CUSTOMER CHURN PREDICTION FOR SUBSCRIPTION SERVICE

DHINAHARAN.S, AP/AI&DS

DINESH A, KISHORE S, KRITHIKA R S, ARASU PANDIAN N

BACHELOR OF TECHNOLOGY – DEPARTMENT OF ARTIFICIAL INTELLIGENCE AND DATA SCIENCE (2ND YEAR)

SRI SHAKTHI INSTITUTE OF ENGINEERING AND TECHNOLOGY

(AUTONOMOUS) COIMBATORE – 641062

ABSTRACT :

Customer churn prediction is a critical task for subscription-based services to maintain long-term profitability and growth. This project aims to develop a predictive model to identify customers at risk of canceling their subscriptions. By analyzing customer behavior, transaction history, usage patterns, and demographic data, we seek to uncover key factors contributing to churn. Machine learning algorithms, such as logistic regression, decision trees, and ensemble methods, will be employed to build and evaluate the model's effectiveness. The goal is to provide actionable insights for customer retention strategies, enabling service providers to take proactive measures to enhance customer satisfaction and reduce churn rates. The project will also explore the potential of integrating advanced techniques like deep learning and natural language processing for refining predictions. Ultimately, the predictive model will help businesses improve customer engagement, optimize marketing efforts, and maximize lifetime value, ensuring sustained growth in competitive subscription markets..

INTRODUCTION

In today's highly competitive market, subscription-based businesses—spanning industries from entertainment and digital services to retail and SaaS—face a pressing challenge: customer churn. Churn refers to the rate at which customers discontinue their subscriptions, and it is a critical metric for measuring the long-term health of subscription models. High churn rates can undermine profitability, reduce customer lifetime value (CLV), and hinder business growth. Therefore, predicting which customers are likely to cancel their subscriptions before they actually do is vital for businesses seeking to implement effective retention strategies.

Customer churn prediction leverages data science and machine learning techniques to analyze past behaviors and identify patterns associated with customer attrition. With early insights into which customers are at risk, businesses can take proactive steps to retain them—through personalized interventions, targeted marketing campaigns, improved service offerings, or customized loyalty programs.

Importance of Churn Prediction

For subscription services, the cost of acquiring new customers often exceeds the cost of retaining existing ones. Thus, reducing churn is often a more efficient strategy for business growth. Churn prediction models help identify risk factors early, allowing businesses to intervene before customers make the decision to cancel. This can significantly reduce the costs associated with customer acquisition and increase overall customer lifetime value (CLV). Furthermore, effective churn prediction models contribute to strategic decision-making, customer segmentation, and the optimization of marketing and customer support efforts.

Data for Churn Prediction

The data used to predict churn typically includes customer demographics, transactional history, engagement metrics, customer service interactions, and usage patterns. For example, if a customer's subscription is dormant for a certain period or they exhibit declining usage, these could be indicative of a higher likelihood of churn. Additionally, behavioral factors such as customer complaints, response to promotional offers, or feedback collected through surveys may further help in identifying at-risk customers.

Methodologies in Churn Prediction

To predict churn, various machine learning techniques are commonly employed. These include:

- Logistic Regression: A simple, interpretable approach often used for binary classification tasks, such as predicting whether a customer will churn or not.

- Decision Trees: A non-linear method that models customer churn based on rules derived from data features, helping to identify important factors influencing churn.

- Random Forests and Gradient Boosting: Ensemble techniques that combine the results of multiple decision trees to improve accuracy and reduce overfitting.

- Support Vector Machines (SVM): A powerful method for high-dimensional data, often used for classification tasks in churn prediction.

- Neural Networks: Deep learning techniques that can capture complex patterns in large datasets, allowing for more nuanced predictions of churn risk.

The performance of these models is evaluated based on metrics such as accuracy, precision, recall, F1 score, and area under the ROC curve (AUC), which assess how well the model predicts churn and minimizes false positives and false negatives.

Challenges in Churn Prediction

Despite the potential of churn prediction models, there are challenges in building accurate systems. These challenges include data quality (missing or incomplete data), class imbalance (where churned customers are fewer than non-churned customers), feature selection, model interpretability, and the incorporation of real-time data. Moreover, customer behavior is dynamic, and factors influencing churn can change over time, requiring constant model updates and retraining.

