**Data Analytics Lifecycle : Customer Churn Prediction**

**Understanding Customer Churn**

Customer churn analysis is crucial for businesses aiming to comprehend customer behavior and enhance retention strategies. Companies can develop predictive models that identify potential churn risks, enabling proactive measures to mitigate customer loss.

**What is Customer Churn?**

Customer churn represents the percentage of customers who discontinue using a company's services within a given timeframe. A high churn rate can negatively impact revenue and brand loyalty, requiring businesses to attract new customers to compensate for lost ones.

**Importance of Customer Churn Analysis**

Analyzing churn helps businesses detect underlying causes and anticipate potential losses. Predictive modeling facilitates proactive engagement with at-risk customers through personalized incentives and service adjustments. Machine learning empowers companies to analyze intricate data patterns, uncovering insights that may otherwise be overlooked.

**Factors Influencing Customer Churn**

Several elements contribute to customer churn, including:

1. **Poor Customer Service**: Unresolved issues or inefficient support can drive customers away.
2. **Lack of Engagement**: Customers disengaged from a brand are more likely to switch.
3. **Competitive Offers**: Attractive offers from rival companies can influence customer decisions.
4. **Product Quality**: Dissatisfaction with product performance can lead to churn.
5. **Price Sensitivity**: Perceived high costs compared to value received can result in customer loss.

A white pipes with a cross on a black background

AI-generated content may be incorrect.

**Types of Customer Churn**

Customer churn falls into two primary categories:

* A diagram of a customer journey

  AI-generated content may be incorrect.**Voluntary Churn**: When customers actively decide to leave due to dissatisfaction or better alternatives.
* **Involuntary Churn**: When customers unintentionally discontinue services due to factors such as payment failures.

**Phases of the Data Analytics Lifecycle**

**Discovery**

The discovery phase involves understanding the problem of customer attrition and defining business objectives. The goal is to analyze customer behavior and predict churn to improve retention strategies. Key questions include: What factors contribute to churn? What patterns can we identify? We also assess data availability, potential challenges, and define the success metrics for our model.

**Data Preparation**

During data preparation, customer-related datasets, including demographics, usage behavior, and transaction history, are collected and cleaned. This involves handling missing values, removing duplicates, and performing exploratory data analysis (EDA) to detect trends and outliers. Feature engineering is also conducted, such as creating new variables from existing data to improve model performance.

**Model Planning**

In this phase, different analytical approaches are considered to build an effective churn prediction model. Techniques such as logistic regression, decision trees, or ensemble learning methods are evaluated. The dataset is split into training and testing sets, and feature selection techniques help identify the most impactful variables. Statistical and visualization methods assist in understanding feature relationships.

**Model Building**

The selected machine learning model is implemented using Python libraries such as Scikit-learn and TensorFlow. The model is trained on historical customer data and fine-tuned using hyperparameter optimization. Various algorithms are tested, and performance metrics like accuracy, precision, recall, and F1-score are analyzed to select the best-performing model.

**Communicate Results**

After evaluating the model, findings are visualized using tools like Matplotlib and Seaborn to present key insights. Reports and dashboards are created to communicate churn trends, significant features influencing customer retention, and model predictions. Business stakeholders are provided with actionable recommendations based on the analysis.

**Operationalize**

The final phase involves deploying the churn prediction model into a real-world environment, integrating it into a company's CRM or decision-making system. Continuous monitoring ensures model performance remains accurate, and periodic retraining is conducted to adapt to changing customer behavior. The model helps businesses proactively engage at-risk customers to reduce churn.

**Problem Statement**  
The problem in this project is **Customer Churn Prediction**, which focuses on identifying customers who are likely to stop using a company's services. Churn is a critical issue for businesses, especially those operating on subscription-based or recurring revenue models. The goal is to analyze customer behavior, demographics, and transactional data to predict which customers are at risk of leaving. By understanding the factors influencing churn, businesses can take proactive measures, such as personalized offers, better customer support, or loyalty programs, to improve retention and reduce revenue loss.

**Research Questions**

What is the relationship between customer tenure and churn rates?

Does total spending influence churn likelihood?

How does contract type affect churn rates?

Do streaming services impact customer retention?

What role does payment method play in churn trends?

**Hypothesis**

**Null Hypothesis (H0):** No significant relationship exists between customer tenure and churn rate.

**Alternative Hypothesis (Ha):** A significant correlation exists between customer tenure and churn rate.

**Steps for Customer Churn Analysis with Python**

**Step 1: Setting up the Python Environmen**These libraries are needed for: Data manipulation (pandas, numpy),Visualization (matplotlib, seaborn),Machine Learning (scikit-learn).

