NAAN MUDHALVAN-IBM DATA ANALYTICS WITH COGNOS

PROJECT PHASE 5: PROJECT DOCUMENTATION AND SUBMISSION

PROJECT TITLE:

CUSTOMER CHURN PREDICTION

PROVIDED DATASET:

https://www.kaggle.com/datasets/blastchar/telco-customer-churn

TEAM MEMBERS:

ANANYA K A – 2021115013

ANULATHA S K - 2021115014

ARADHYA M - 2021115015

SAKTHIVEL K – 2021115088

BALAJI P - 2021115323

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:

1. 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.

2. 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.

3. 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.

4. 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.

5. 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.

6. 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.

7. 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

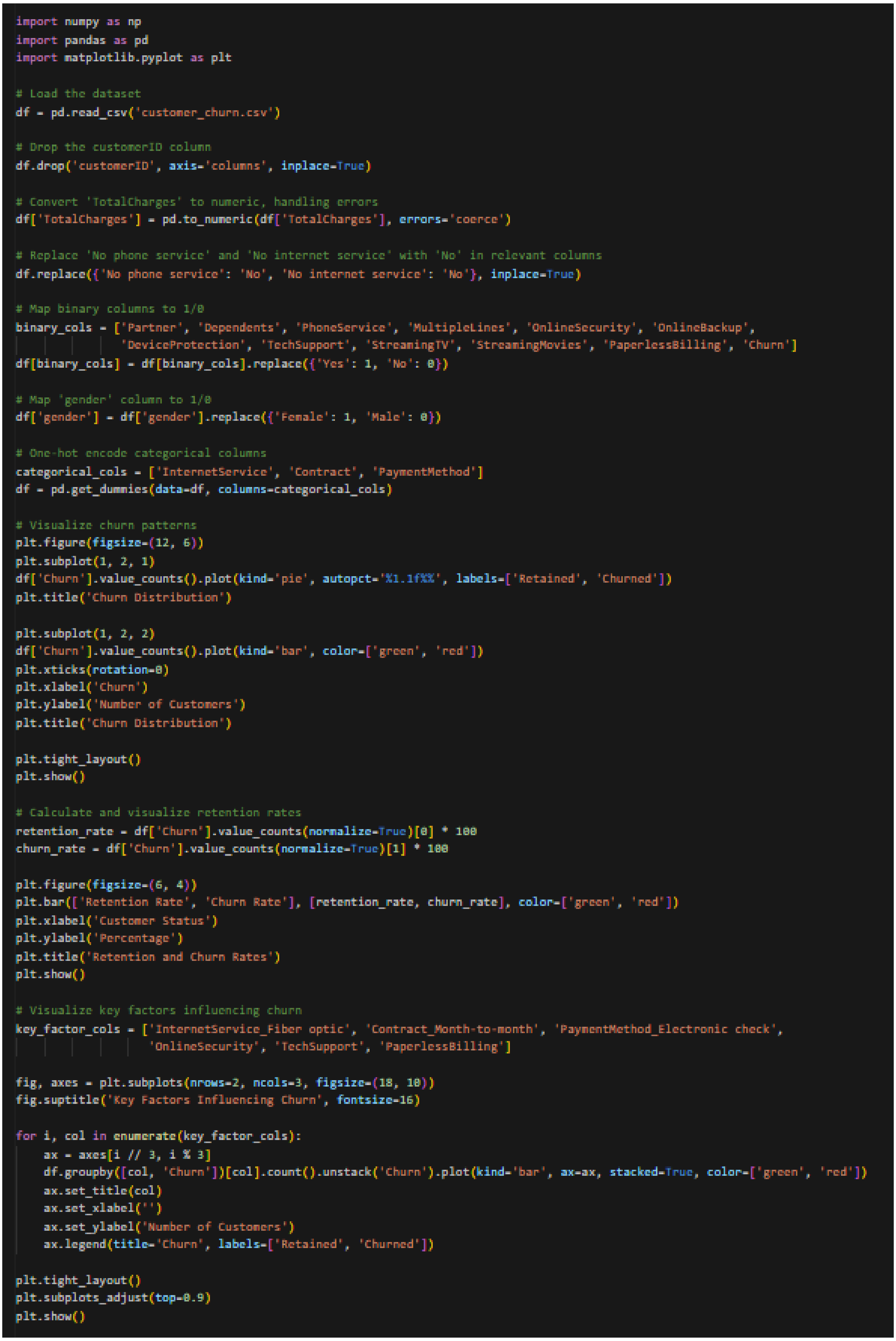
Data visualization is the use of graphical elements such as charts, graphs, and maps to represent data and information visually. The use of visualization tools provides an accessible way to see and understand trends, outliers, and patterns in data.

