Interpretability:** Analyze model outputs to understand how different factors influence traffic flow and congestion.

2. Industry Problem: Personalized Health Recommendations Using Wearable Devices

Business Problem:

With the rise of wearable health devices, users are generating a vast amount of health data. However, individuals often struggle to understand how to use this data for personalized health and wellness recommendations.

Data Science Problem:

Objective: Develop a system that analyzes data from wearable devices to provide personalized health and wellness recommendations, such as exercise routines, diet suggestions, and health alerts.

Requirements:

1. **Wearable Data:** Continuous health metrics from wearables, including heart rate, sleep patterns, activity levels, and calorie expenditure.
2. **User Profiles:** Demographic and health information, including age, weight, medical history, and fitness goals.
3. **Lifestyle Data:** Information on user lifestyle habits, such as dietary preferences and exercise routines.
4. **Health Data:** Historical health records, including medical conditions and previous recommendations.

Data Science Steps:

1. **Data Collection:** Aggregate data from wearable devices, user profiles, and health records.
2. **Data Preprocessing:** Clean and preprocess data, handle missing values, and normalize metrics.
3. **Exploratory Data Analysis (EDA):** Explore patterns and relationships in health data to understand user behavior and health needs.
4. **Feature Engineering:** Create features from wearable data related to activity levels, sleep quality, and health metrics.
5. **Model Selection:** Use recommendation algorithms (e.g., collaborative filtering, content-based filtering) and predictive models for personalized health suggestions.
6. **Model Training:** Train the recommendation system using user data and health metrics.
7. **Model Evaluation:** Evaluate the system's recommendations based on user feedback and health outcomes.
8. **Hyperparameter Tuning:** Optimize the recommendation model for better accuracy and relevance.

9. **Model Deployment:** Deploy the system on wearable devices or health apps for real-time recommendations.
10. **Interpretability:** Provide insights into how recommendations are generated and how different health metrics influence the advice.

---

3. Industry Problem: Autonomous Drone Navigation for Disaster Response

Business Problem:

During natural disasters, rapid and accurate assessment of affected areas is crucial. However, it is challenging to navigate drones effectively through dynamic and hazardous environments.

Data Science Problem:

Objective: Develop an autonomous drone navigation system using computer vision and machine learning to navigate through disaster zones, avoiding obstacles and identifying areas of interest.

Requirements:

1. **Drone Sensor Data:** Data from cameras, LiDAR, and other sensors for real-time environment mapping.
2. **Disaster Data:** Information on disaster zones, including damage reports and hazard maps.
3. **Navigation Data:** Historical navigation data from drones in similar environments.
4. **Weather Data:** Real-time weather conditions that may impact drone operations.

Data Science Steps:

1. **Data Collection:** Collect sensor data from drones, disaster zone maps, and weather conditions.
2. **Data Preprocessing:** Clean and preprocess sensor data, handle noise, and integrate data from multiple sources.
3. **Exploratory Data Analysis (EDA):** Analyze environmental data and obstacle types to understand navigation challenges.
4. **Feature Engineering:** Extract features related to obstacles, terrain, and environmental conditions from sensor data.
5. **Model Selection:** Use computer vision models (e.g., object detection, semantic segmentation) and reinforcement learning for autonomous navigation.
6. **Model Training:** Train models on annotated disaster zone data and simulated environments.
7. **Model Evaluation:** Evaluate navigation performance using metrics like obstacle avoidance success rate and route optimization.
8. **Hyperparameter Tuning:** Optimize model parameters for better navigation accuracy and safety.

9. **Model Deployment:** Deploy the system on drones for autonomous disaster response missions.
10. **Interpretability:** Analyze the decision-making process of the navigation system to understand how it handles obstacles and navigates through disaster zones.

---

4. Industry Problem: Dynamic Pricing for Event Ticket Sales

Business Problem:

Event organizers want to optimize ticket pricing dynamically to maximize revenue and attendance. Traditional pricing models often fail to adjust for real-time demand changes and competitor pricing.

Data Science Problem:

Objective: Develop a dynamic pricing model for event tickets that adjusts prices in real-time based on demand, competitor pricing, and market conditions.

Requirements:

1. **Ticket Sales Data:** Historical ticket sales data, including prices, sales volume, and timings.
2. **Demand Data:** Real-time data on customer interest, including search queries and social media mentions.
3. **Competitor Data:** Pricing and sales information from similar events or competitors.
4. **Event Data:** Information about the event, such as location, date, and expected attendance.

Data Science Steps:

1. **Data Collection:** Gather historical sales, demand, competitor, and event data.
2. **Data Preprocessing:** Clean and preprocess data, handle missing values, and normalize pricing information.
3. **Exploratory Data Analysis (EDA):** Analyze trends and patterns in ticket sales and demand fluctuations.
4. **Feature Engineering:** Create features representing demand indicators, competitor pricing, and event characteristics.
5. **Model Selection:** Use dynamic pricing models (e.g., time-series forecasting, demand prediction) and optimization algorithms.
6. **Model Training:** Train models on historical sales and demand data to predict optimal pricing.
7. **Model Evaluation:** Evaluate pricing strategies based on revenue maximization and attendance metrics.
8. **Hyperparameter Tuning:** Optimize model parameters to enhance pricing accuracy and responsiveness.

9. **Model Deployment:** Implement the dynamic pricing model in ticketing systems for real-time adjustments.
10. **Interpretability:** Analyze pricing decisions to understand factors influencing price changes and revenue outcomes.

---

5. Industry Problem: Ethical AI Monitoring for Bias Detection

Business Problem:

As AI systems become more widespread, ensuring that these systems operate fairly and without bias is crucial. There is a growing need for tools to monitor and evaluate AI models for ethical compliance.

Data Science Problem:

Objective: Develop a monitoring system that detects and mitigates biases in AI models by analyzing their predictions and ensuring they meet ethical standards.

Requirements:

1. **Model Data:** Predictions and decision-making data from AI models, including input features and output results.
2. **Ethical Standards:** Guidelines and criteria for identifying biases and ensuring fairness.
3. **Historical Data:** Data on previous model predictions and known instances of bias or unfair outcomes.
4. **Demographic Data:** Information on demographic groups affected by the model's decisions.

Data Science Steps:

1. **Data Collection:** Collect model predictions, ethical guidelines, and demographic data.
2. **Data Preprocessing:** Clean and preprocess data, handle missing values, and encode sensitive attributes.
3. **Exploratory Data Analysis (EDA):** Analyze prediction patterns and identify potential biases.
4. **Feature Engineering:** Create features to assess fairness and bias, such as group-specific error rates and disparity metrics.
5. **Model Selection:** Use fairness and bias detection algorithms (e.g., disparate impact analysis, fairness-aware learning) to evaluate AI models.
6. **Model Training:** Train models to detect and mitigate biases based on ethical guidelines.
7. **Model Evaluation:** Evaluate bias detection performance using fairness metrics and compliance with ethical standards.

8. **Hyperparameter Tuning:** Optimize parameters for better bias detection and mitigation.
9. **Model Deployment:** Implement the monitoring system in production environments to continuously assess AI model fairness.
10. **Interpretability:** Provide insights into how biases are detected and mitigated, and ensure compliance with ethical standards.