Objective of the Project

The objective of this project is to develop an effective customer churn prediction model for a subscription-based service. By leveraging machine learning algorithms and customer data, the project will focus on:

1. Identifying key features: Understanding the data and selecting the most relevant features that impact churn.

2. Building a predictive model: Implementing various machine learning algorithms and evaluating their performance.

3. Providing actionable insights: Generating insights that can guide customer retention strategies and optimize marketing efforts.

Through this, the project aims to provide businesses with a tool to proactively address churn, improve customer retention, and ultimately drive revenue growth. In the long term, the predictive model can be refined and scaled to handle more complex datasets and address the dynamic nature of customer behavior in subscription services.

LITERATURE REVIEW :

Eric Siegel's paper provides a comprehensive introduction to predictive analytics, making it a valuable resource for those interested in churn prediction. Siegel explains how predictive analytics models can forecast individual behavior based on historical data, which is central to predicting customer churn. The paper covers a wide range of techniques, such as data mining, machine learning, and statistical modeling, and applies these to real-world scenarios, including churn prediction in telecommunications and finance.

Provost and Fawcett’s paper is a foundational text for understanding data science and its application to business contexts, including churn prediction. It provides a solid grounding in data-analytic thinking, with a focus on building and evaluating predictive models. The authors cover classification, decision trees, and clustering—techniques widely used in churn prediction models. Additionally, the paper explores customer segmentation and lifetime value analysis, both of which are crucial for understanding churn dynamics

1. Customer Relationship Management

This paper focuses on customer relationship management (CRM) and how companies can use CRM strategies to enhance customer retention, thus reducing churn. Kumar and Reinartz explore the factors that influence customer behavior, including satisfaction, loyalty, and perceived value. They also discuss CRM data analysis techniques that help companies build effective churn prediction models.

Andrew Ng's Machine Learning Yearning provides a practical guide to applying machine learning techniques, including how to build effective models for complex problems such as churn prediction. Although this book is not solely focused on churn prediction, it offers valuable insights into building robust machine learning models, improving their performance, and reducing errors. Ng’s book covers key considerations such as data quality, feature selection, and evaluation metrics, which are critical when developing churn prediction models.

Customer Analytics for Dummies by Jeff Sauro and Alex Burkett:

This paper provides a beginner-friendly introduction to customer analytics, with a specific focus on understanding customer Behaviour, segmentation, and churn prediction. Sauro and Burkett cover key analytical methods, including predictive modelling, cohort analysis, and survival analysis, that can be applied to churn prediction.

2.Competing on Analytics:

The New Science of Winning by Thomas H. Davenport and Jeanne G. Harris. This paper discusses key analytical techniques that can be applied to identify at-risk customers in subscription-based services. It emphasizes a structured approach to analytics maturity, helping businesses evolve from basic reporting to advanced predictive modelling

Marketing Analytics: This paper focuses on analytics to understand and predict customer behavior, with specific applications for churn prediction. It covers techniques such as logistic regression, decision trees, and clustering, all of which are essential for modeling churn. The book also emphasizes the importance of feature selection and engineering, particularly for high-dimensional customer data, which is common in subscription services.

Eyal’s Hooked delves into the psychology of consumer habits and how companies can create products that encourage long-term customer engagement. Though it’s not focused on churn prediction directly, it provides a valuable perspective on how to design subscription services that reduce churn by fostering habitual use.

This paper isn’t directly focused on churn prediction, it offers valuable insights into customer communication, engagement, and loyalty—key factors in reducing churn. Miller emphasizes creating a compelling brand narrative that aligns with customer needs, which can enhance customer retention by building emotional connections. For subscription-based services, understanding customer motivations and pain points can inform not only engagement strategies but also improve the design of churn prediction models by incorporating sentiment analysis and behavioral data.

Raschka and Mirjalili's Python Machine Learning is a technical guide that covers essential machine learning techniques, including supervised learning methods commonly used for churn prediction. The book provides hands-on examples, particularly in Python, making it ideal for practitioners aiming to build churn prediction models. The book also includes a discussion on deep learning architectures, which are gaining popularity for predicting complex patterns of customer behavior in subscription models.