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

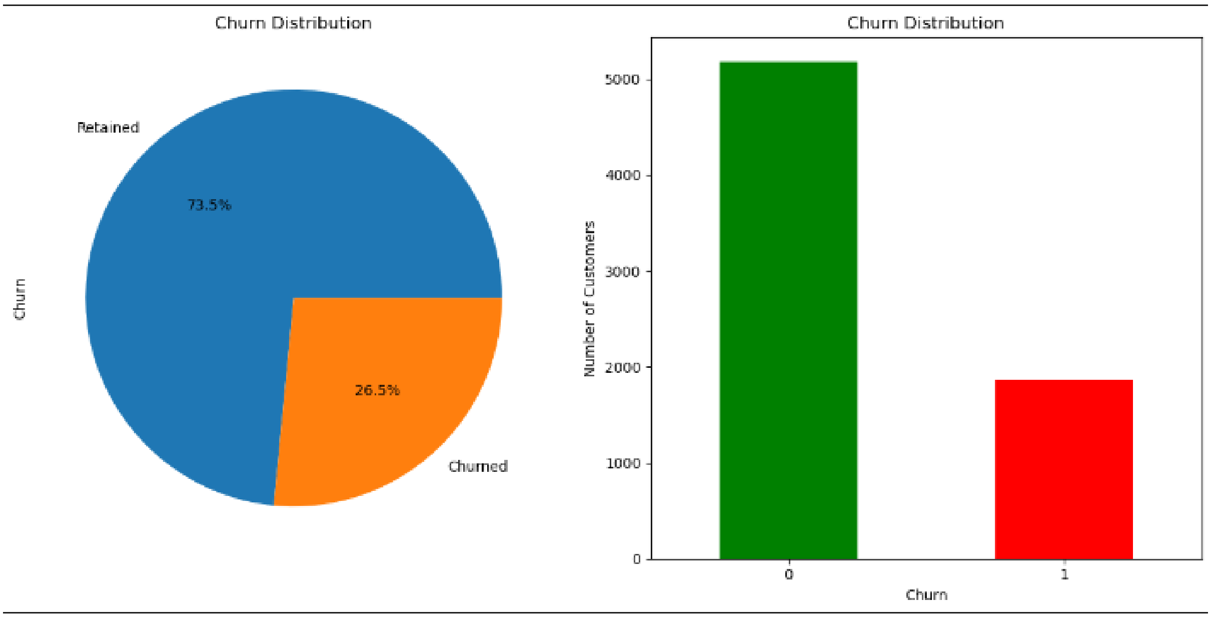
It helps to gain insights, identify patterns, and communicate your findings effectively.

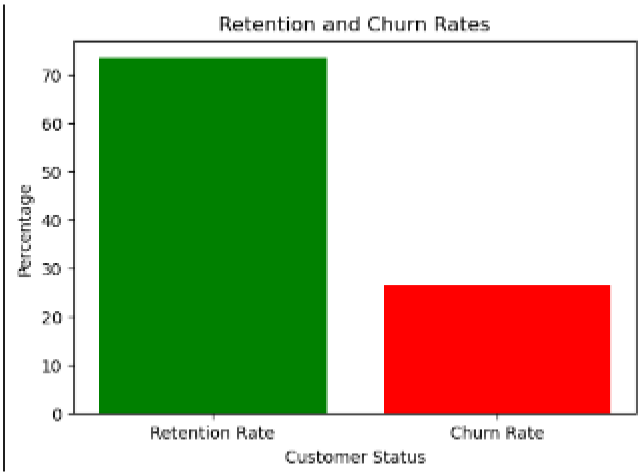
There are various techniques in data visualization. Few of them are described below,

