**NAAN MUDHALVAN-IBM DATA ANALYTICS WITH COGNOS**

# **PHASE 5:** CUSTOMER CHURN PREDICTION

**TEAM MEMBERS:**

Akshaya K 2021115011

Akshaya S M 2021115012

Preetham V 2021115077

Praveen M- 2021115075

Akash M 2021115324

ABSTRACT

Customer churn is a critical concern for businesses across various industries. Losing customers can be costly and detrimental to a company's long-term success. Predicting customer churn using artificial intelligence (AI) can provide valuable insights into customer behavior, enabling businesses to take proactive measures to retain their customers. This project aims to develop a customer churn prediction system using AI techniques.

PROBLEM STATEMENT

In today's highly competitive business world, retaining customers is highly important for sustainable growth and profitability. Customer churn poses a significant threat to businesses across various industries. To address this challenge effectively, there is a critical need for the development of a customer churn prediction system bringing in artificial intelligence (AI). The problem at hand revolves around the uncertainty of when and why customers decide to leave a business. Companies are seeking a proactive approach to predict and prevent customer churn by using the power of AI.

PROJECT DESIGN

A customer churn prediction project is a common and valuable application of data analytics, particularly in the fields of marketing and customer relationship management. Churn prediction involves identifying customers who are likely to stop doing business with a company in the near future. By identifying these customers early, a company can take proactive measures to retain them, which is often more cost-effective than acquiring new customers. Here's an overview of the key steps involved in a customer churn prediction project:

1. Understanding the Problem

Define the problem: Clearly state the objective of your project, which is to predict customer churn.

Understand the business context: Familiarize yourself with the industry, the company, and its specific customer churn challenges. This understanding will guide your analysis.

2. Data Collection

Gather data: Collect relevant data, such as customer information, transaction history, customer interactions, and any other data that might be indicative of churn.

3. Data Preprocessing

Data cleaning: Address missing values and outliers in the dataset.

Feature engineering: Create or modify features that may be relevant for predicting churn. For example, you might calculate metrics like customer lifetime value or churn propensity.

Data transformation: Encode categorical variables, scale numerical features, and perform other preprocessing steps.

4. Exploratory Data Analysis (EDA)

Conduct EDA to gain insights into the data. Explore the distribution of variables and relationships between features and the target variable (churn).

5. Data Splitting

Divide your dataset into training and testing sets to evaluate the model's performance accurately.

6. Model Selection

Choose appropriate machine learning or predictive modeling algorithms. Common choices include logistic regression, decision trees, random forests, support vector machines, and gradient boosting.

7. Model Training

Train your selected model(s) on the training data. Consider using techniques like cross-validation for hyperparameter tuning and model selection.

8. Model Evaluation

Assess the model's performance using appropriate evaluation metrics, such as accuracy, precision, recall, F1-score, and AUC-ROC. These metrics will help you understand how well your model predicts customer churn.

9. Model Interpretation

Interpret the model to understand which features are most influential in predicting churn. This insight can guide business decisions.

10. Deployment

If the model performs well, deploy it into a production environment, such as a customer relationship management (CRM) system, to make real-time predictions.

11. Monitoring and Iteration

Continuously monitor the model's performance in a production environment and update it as needed. Customer behavior and patterns may change over time.

12. Report and Communication

Summarize your findings, insights, and model performance in a clear and concise report. Present your results to relevant stakeholders in the organization.

Actionable Insights

Provide actionable recommendations based on your analysis. This could include strategies for retaining at-risk customers, personalized marketing, or other initiatives to reduce churn.

Model Selection

Model Selection is a crucial step in the process of building a customer churn prediction model. It involves choosing an appropriate algorithm or method to train a model that can effectively predict customer churn. Below, I'll discuss some commonly used models for churn prediction and justify their selection.

1. Logistic Regression

Justification: Logistic regression is a simple and interpretable model that is often a good starting point for churn prediction. It's a binary classification algorithm that models the probability of a customer churning. Here's why it's a good choice:

Interpretability: Logistic regression provides clear insights into the impact of each feature on the probability of churn. This interpretability is important for understanding the drivers of churn.

Simplicity: Logistic regression is relatively simple and computationally efficient, making it a good choice for quick prototyping and model building.