PYTHON: AN OVERVIEW

Python is one of the most widely used programming languages in data science and machine learning due to its simplicity, flexibility, and the extensive set of libraries that support various data analysis and modeling tasks. For a project focused on customer churn prediction in subscription-based services, Python provides an ideal environment to develop and deploy machine learning models efficiently. Below, we will explore how Python is used in the context of this project, including key libraries, tools, and techniques that will enable the development of accurate and scalable churn prediction models.

Key Features of python

Pandas:

- **Purpose**: Data manipulation and analysis.
- **How it’s used**: Pandas provides powerful data structures like DataFrames that make it easy to load, clean, and manipulate data. It allows us to perform tasks like handling missing values, encoding categorical variables, aggregating data, and creating feature matrices for machine learning models.
- **Example**: Loading customer data from CSV files or databases, handling missing values, and performing exploratory data analysis

(EDA) to understand the distribution and relationships within the data.

2. **NumPy**:

- **Purpose**: Numerical computing and array manipulation.
- **How it's used**: NumPy is essential for handling arrays and performing efficient numerical operations. It supports vectorized operations, which are useful for speeding up calculations, such as computing statistical metrics or preparing input data for machine learning models.
- **Example**: Working with matrices or arrays of customer data (e.g., usage statistics) and performing mathematical operations for feature scaling or normalization.

Scikit-learn:

- **Purpose**: Machine learning and statistical modeling.
- **How it's used**: Scikit-learn is the cornerstone of most machine learning tasks in Python. It offers a wide range of algorithms for classification, regression, and clustering, along with tools for model evaluation, cross-validation, and feature selection.
- **Example**: Implementing churn prediction models using algorithms like logistic regression, decision trees, random forests, and gradient boosting. It also provides utilities to measure model performance (e.g., accuracy, precision, recall) and handle data splitting.

4 **Matplotlib & Seaborn**:

- **Purpose**: Data visualization.
- **How it's used**: Visualizing data helps in understanding the underlying patterns, distributions, and relationships between features that influence customer churn. Matplotlib and Seaborn are widely used for creating plots, histograms, scatter plots, and heatmaps, which can help in exploratory data analysis (EDA) and model diagnostics.
- **Example**: Visualizing the distribution of customer demographics, churn rates, and model performance metrics.

XGBoost & LightGBM:

- **Purpose**: Gradient boosting algorithms for classification and regression tasks.
- **How it's used**: XGBoost and LightGBM are popular gradient boosting libraries that offer high performance, scalability, and efficiency. These algorithms work particularly well for predictive modeling tasks, such as churn prediction, and can outperform other machine learning methods in many cases.
- **Example**: Implementing advanced boosting models like XGBoost for churn prediction, tuning hyperparameters, and using cross-validation for model optimization.

TensorFlow/Keras:

- **Purpose**: Deep learning frameworks.
- **How it's used**: In cases where more complex patterns in data need to be captured, deep learning techniques using TensorFlow or Keras may be applied. These libraries allow for the creation of neural networks that can model intricate

relationships between input features and churn behavior, especially when large datasets are involved.

- **Example**: Building a neural network model to predict churn based on customer interactions and behavioral features.

Model Evaluation and Tuning

Python also supports comprehensive evaluation and tuning of machine learning models. Tools like **GridSearchCV** and **RandomizedSearchCV** (from Scikit-learn) enable automated hyperparameter optimization. Cross-validation techniques such as **K-fold cross-validation** can help in validating the model's generalizability and robustness.

FlutterFlow and Flutter Integration

FlutterFlow is a visual app-building platform that enables developers to create Flutter applications through a drag-and-drop interface. It simplifies the process of building apps by providing a no-code environment, making Flutter accessible to both technical and non-technical users. FlutterFlow supports integration with Firebase, API connections, and custom code blocks, giving users the flexibility to add backend functionality and custom logic to their apps. FlutterFlow also generates Flutter code, which can be exported and modified in any code editor for further customization. This approach makes it an excellent choice for rapid prototyping and application development, particularly for teams working with Flutter.