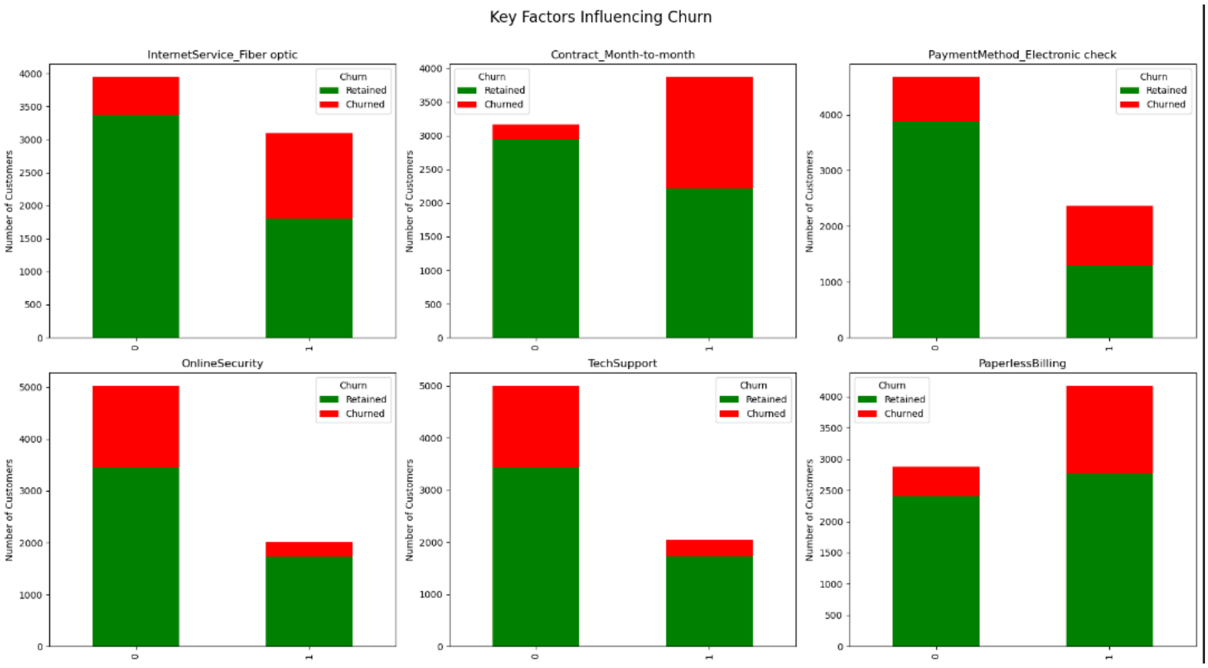
- ***Histograms***: Plot the frequency distribution of numerical variables to identify patterns and distributions.
- ***Bar Charts***: Used for categorical data to show the frequency of different categories.
- ***Pie charts:*** Display the correlation between variables using pie circles and color gradients.



```python
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt

# Load the dataset
df = pd.read_csv('customer_churn.csv')

# Drop the customerID column
df.drop('customerID', axis='columns', inplace=True)

# Convert 'TotalCharges' to numeric, handling errors
df['TotalCharges'] = pd.to_numeric(df['TotalCharges'], errors='coerce')

# Replace 'No phone service' and 'No internet service' with 'No' in relevant columns
df.replace({'No phone service': 'No', 'No internet service': 'No'}, inplace=True)

# Map binary columns to 1/0
binary_cols = ['Partner', 'Dependents', 'PhoneService', 'MultipleLines', 'OnlineSecurity', 'OnlineBackup',
               'DeviceProtection', 'TechSupport', 'StreamingTV', 'StreamingMovies', 'PaperlessBilling', 'Churn']
df[binary_cols] = df[binary_cols].replace({'Yes': 1, 'No': 0})

# Map 'gender' column to 1/0
df['gender'] = df['gender'].replace({'Female': 1, 'Male': 0})

# One-hot encode categorical columns
categorical_cols = ['InternetService', 'Contract', 'PaymentMethod']
df = pd.get_dummies(data=df, columns=categorical_cols)

# Visualize churn patterns
plt.figure(figsize=(12, 6))
plt.subplot(1, 2, 1)
df['Churn'].value_counts().plot(kind='pie', autopct='%1.1f%%', labels=['Retained', 'Churned'])
plt.title('Churn Distribution')

plt.subplot(1, 2, 2)
df['Churn'].value_counts().plot(kind='bar', color=['green', 'red'])
plt.xticks(rotation=0)
plt.xlabel('Churn')
plt.ylabel('Number of Customers')
plt.title('Churn Distribution')

plt.tight_layout()
plt.show()

# Calculate and visualize retention rates
retention_rate = df['Churn'].value_counts(normalize=True)[0] * 100
churn_rate = df['Churn'].value_counts(normalize=True)[1] * 100

plt.figure(figsize=(6, 4))
plt.bar(['Retention Rate', 'Churn Rate'], [retention_rate, churn_rate], color=['green', 'red'])
plt.xlabel('Customer Status')
plt.ylabel('Percentage')
plt.title('Retention and Churn Rates')
plt.show()

# Visualize key factors influencing churn
key_factor_cols = ['InternetService_Fiber optic', 'Contract_Month-to-month', 'PaymentMethod_Electronic check',
                   'OnlineSecurity', 'TechSupport', 'PaperlessBilling']

fig, axes = plt.subplots(nrows=2, ncols=3, figsize=(18, 10))
fig.suptitle('Key Factors Influencing Churn', fontsize=16)

for i, col in enumerate(key_factor_cols):
    ax = axes[i // 3, i % 3]
    df.groupby([col, 'Churn'])[col].count().unstack('Churn').plot(kind='bar', ax=ax, stacked=True, color=['green', 'red'])
    ax.set_title(col)
    ax.set_xlabel('')
    ax.set_ylabel('Number of Customers')
    ax.legend(title='Churn', labels=['Retained', 'Churned'])

plt.tight_layout()
plt.subplots_adjust(top=0.9)
plt.show()
```


