Shri Vile Parle Kelavani Mandal's
DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING
(Autonomous College Affiliated to the University of Mumbai)
NAAC Accredited with "A" Grade (CGPA : 3.18)

**Department of Computer Science and Engineering (Data Science)**

**Subject: Applied Data Science (DJ19DSL703)**

**Experiment: 1**

**(Data Science Problem)**

**Name:Rishabh Patil**
**Sapid:60009200056**
**Div:D12**

**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 through 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

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.

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:

Objective: Develop a machine learning model that predicts customer churn to help the 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)Inventory Optimization for Retailers:

**Business Problem**: A retail chain with multiple stores is struggling with inefficient inventory management. They face issues such as overstocking certain products while frequently running out of others. This results in lost sales due to out-of-stock items and ties up capital in excess inventory. The company wants to optimize its inventory by leveraging historical sales data, seasonality, and demand fluctuations to ensure the right products are available at the right stores and at the right times. This would help maximize sales, minimize carrying costs, and improve overall profitability.

**Objective:** Build a data-driven inventory optimization model that uses historical sales data, seasonality, and demand patterns to minimize overstock and out-of-stock situations, thereby maximizing sales and minimizing carrying costs for the retail chain.

**Requirements:**

- Historical sales data: Including product sales data for each store over time.
- Inventory data: Current stock levels, purchase orders, and supplier lead times.
- Seasonality information: Data on sales patterns, promotions, and events that impact demand.
- Supply chain data: Information on supplier performance and logistics

**Data Science Steps:**

- Data Collection:Gather historical sales data, inventory data, seasonality information, and supply chain data from various sources.
- Data Preprocessing:Clean and format the data, handle missing values, and standardize units of measurement.
- Exploratory Data Analysis (EDA):Analyze the data to understand sales trends, demand patterns, and inventory fluctuations.
- Feature Engineering:Create features such as demand forecasts, lead times, and reorder points to support inventory optimization.
- Model Selection:Choose appropriate forecasting models (e.g., time series models) for demand prediction and inventory optimization.
- Model Training:Split the data into training and validation sets and train the selected models.
- Model Evaluation:Evaluate model performance using metrics like Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE).
- Optimization:Implement inventory management policies based on the forecasting models and optimization algorithms.

## 2)Fraud Detection in Financial Services

**Business Problem**: A financial institution is facing a growing problem of fraudulent transactions, which lead to significant financial losses and erode customer trust. The company needs to implement a robust fraud detection system. They want to analyze transaction data, customer behavior, and transaction patterns to identify suspicious activities in real-time. By promptly detecting and blocking fraudulent transactions, they aim to reduce financial losses, enhance security, and maintain the confidence of their customers

**Objective:** Create a real-time fraud detection system that analyzes transaction data, customer behavior, and transaction patterns to identify and block suspicious activities, thus minimizing financial losses and enhancing security in the financial institution.

**Requirements:**

- Transaction data: Records of all financial transactions, including amounts, timestamps, and transaction types.
- Customer behavior data: User login times, IP addresses, and transaction history.
- Historical fraud cases: Data on previously identified fraudulent transactions for model training.

**Data Science Steps:**

- Data Collection:Gather transaction data, customer behavior data, and historical fraud cases.
- Data Preprocessing:Clean and preprocess transaction data, handle imbalanced datasets, and encode categorical variables.
- Exploratory Data Analysis (EDA):Explore transaction patterns and customer behavior to detect anomalies.
- Feature Engineering:Create features such as transaction frequency, transaction amounts, and customer behavior patterns.
- Model Selection:Choose appropriate anomaly detection or classification models (e.g., Isolation Forest, Random Forest) for fraud detection.
- Model Training:Split the data into training and testing sets and train the chosen fraud detection model.
- Model Evaluation:Evaluate model performance using metrics like precision, recall, F1-score, and Receiver Operating Characteristic (ROC) curves.
- Deployment:Deploy the fraud detection system for real-time monitoring of financial transactions.

## 3)Predictive Maintenance in Manufacturing:

**Business Problem**: A manufacturing facility is experiencing unexpected equipment breakdowns, causing production delays and high maintenance costs. To address this issue, the company aims to implement predictive maintenance. They want to utilize sensor data, machine learning models, and historical maintenance records to predict when equipment is likely to fail. By proactively scheduling maintenance based on these predictions, they can reduce downtime, lower repair costs, and improve overall production efficiency.

**Objective:** Develop a predictive maintenance model using sensor data and historical maintenance records to forecast equipment failures and schedule proactive maintenance, reducing downtime and maintenance costs in the manufacturing facility.