Why Use Flutter for the Smart Classroom Management System?

Flutter's cross-platform capabilities, customizable widgets, and high performance make it ideal for building a comprehensive app like the Smart Classroom Management System. With Flutter, the app can be deployed across Android, iOS, and even web platforms, ensuring accessibility for students and teachers on various devices. Additionally, Flutter’s real-time hot reload feature will enable developers to test and iterate on features quickly, which is valuable for refining complex functionalities like automated attendance tracking, real-time notifications, and AI-driven student assistance.

METHODOLOGY:

Churn prediction is a critical task for subscription-based businesses (e.g., streaming platforms, SaaS, telecom, etc.) to anticipate customer attrition and take preventive actions. Here’s a step-by-step methodology for predicting churn

1. **Problem Definition**

   Objective: Predict whether a customer will churn (i.e., cancel their subscription) within a given time frame (e.g., next 30 days).

   Business Impact: Understanding churn allows the business to take proactive measures (e.g., targeted retention campaigns) to improve customer retention and reduce revenue loss.

2. **Data Collection:** Collect historical data on customer behavior, interactions, and subscription details. Some essential data sources include Customer Profile Information, Subscription Data, Customer Interactions, Payment History etc..

3. **Data Preprocessing:**

   Prepare the data for modelling:

   Data Cleaning: Remove or handle missing values, outliers, and duplicates.

Feature Engineering: Create new features that may better capture churn behaviour.

4. **Exploratory Data Analysis (EDA)**

Perform a thorough analysis of the data to understand patterns and relationships:

Churn Rate: Analyse the proportion of customers who churned over time.

Churn Characteristics: Identify patterns in churn behaviour.

5. **Feature Selection:**

Select the most relevant features for the prediction model:

Correlation and Mutual Information: Use statistical tests to assess which features are most correlated with churn.

Feature Importance: Use machine learning techniques like decision trees or random forests to rank feature importance.

Dimensionality Reduction: In cases with a high number of features, techniques like PCA (Principal Component Analysis) can reduce dimensionality.

A 7-step guide to developing customer churn prediction software

Step 1.
Conduct a thorough analysis to understand the churn

Step 2.
Define specific objectives for your churn prediction software

Step 3.
Collect and prepare data

Step 4.
Implement the core features and functionalities

Step 5.
Build and train the predictive model

Step 6.
Test the model on real users

Step 7.
Deploy and collect the feedback

Workflow :

- The workflow for a customer churn prediction project can be broken down into several stages, from understanding the problem and gathering data to model deployment and evaluation. Below is a detailed step-by-step workflow that outlines the typical tasks involved in the churn prediction process using Python:.

. Problem Understanding and Data Collection

Goal: Define the problem, understand business objectives, and gather relevant data.

- **Understand Business Objective**: Clearly define the problem—predicting whether a customer will churn (cancel their subscription) within a given time frame.
- **Identify Data Sources**: Identify and collect data relevant to customer behavior and characteristics (e.g., customer demographics, usage patterns, interactions with support, transaction history).
  - ○ Data could come from internal databases, CRM systems, user logs, etc.
- **Data Types**: Data can include structured (e.g., CSV, SQL databases) or unstructured (e.g., customer feedback, call center logs) formats.

---

2. Data Preprocessing and Cleaning

Goal: Clean the data to ensure it's ready for analysis and modeling.

- **Handle Missing Values**: Check for missing or null values in key variables, and decide how to handle them (e.g., imputation, removal, or leaving as-is).
- **Remove Duplicates**: Ensure no duplicate rows are present that could skew the analysis.
- **Handle Outliers**: Identify and handle outliers, especially in continuous numerical

variables (e.g., extreme values for age or usage).

- **Feature Engineering**: Create new features that might help the model. For instance:
  - Calculate **customer tenure** (time since signup).
  - Create **usage frequency** metrics (e.g., number of logins, number of interactions).
- **Data Transformation**:
  - **Scaling and Normalization**: Standardize or normalize numerical features (e.g., age, income, usage).
  - **Encoding Categorical Variables**: Convert categorical variables (e.g., gender, region) to numerical representations using techniques like **One-Hot Encoding** or **Label Encoding**

RESULT :

Churn Probability Scores: Each customer is assigned a probability score that indicates the likelihood of them churning within a specified period (e.g., next 30 days).