Churn Distribution
Retained
73.5%
26.5%
Churned
Churn
Churn Distribution
Number of Customers
5000
4000
3000
2000
1000
0
0
1
Churn


Retention and Churn Rates
Percentage
70
60
50
40
30
20
10
0
Retention Rate
Churn Rate
Customer Status


Key Factors Influencing Churn
InternetService_Fiber optic
Contract_Month-to-month
PaymentMethod_Electronic check
OnlineSecurity
TechSupport
PaperlessBilling
Number of Customers
Churn
Retained
Churned
0
1

KEY FINDINGS AND INSIGHTS

Churn prediction, often used in the context of customer churn, aims to identify customers or users who are likely to stop using a product or service. The analysis of churn data can provide valuable insights into customer behavior and help businesses make informed decisions to reduce churn and retain valuable customers. Here are some key findings and insights in churn prediction:

1. Churn Rate: Calculate the overall churn rate, which is the percentage of customers who have churned in a given period. This helps in understanding the scale of the problem.

2. Demographics Analysis: Examine whether certain demographic characteristics (e.g., age, gender, location) are correlated with higher churn rates. This can help tailor marketing efforts to specific customer segments.

3. Usage Patterns: Analyze how often and how intensively customers use your product or service. Customers who use a product more frequently or extensively may have lower churn rates.

4. Customer Feedback: Study customer feedback, surveys, and support tickets to identify common issues or complaints that could lead to churn. Addressing these pain points can reduce churn.

5. Contract Analysis: For subscription-based services, analyze the types of contracts customers are on. Longer-term contracts or commitments may lead to lower churn rates.

6. Competitor Analysis: Consider competitive pressures. Customers might be leaving for a competitor with better features, pricing, or customer service.

7. Engagement Metrics: Track customer engagement metrics, such as login frequency, feature usage, and interactions with your platform. A drop in engagement may indicate a higher likelihood of churn.

8. Customer Lifetime Value (CLV): Calculate CLV to identify which customers are the most valuable to your business. Focus on retaining high CLV customers.

9. Predictive Models: Use machine learning models to predict which customers are most likely to churn. Common models include logistic regression, decision trees, and random forests. Insights from these models can help target at-risk customers.

10. Retention Strategies: Develop and implement customer retention strategies. For example, offer personalized incentives, loyalty programs, or targeted marketing campaigns to retain customers.

11. A/B Testing: Experiment with different strategies to see what works best for retaining customers. A/B testing can help determine which interventions are most effective.

12. Feedback Loop: Continuously gather feedback from customers who have churned. This information can help improve your product or service to prevent future churn.

13. Customer Segmentation: Segment your customer base into different groups based on behavior, needs, or characteristics. Tailor retention strategies to each segment.

14. Early Warning Systems: Implement early warning systems to identify customers who show early signs of disengagement. Timely intervention can prevent churn.

15. Churn Costs: Calculate the cost of acquiring new customers versus retaining existing ones. This can help in making informed decisions about investing in customer retention.

16. Customer Satisfaction Surveys: Regularly conduct customer satisfaction surveys to gauge how satisfied your customers are with your product or service. Address areas of dissatisfaction promptly.

17. Product Updates and Enhancements: Keep your product or service up to date with the latest features and enhancements to maintain customer interest and satisfaction.

18. Customer Support: Offer excellent customer support to address issues and concerns promptly. Good support can significantly reduce churn.