**Step 2: Importing and Understanding the Dataset**

**Step 3: Exploring and Preparing the Data**

Before diving into analysis, we’ll first take a close look at the dataset to understand its structure. This involves checking for missing values, identifying inconsistencies, and ensuring that all data types are appropriate for further processing. Any necessary adjustments, such as filling in gaps or converting data types, will be made to ensure smooth analysisThe **TotalCharges** column includes some non-numeric values and missing data. To ensure consistency, we’ll convert it to a numeric format and handle any missing values by replacing them with the median, which helps maintain data balance without being affected by extreme values.

Since the **customerID** column doesn’t contribute to the prediction process, we’ll remove it from the dataset to streamline our analysis and focus on the more relevant features.

To prepare categorical columns like **Contract** and **PaymentMethod** for machine learning, we’ll convert them into dummy variables. This transformation turns the categories into numerical values, making them compatible with machine learning algorithms while preserving the information they contain.

**Step 4: Exploratory Data Analysis (EDA)**

Next, we’ll dive deeper into the dataset by visualizing the distribution of **churn**. This will give us a clearer picture of how customer churn is spread across the dataset, helping us understand any potential imbalances or patterns in the data.



**Step 5: Splitting the Data**

We’ve now separated the target variable (**Churn\_Yes**) from the features (all the other columns). After that, we split the dataset into training and testing sets to ensure we can train our model on one portion of the data and evaluate its performance on another, unseen set.

**Step 6: Scaling the Features**

To ensure the machine learning model performs at its best, we’ll scale the feature values so they all lie within a similar range. This helps prevent any single feature from dominating the model due to differing magnitudes, allowing for a more balanced and accurate prediction.

**Step 7: Model Training**

We’ll train the model using a **Random Forest Classifier** and generate predictions. This powerful ensemble learning method combines multiple decision trees to improve accuracy and handle complex data patterns effectively.



**Step 8: Assessing Model Performance**

To see how well the model is performing, we’ll take a look at key metrics like **accuracy**, the **confusion matrix**, and the **classification report**. These will help us understand how accurately the model is predicting customer churn and where it might need improvement.

**Step 9: Feature Importance**

We’ll plot the **feature importance** to identify which features have the greatest impact on predicting customer churn. This visualization will help us understand which variables the model is relying on most, giving us valuable insights into the factors influencing churn.

**Step 10: Real-Time Prediction Dashboard**

We’ve created an easy-to-use, interactive interface that lets users predict customer churn by entering values for different features. With sliders and dropdown menus, users can interact with the model and get instant predictions, making it a simple and engaging way to explore how different factors influence the outcomes.

To provide more context for the user’s input, we’ll visualize the distribution of **tenure** and **MonthlyCharges**. These visualizations will help users better understand how these features are distributed in the dataset, adding valuable insights when making predictions.



We’ve set up sliders and dropdown menus, allowing users to easily interact with the model. These controls give users the flexibility to adjust input values and see how different factors affect the churn prediction, making the experience more dynamic and engaging.



This will show probabilities for each class (churn and not churn). Predicting customer churn is essential for businesses to improve customer retention strategies. By leveraging Python, data analysis, and machine learning models, we can identify patterns and take proactive steps to reduce churn rates.

A diagram of a customer service

AI-generated content may be incorrect.

**Example 2**

**Problem Description and Objectives**

Customer churn is a significant concern for banks and financial institutions as it directly impacts profitability and long-term customer relationships. This project aims to predict whether a customer will leave the bank or remain, based on the well-known Kaggle dataset, "Bank Customer Churn Dataset."

The primary objectives of this study are:

* To analyze key factors contributing to customer churn.
* To build and evaluate predictive models using various machine learning algorithms.
* To improve prediction accuracy using feature engineering techniques.
* To identify the most effective algorithm for predicting customer churn.

**Methodology and Technology Used**

**Methodology**

1. Data Collection: The dataset is sourced from Kaggle.
2. Data Preprocessing:
   * Dropping irrelevant columns such as Customer ID and Surname.
   * Encoding categorical variables like Gender and Geography.
   * Scaling numerical features using MinMaxScaler.
3. Exploratory Data Analysis (EDA):
   * Visualizing relationships between features and customer churn.
   * Identifying imbalances in the dataset.
4. Feature Engineering:
   * Converting categorical variables to numerical values.
   * Normalizing and scaling features.
5. Model Selection and Training:
   * Logistic Regression
   * K-Nearest Neighbors (KNN)
   * Artificial Neural Network (ANN)
   * Random Forest Classifier
6. Model Evaluation:
   * Accuracy, Precision, Recall, F1-Score, and Confusion Matrix.
7. Comparative Analysis:
   * Identifying the best-performing model based on metrics.