Works well with small to moderate-sized datasets: If you have a limited amount of data, logistic regression can still perform well without overfitting.

2. Decision Trees

Justification: Decision trees are another common choice for churn prediction. They are a non-linear model that can capture complex interactions between features.

Interpretability: Although not as interpretable as logistic regression, decision trees can still provide insights into the features that influence churn. You can easily visualize the tree structure.

Handles non-linear relationships: Decision trees can handle non-linear relationships between features and the target variable, which may be important in churn prediction where the factors affecting customer decisions can be complex.

3. Random Forests

Justification: Random forests are an ensemble learning technique that combines multiple decision trees, providing improved predictive performance and robustness.

Improved predictive performance: Random forests reduce overfitting compared to individual decision trees and generally provide better generalization to unseen data.

Feature importance: Random forests can measure feature importance, helping you understand which features are most relevant in predicting churn.

Robustness: They are less sensitive to outliers and noise in the data, which can be beneficial in real-world scenarios.

4. Gradient Boosting (e.g., XGBoost or LightGBM)

Justification: Gradient boosting algorithms like XGBoost or LightGBM have become increasingly popular for churn prediction due to their high predictive accuracy and versatility.

State-of-the-art performance: These algorithms often achieve top performance in various machine learning competitions and churn prediction challenges.

Handling imbalanced data: Churn prediction datasets are often imbalanced, with more non-churn cases. Gradient boosting can handle imbalanced data effectively by weighting classes or using sampling techniques.

Automatic feature selection: Gradient boosting models can automatically select important features, reducing the need for extensive feature engineering.

5. Support Vector Machines (SVM)

Justification: SVMs are a good choice when you want to find a hyperplane that maximizes the margin between churn and non-churn cases in a high-dimensional space.

Effective in high-dimensional spaces: Churn prediction often involves many features, and SVMs are effective in such scenarios.

Good generalization: SVMs aim to find the best separating hyperplane, which often results in good generalization to unseen data.

Flexibility in kernel selection: SVMs can use various kernel functions to capture complex relationships between features.

The choice of model should be driven by the specific characteristics of your dataset, your computing resources, and your objectives. It's often a good practice to start with simpler models like logistic regression and then progressively explore more complex models to see if they offer improved performance. Additionally, you can use techniques like cross-validation to compare model performance and select the one that works best for your churn prediction problem.

XGBoost (Extreme Gradient Boosting)

It is a popular and powerful choice for a customer churn prediction project for several compelling reasons. It has gained prominence in the machine learning community and is often a top choice for various predictive modeling tasks, including churn prediction. Here's why XGBoost is a better choice for this project:

High Predictive Performance:

XGBoost is known for its outstanding predictive performance. It consistently ranks among the top-performing algorithms in machine learning competitions and real-world applications. In the context of customer churn prediction, its ability to capture complex, non-linear relationships between features can lead to more accurate predictions.

Regularization Techniques:

XGBoost incorporates L1 (Lasso) and L2 (Ridge) regularization techniques, which help prevent overfitting. This is particularly useful when dealing with noisy or high-dimensional datasets common in churn prediction.

Handles Imbalanced Data:

Customer churn datasets often suffer from class imbalance, where the number of customers who don't churn far outweighs those who do. XGBoost can handle imbalanced datasets effectively by assigning different weights to classes or using sampling techniques.

Feature Importance:

XGBoost provides a feature importance score, allowing you to identify the most influential features in predicting churn. This is essential for understanding the drivers of churn and can inform business decisions.

Efficient Parallel and Distributed Computing:

XGBoost is optimized for efficiency and can take advantage of parallel and distributed computing. This is valuable when working with large datasets, as it can significantly reduce training time.

Flexibility in Loss Functions:

XGBoost supports a variety of loss functions, making it adaptable to different types of classification problems. You can select a loss function that aligns with the specific characteristics of your churn prediction problem.

Gradient Boosting:

XGBoost is a gradient boosting algorithm, which means it builds a series of decision trees sequentially, learning from the errors of previous trees. This approach often leads to improved model accuracy, as the model adjusts its predictions to focus on the most challenging cases.

VISUALIZATION

Data visualization is the use of graphical elements such as charts, graphs, and maps to represent data and information visually. The use of visualization tools provides an accessible way to see and understand trends, outliers, and patterns in data.