**Requirements:**

- Sensor data: Real-time sensor readings from manufacturing equipment.
- Maintenance records: Historical data on equipment failures, maintenance actions, and repair times.
- Equipment specifications: Technical details and specifications for the manufacturing machinery.

**Data Science Steps:**

- Data Collection:Collect real-time sensor data, historical maintenance records, and equipment specifications.
- Data Preprocessing:Clean and preprocess sensor data, handle outliers, and convert timestamps to a consistent format.
- Exploratory Data Analysis (EDA):Investigate sensor data to identify patterns leading to equipment failures.
- Feature Engineering:Extract relevant features from sensor data, such as rolling statistics and trend analysis.
- Model Selection:Choose appropriate machine learning models (e.g., regression, classification) for predicting equipment failures.
- Model Training:Split the data into training and testing sets and train the chosen predictive maintenance model.
- Model Evaluation:Evaluate model performance using metrics like accuracy, precision, recall, and F1-score.
- Implementation:Implement the model into the manufacturing process for real-time equipment failure prediction and maintenance scheduling.

## 4)Personalized Content Recommendations for Media Streaming:

**Business Problem**: A streaming platform is struggling to retain and engage its subscribers due to content discovery issues. Many users complain about not finding content that matches their preferences, leading to increased churn. The company wants to develop a personalized content recommendation system. They plan to analyze user viewing history, ratings, and demographic information to provide tailored content suggestions. By offering relevant recommendations, they aim to increase user engagement, extend subscription durations, and ultimately boost revenue.

**Objective:** Implement a content recommendation system that utilizes user viewing history, ratings, and demographic information to provide personalized content suggestions, increasing user engagement, extending subscription durations, and maximizing revenue for the streaming platform.

**Requirements:**

- User data: User profiles, including demographics (age, gender), preferences, and viewing history.
- Content metadata: Information about available content, including genre, ratings, and release dates.
- User ratings and reviews: Feedback and ratings provided by users for content.

**Data Science Steps:**

- Data Collection:Collect user data, content metadata, and user ratings and reviews.
- Data Preprocessing:Clean and preprocess user data, handle missing values, and transform user behavior data into meaningful features.
- Exploratory Data Analysis (EDA):Analyze user preferences and content popularity to understand user engagement.
- Feature Engineering:Create user profiles and content embeddings to improve recommendation quality.
- Model Selection:Choose recommendation algorithms (e.g., collaborative filtering, content-based filtering) for content recommendation.
- Model Training:Split the data into training and testing sets and train the chosen recommendation model.
- Model Evaluation:Evaluate recommendation model performance using metrics like Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE).
- Deployment:Implement the recommendation system in the streaming platform to provide personalized content suggestions.

## 5)Supply Chain Visibility for E-commerce:

**Business Problem**: An e-commerce company faces supply chain disruptions, leading to delayed deliveries, increased shipping costs, and customer dissatisfaction. The company wants to improve supply chain visibility by tracking the movement of products from manufacturers to customers. They aim to implement a system that uses real-time tracking data, weather forecasts, traffic information, and historical logistics data to predict potential disruptions. By proactively addressing supply chain issues, they hope to ensure on-time deliveries, reduce operational costs, and enhance customer satisfaction.

**Objective:** Develop a supply chain visibility solution that utilizes real-time tracking data, weather forecasts, traffic information, and historical logistics data to predict and mitigate potential disruptions, ensuring on-time deliveries, reducing operational costs, and improving customer satisfaction for the e-commerce company

**Requirements:**

- Real-time tracking data: Location and status of shipments, delivery times, and potential delays.
- Weather data: Weather forecasts and historical weather patterns in the regions of operation.
- Traffic data: Information on traffic conditions, road closures, and congestion.
- Historical logistics data: Past shipment routes, delivery times, and issues.

**Data Science Steps:**

- Data Collection:Collect real-time tracking data, weather data, traffic data, and historical logistics data.
- Data Preprocessing:Clean and preprocess tracking data, handle missing location information, and format timestamps.
- Exploratory Data Analysis (EDA):Analyze tracking data and historical logistics information to identify supply chain patterns and potential disruptions.
- Feature Engineering:Create features such as estimated delivery times, route optimization, and disruption risk scores.
- Model Selection:Choose appropriate models (e.g., predictive modeling, optimization algorithms) for supply chain visibility and disruption prediction.
- Model Training:Split the data into training and validation sets and train the chosen models.
- Model Evaluation:Evaluate model performance using metrics like accuracy, reliability, and timeliness of disruption predictions.
- Implementation:Implement the supply chain visibility solution to monitor shipments, predict disruptions, and optimize logistics operations.