**Technology Used**

* **Programming Language**: Python
* **Libraries**: Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn
* **Machine Learning Techniques**: Supervised Learning, Feature Engineering, Normalization, Hyperparameter Tuning

**Technical Implementation**

**Data Preprocessing**

* Removed irrelevant columns (CustomerId, Surname).
* Encoded categorical variables (Gender, Geography).
* Normalized numerical values using MinMaxScaler.
* Created dummy variables for Geography and dropped one to prevent multicollinearity.

**Model Training**

* Split data into 75% training and 25% testing.
* Implemented the following models:
  + **Logistic Regression**: Baseline model.
  + **KNN**: Distance-based classification.
  + **ANN**: Deep learning-based approach.
  + **Random Forest**: Ensemble learning for better accuracy.

**Performance Evaluation**

* Evaluated models based on accuracy, precision, recall, and F1-score.
* Observed that **Random Forest performed the best with 86% accuracy**.
* Random Forest demonstrated a balance in recall and precision, making it the most effective for imbalanced data.

**Achieved Results**

* **Random Forest outperformed other models**, achieving **86% accuracy**.
* **Feature engineering improved model performance** by encoding categorical variables and scaling numerical values.
* **Customer churn insights**: Key factors affecting churn included Age, Balance, NumOfProducts, and Geography.

**Dataset contains following columns :** CustomerId,Surname,CreditScore,Geography ( country of customer),Gender,Age,Tenure (Total number of years with bank),Balance, NumOfProducts( Number of products or services utilising from bank)**,**HasCrCard ( Utilising credit card or not ( Card - 1 , No card - 0 )),IsActiveMember( Metric defining the member on the basis of his/her transactions (Active - 1 , Inactive - 0 ))**,**EstimatedSalary

**Based on above columns we are predicting Exited column :**  
Exited - Left bank or not ( left bank - 1 , retained by bank - 0 )

**Step 1: Setting up the Python Environmen**



A graph showing a customer exit

AI-generated content may be incorrect.**Step 2: Loading and Exploring the Dataset**

Bank has the highest customers in France and many exited from France and Germany.



This data indicates that females are more likely to leave the bank than males. Additionally, individuals over the age of 45 tend to exit more frequently than those under 45. Furthermore, active members are more inclined to leave the bank compared to inactive members. People are more likely to leave the bank after one year, while the number of departures significantly decreases after nine years. Individuals with a salary between 100,000 and 150,000 tend to leave the bank more frequently.

A diagram of a number of blue and orange dots

AI-generated content may be incorrect.A graph of a customer exit

AI-generated content may be incorrect.

A graph of a customer exit

AI-generated content may be incorrect.A graph of a customer exit

AI-generated content may be incorrect.

A diagram of a number of blue and orange dots

AI-generated content may be incorrect.

**Step 3: Feature engineering**



**Step 4: Training and testing using Artificial Neural Networks , Logistic regression , K nearest neighbours algorithm and Random forest algorithm**



**Step 5 : Interactive Prediction**



**Conclusion :**

Based on the evaluation metrics, Random Forest outperforms ANN, KNN, and Logistic Regression in multiple key areas. It achieves the highest recall (48%) for the minority class (Class 1), significantly improving its ability to correctly classify positive instances compared to KNN (27%) and Logistic Regression (20%). Additionally, Random Forest maintains a strong overall accuracy of 86%, outperforming KNN and Logistic Regression (both at 81%) and performing similarly to ANN.

Another critical factor is the F1-score for Class 1, where Random Forest achieves 0.59, making it far more balanced than KNN (0.37) and Logistic Regression (0.31), demonstrating a better trade-off between precision and recall. This makes it a more suitable model for imbalanced datasets where capturing positive instances is essential.

Overall, the combination of higher recall, strong F1-scores, and improved generalization makes Random Forest the best-performing model in this scenario.

**A screenshot of a computer

AI-generated content may be incorrect.A screenshot of a computer

AI-generated content may be incorrect.**

**A screenshot of a computer

AI-generated content may be incorrect.A screenshot of a computer program

AI-generated content may be incorrect.**

**Code :** [**andetazeqiri/Data-Analytics-Lifecycle-Customer-Churn-Prediction**](https://github.com/andetazeqiri/Data-Analytics-Lifecycle-Customer-Churn-Prediction)