In machine learning (ML), visualization is a crucial tool for understanding data, model performance, and results.

It helps to gain insights, identify patterns, and communicate your findings effectively.

There are various techniques in data visualization. Few of them are described below:

Histograms

Histograms: Plot the frequency distribution of numerical variables to identify patterns and distributions.

Bar Charts

Bar Charts: Used for categorical data to show the frequency of different categories.

Pie Charts

Pie charts: Display the correlation between variables using pie circles and color gradients.

Word Clouds

Word Clouds: Word clouds are a visual representation of text data where words are displayed in varying sizes, with the size indicating the frequency of each word in the text. They are often used to quickly identify the most common words or themes in a large body of text.

Scatter Plots

Scatter Plots: Scatter plots are used to visualize the relationship between two numerical variables. Each data point is represented as a dot on the plot, and the position of the dot reflects the values of the two variables, making it easy to identify patterns, correlations, or outliers.

Line Charts

Line Charts: Line charts are used to display data points over a continuous interval, such as time. They are useful for showing trends and changes in data over time.

Heatmaps

Heatmaps: Heatmaps are graphical representations of data where values in a matrix are represented as colors. They are often used to visualize the correlation between variables, making it easy to identify relationships in large datasets.

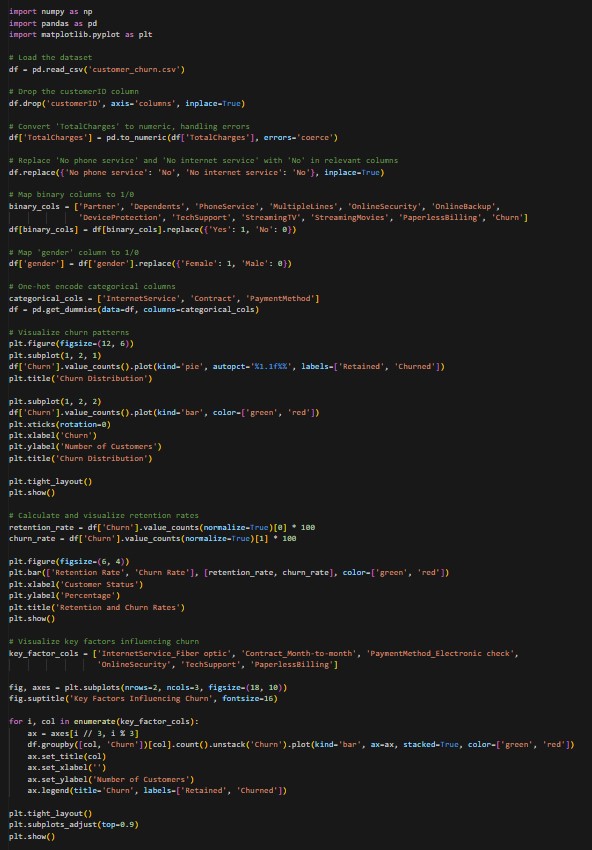
Box Plots

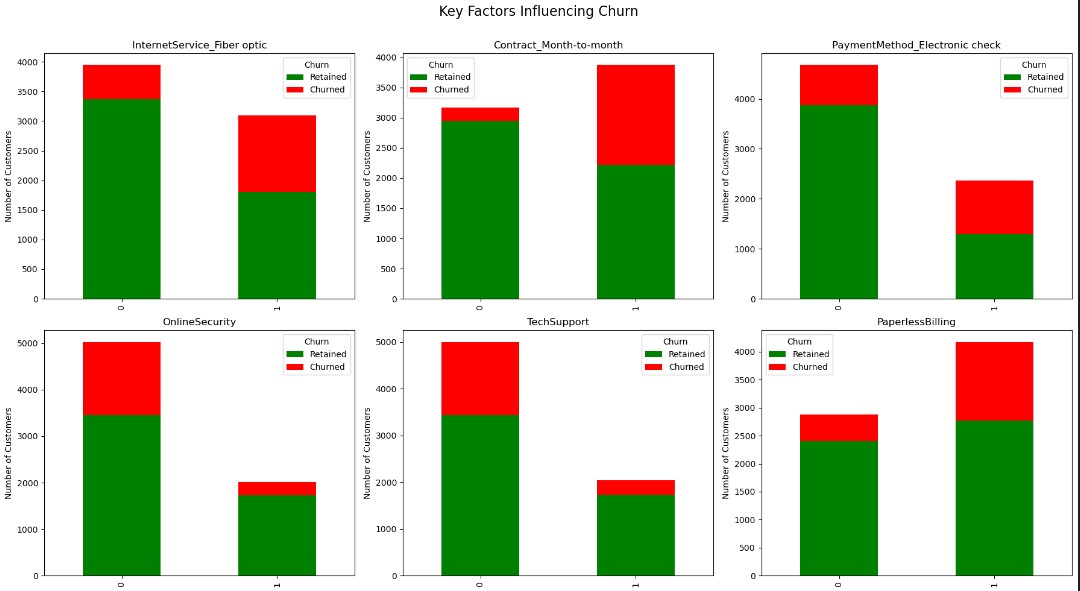
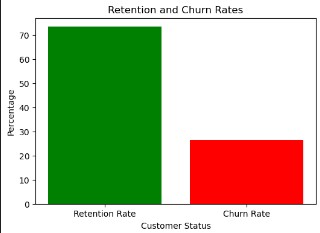
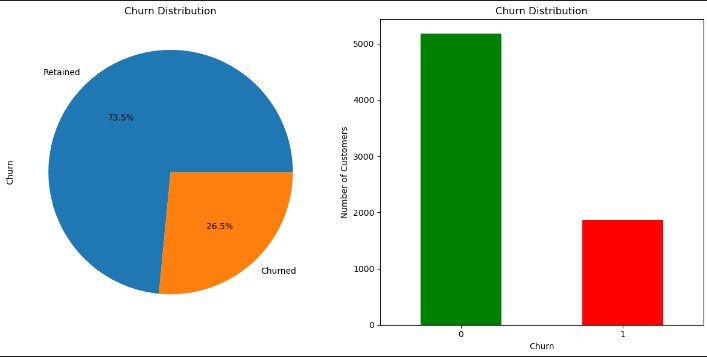
Box Plots: Box plots, also known as box-and-whisker plots, are used to display the distribution of numerical data and identify outliers, quartiles, and the median.

Violin Plots

Violin Plots: Violin plots are similar to box plots but also include a rotated kernel density plot on each side. They provide a more detailed view of the data distribution, especially for multimodal or complex distributions.