19. Loyalty Programs: Implement loyalty programs that reward long-term customers and encourage them to stay.

20. Monitoring and Feedback Loop: Continuously monitor and adjust your churn prediction model and customer retention efforts based on ongoing data and feedback.

RECOMMENDATIONS

Churn prediction is a crucial task for businesses looking to retain customers and reduce revenue loss. To build effective churn prediction models, consider the following recommendations:

1. Data Collection and Preparation:

- Collect comprehensive customer data, including demographic information, usage patterns, transaction history, and customer feedback.
- Ensure data quality and consistency by handling missing values, outliers, and data normalization.

2. Feature Engineering:

- Create relevant features, such as customer tenure, frequency of product usage, customer lifetime value (CLV), and customer behavior metrics.
- Incorporate sentiment analysis of customer feedback and support ticket data.

3. Label Definition:

- Define a clear and accurate churn label. Churn can be binary (e.g., churned or not churned) or multi-class (e.g., low churn risk, medium churn risk, high churn risk).

4. Data Splitting:

- Split the data into training, validation, and test sets to assess model performance accurately.
- Use time-based splitting to mimic real-world scenarios, where models are trained on historical data and tested on future data.

5. Model Selection:
   - Experiment with various machine learning algorithms, including logistic regression, decision trees, random forests, gradient boosting, and neural networks.
   - Consider using ensemble models or stacking techniques for improved performance.

6. Hyperparameter Tuning:
   - Perform hyperparameter optimization to fine-tune model parameters for better predictive accuracy.

7. Evaluation Metrics:
   - Choose appropriate evaluation metrics, such as accuracy, precision, recall, F1-score, and area under the ROC curve (AUC), depending on the business's specific needs and the class distribution in the data.

8. Feature Importance Analysis:
   - Analyze feature importance to understand which factors most significantly contribute to churn prediction. This insight can guide targeted retention efforts.

9. Ensemble Methods:
   - Combine the predictions of multiple models (e.g., by using a voting classifier or stacking) to improve model robustness and reduce overfitting.

10. Regular Model Updating:
    - Regularly retrain and update churn prediction models to account for changing customer behaviors and evolving market conditions.

11. Feedback Loop:
   - Establish a feedback loop to continuously gather customer feedback, especially from customers who have churned. This data can help refine the model and identify areas for product or service improvement.

12. Early Warning Systems:
   - Develop early warning systems to detect and alert teams to customers displaying early signs of disengagement, enabling proactive intervention.

13. Customization:
   - Customize retention strategies for different customer segments based on behavior, preferences, and characteristics.

14. A/B Testing:
   - Conduct A/B tests to validate the effectiveness of retention strategies and interventions. Test different approaches to identify the most successful ones.

15. Loyalty Programs:
   - Implement loyalty programs, discounts, or rewards for long-term and high-value customers to encourage loyalty and reduce churn.

16. Customer Support:
   - Provide exceptional customer support to address customer concerns and issues promptly. Satisfied customers are less likely to churn.

17. Monitor Key Metrics:
   - Continuously monitor key performance metrics, including churn rate, customer satisfaction, and CLV, to gauge the effectiveness of your retention efforts.

CONCLUSION

Churn prediction plays a vital role in today's highly competitive business landscape. It enables organizations to proactively address customer attrition by identifying and targeting customers at risk of leaving. In this process, businesses harness the power of data analytics to gain a deeper understanding of customer behaviors and preferences.

The key insights derived from churn prediction models help businesses make informed decisions and tailor their strategies to reduce churn rates. By taking a data-driven approach, companies can focus their resources on retaining valuable customers, improving customer satisfaction, and ultimately sustaining long-term growth.

Churn prediction goes beyond the technical aspect of machine learning models. It involves cross-functional collaboration, where data scientists, marketing teams, customer support, and product development work together to implement effective retention strategies. Regular model updates and a feedback loop that collects input from churned customers are essential components of this process.

Furthermore, early warning systems and customer segmentation are used to detect and respond to signs of customer disengagement. These measures enable businesses to intervene in a timely manner and prevent churn.

In conclusion, churn prediction is not merely a predictive analytics task; it's a holistic approach to understanding and engaging with customers. It empowers organizations to navigate the challenges of customer attrition, adapt to changing market conditions, and prioritize customer satisfaction. By leveraging data and advanced analytics, businesses can achieve higher customer retention rates, lower acquisition costs, and ultimately, long-term success in their respective industries.