Customer Segmentation: Based on these probabilities, customers are grouped into different risk categories (e.g., high, medium, low). This segmentation allows for targeted interventions, such as offering discounts to high-risk customers or engaging with low-risk customers to maintain satisfaction.

Model Performance Metrics:

Accuracy: The overall correctness of predictions (though this is not always the best metric for imbalanced datasets).

Precision, Recall, F1-Score: These metrics provide a deeper understanding of how well the model identifies actual churners (recall) and how many of the predicted churners actually churn (precision).

ROC-AUC: Indicates the model's ability to distinguish between churners and non-churners at various decision thresholds.

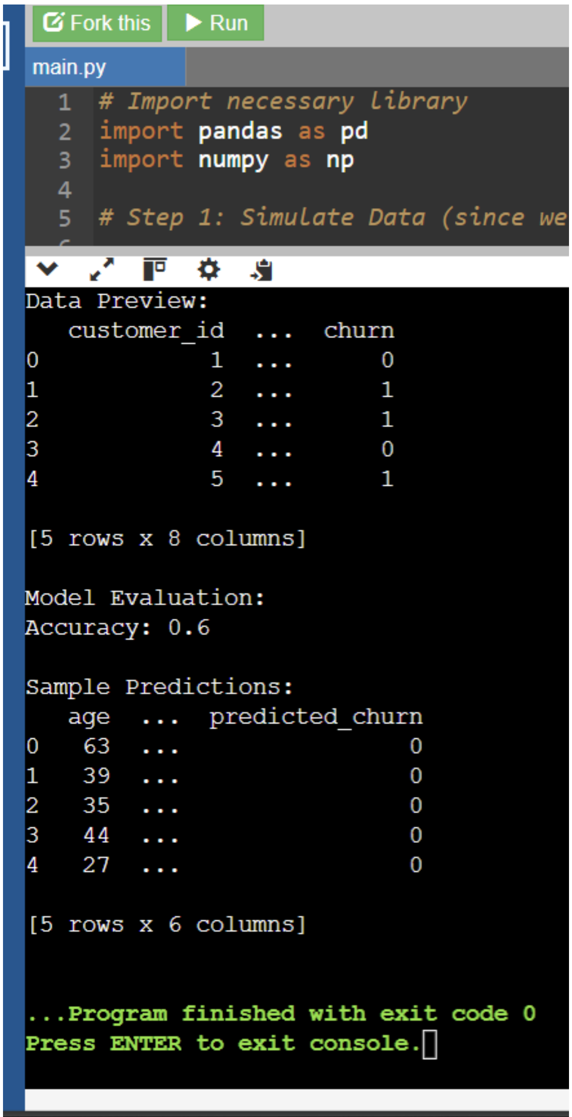
Insights for Retention Strategy:

Key Churn Drivers: Insights into which features (e.g., low usage, customer complaints, payment issues) most strongly correlate with churn.

Tailored Retention Campaigns: Based on the predicted churn, businesses can design personalized retention strategies (e.g., targeted offers or customer service outreach) to reduce churn and improve customer lifetime value.

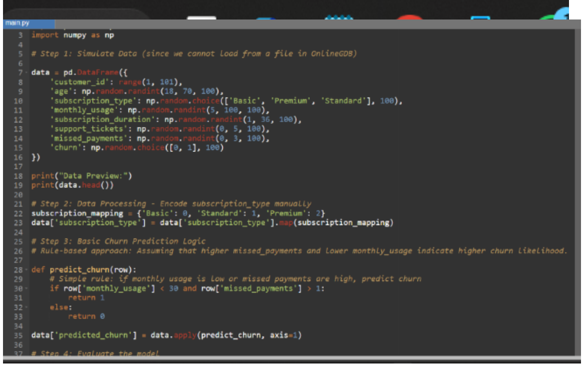
Business Impact: Ultimately, by predicting churn accurately, businesses can allocate resources more effectively, proactively intervene to reduce churn, and ultimately improve customer retention rates, leading to increased revenue and profitability.