These visualization techniques can be valuable for exploring and presenting the results of a customer churn prediction project. For example, you can use scatter plots to visualize the relationship between customer lifetime value and the likelihood of churn or create heatmaps to explore the correlations between various customer attributes and churn. Additionally, you can use word clouds to display the most common reasons cited by customers for leaving, providing a quick overview of customer feedback. Data visualization not only aids in understanding the data but also in effectively communicating your findings to stakeholders within the organization.





ONCLUSION

CONCLUSION

In today's fiercely competitive business environment, churn prediction has become a critical tool for organizations. It allows them to take a proactive stance in addressing customer attrition by identifying and targeting customers at risk of leaving. This process harnesses the potential of data analytics to gain profound insights into customer behaviors and preferences.

The valuable insights obtained through churn prediction models empower businesses to make well-informed decisions and tailor their strategies to reduce churn rates. By adopting a data-driven approach, companies can allocate resources strategically to retain their most valuable customers, enhance customer satisfaction, and ensure sustainable long-term growth.

Churn prediction extends beyond the realm of technical machine learning models. It involves cross-functional collaboration, bringing together data scientists, marketing teams, customer support, and product development departments to implement effective customer retention strategies. Regular model updates and a feedback mechanism that collects input from departed customers are essential components of this process.

Moreover, early warning systems and customer segmentation are employed to detect and respond to indicators of customer disengagement promptly. These measures enable businesses to intervene in a timely manner and avert churn.

In summary, churn prediction is more than just a predictive analytics task; it represents a comprehensive approach to comprehending and engaging with customers. It empowers organizations to navigate the challenges associated with customer attrition, adapt to evolving market conditions, and prioritize customer satisfaction. By leveraging data and advanced analytics, businesses can achieve higher customer retention rates, reduce acquisition costs, and ultimately secure long-term success in their respective industries.