**Data Analytics Lifecycle : Customer Churn Prediction**

Customer churn analysis is essential for companies looking to understand why their customers leave and how they can retain them. With Python and machine learning, we can create a powerful predictive model to help businesses identify potential churn risks before they happen, allowing them to take preemptive actions.

**What is Customer Churn?**

Customer churn refers to the percentage of customers that stop doing business with a company over a specific period. Churn is a critical metric for businesses to track, as a high churn rate means that a company may be losing revenue or brand loyalty, and will likely need to acquire more customers to replace those who leave.

**Why is Customer Churn Analysis Important?**

Identifying factors that lead to customer churn and predicting potential churners can significantly impact a company’s bottom line. By predicting churn, companies can proactively reach out to at-risk customers, offer tailored incentives, or modify their service offerings to retain these customers. Using machine learning for churn prediction allows us to process complex data patterns and identify insights that might not be immediately obvious.

**Causes of Customer Churn**

Several factors can contribute to customer churn, including:

1. **Poor Customer Service**: Negative experiences with customer support can lead to dissatisfaction and eventual churn.
2. **Lack of Engagement**: Customers who do not feel engaged with a brand are more likely to leave.
3. **Competitive Offers**: Attractive offers from competitors can entice customers to switch.
4. **Product Quality**: If the product does not meet customer expectations, they may seek alternatives.
5. **Price Sensitivity**: Customers may leave if they perceive the price as too high compared to the value received.

A white pipes with a cross on a black background

AI-generated content may be incorrect.

**Understanding Customer Churn**

Customer churn can be categorized into two types: voluntary and involuntary. Voluntary churn occurs when customers choose to leave, often due to dissatisfaction with the product or service. Involuntary churn happens when customers are unable to continue their relationship with the company due to reasons such as payment failures or changes in circumstances.

A diagram of a customer journey

AI-generated content may be incorrect.

**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**

1. How does customer tenure relate to churn rates?
2. Is there a correlation between total charges and churn rates?
3. What is the impact of contract type on churn rates?
4. Do streaming services influence customer churn?
5. How does the choice of payment method impact churn rates?

**Hypothesis**

* **Null Hypothesis (H0)**: There is no significant relationship between customer tenure and churn rate.
* **Alternative Hypothesis (Ha)**: There is a significant relationship 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: Loading and Exploring the Dataset**



**Step 3: Data Inspection and Preprocessing**

We will inspect the dataset to understand its structure and perform necessary preprocessing steps like handling missing values and converting data types.



The TotalCharges column contains non-numeric values and missing data. We will convert it to numeric and replace any missing values with the median.



The customerID column is not relevant for prediction, so it will be dropped.



Categorical columns like Contract and PaymentMethod need to be converted to dummy variables to make them suitable for machine learning.



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

We will explore the dataset further by visualizing the distribution of churn in the dataset.



**Step 5: Splitting the Data**

We have separated the target variables (Churn\_Yes) and the features (all other columns), then split the data into training and testing sets.



**Step 6: Scaling the Features**

We will scale the feature values to bring them to a common scale, ensuring that the machine learning model performs optimally.



**Step 7: Model Training (Random Forest)**

We will use a Random Forest Classifier to train the model and make predictions.



**Step 8: Model Evaluation**

We will evaluate the model’s performance using accuracy, confusion matrix, and classification report.



**Step 9: Feature Importance**

We will plot the feature importance to understand which features most contribute to predicting customer churn.



**Step 10: Interactive Prediction Interface**

We have created an interactive widget-based user interface to predict customer churn based on user input for various features.



We will visualize additional context for the user’s input, such as the distribution of tenure and MonthlyCharges.



We have created sliders and dropdowns for users to interact with the model.



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 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.

**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)