OUTPUT:



```
# Import necessary library
import pandas as pd
import numpy as np

# Step 1: Simulate Data (since we
```

```
Data Preview:
   customer_id  ...  churn
0            1  ...      0
1            2  ...      1
2            3  ...      1
3            4  ...      0
4            5  ...      1

[5 rows x 8 columns]

Model Evaluation:
Accuracy: 0.6

Sample Predictions:
   age  ...  predicted_churn
0   63  ...                0
1   39  ...                0
2   35  ...                0
3   44  ...                0
4   27  ...                0

[5 rows x 6 columns]


...Program finished with exit code 0
Press ENTER to exit console.
```



```
import numpy as np

# Step 1: Simulate Data (since we cannot load from a file in OnlineGDB)

data = pd.DataFrame({
    'customer_id': range(1, 101),
    'age': np.random.randint(18, 70, 100),
    'subscription_type': np.random.choice(['Basic', 'Premium', 'Standard'], 100),
    'monthly_usage': np.random.randint(5, 100, 100),
    'subscription_duration': np.random.randint(1, 36, 100),
    'support_tickets': np.random.randint(0, 5, 100),
    'missed_payments': np.random.randint(0, 3, 100),
    'churn': np.random.choice([0, 1], 100)
})

print("Data Preview:")
print(data.head())

# Step 2: Data Processing - Encode subscription_type manually
subscription_mapping = {'Basic': 0, 'Standard': 1, 'Premium': 2}
data['subscription_type'] = data['subscription_type'].map(subscription_mapping)

# Step 3: Basic Churn Prediction Logic
# Rule-based approach: Assuming that higher missed_payments and lower monthly_usage indicate higher churn likelihood.

def predict_churn(row):
    # Simple rule: if monthly usage is low or missed payments are high, predict churn
    if row['monthly_usage'] < 30 and row['missed_payments'] > 1:
        return 1
    else:
        return 0

data['predicted_churn'] = data.apply(predict_churn, axis=1)

# Step 4: Evaluate the model
```

REFERENCE :

1. Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die by Eric Siegel - 2013

   Eric Siegel's book provides a comprehensive introduction to predictive analytics, making it a valuable resource for those interested in churn prediction.

2. Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking by Foster Provost and Tom Fawcett - 2013

   Provost and Fawcett’s book is a foundational text for understanding data science and its application to business contexts, including churn prediction.

3. Customer Relationship Management: Concept, Strategy, and Tools by V. Kumar and Werner Reinartz - 2012

   This book focuses on customer relationship management (CRM) and how companies can use CRM strategies to enhance customer retention, thus reducing churn

4. Machine Learning Yearning: Technical Strategy for AI Engineers in the Era of Deep Learning by Andrew Ng - 2018

   Andrew Ng's Machine Learning Yearning provides a practical guide to applying machine learning techniques, including

how to build effective models for complex problems such as churn prediction.

5. Customer Analytics for Dummies by Jeff Sauro and Alex Burkett - 2015

   This book provides a beginner-friendly introduction to customer analytics, with a specific focus on understanding customer behavior, segmentation, and churn prediction.
6. Competing on Analytics: The New Science of Winning by Thomas H. Davenport and Jeanne G. Harris

   This book discusses key analytical techniques that can be applied to identify at-risk customers in subscription-based services.

7. Marketing Analytics: A Practical Guide to Improving Consumer Insights Using Data Techniques by Mike Grigsby

   This book focuses on using data analytics to understand and predict customer behavior, with specific applications for churn prediction.

8. *Hooked: How to Build Habit-Forming Products* by Nir Eyal

   This book provides a valuable perspective on how to design subscription services that reduce churn by fostering habitual use.

9. Building a StoryBrand: Clarify Your Message So Customers Will Listen by Donald Miller

   This book isn't directly focused on churn prediction, it offers valuable insights into customer communication, engagement, and loyalty—key factors in reducing churn.

10. Python Machine Learning by Sebastian Raschka and Vahid Mirjalili