Telecommunication Customer Churn Prediction Report

Problem Statement

The telecommunication industry faces challenges in maintaining customer loyalty, with churn rates reflecting customers who discontinue services such as mobile plans, internet subscriptions and broadband connections.

The report outlines the comprehensive analysis conducted on the Telecommunication Customer Churn Prediction Dataset from Kaggle, aimed at enhancing customer retention and satisfaction within the telecommunication sector.

Addressing the issue is critical for :

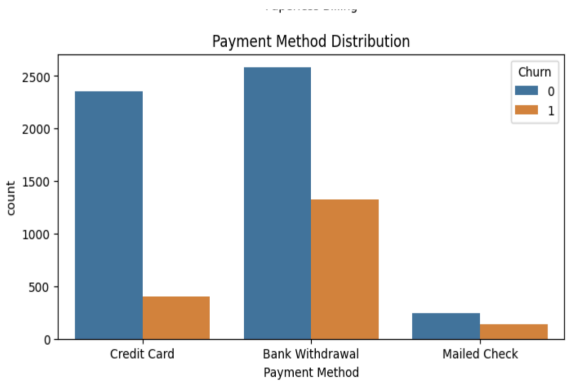
1. Revenue Retention: Retaining existing customers is more cost-effective than acquiring new ones, making churn reduction vital for steady revenue streams.
2. Brand Reputation: High customer retention rates signify reliability and profitability, enhancing brand perception in the market.
3. Data-Driven Insights: Analyzing customer data for churn prediction provides insights that guide strategic initiatives, product/service improvements, and targeted marketing efforts.

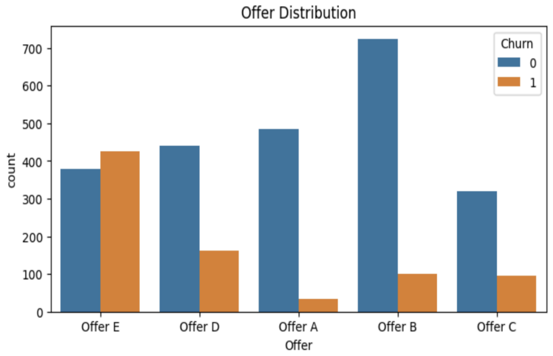
Dataset Overview: The dataset under analysis comprises information on 7043 customers from a telecommunication company in California for Q2 2022. It includes demographics, location, tenure, subscriptions, services, and customer status for the quarter. The dataset consists of 38 columns. Out of the 38 columns, 15 are numerical type and 23 are categorical type values.

Exploratory Data Analysis

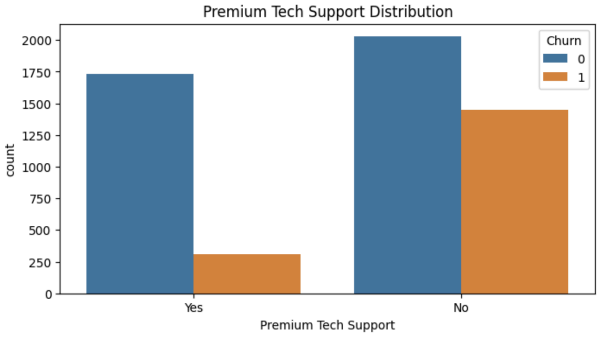
Our EDA analysis on Telecommunication Customer Churn Dataset identified critical factors influencing customer churn, enabling targeted strategies for enhanced retention. Notably:

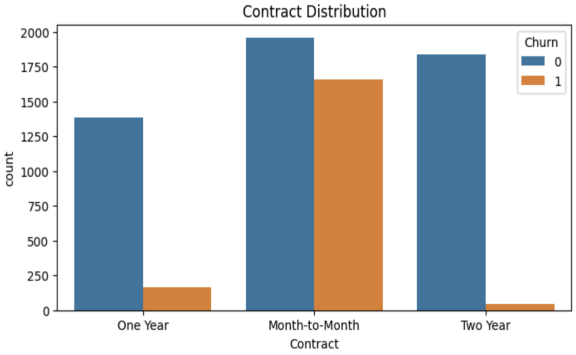
1. Offer Type and Payment Method: These 2 features significantly affect churn rates, with certain offers and automatic billing via credit cards correlating to higher retention.





2. Premium Tech Support and Longer Contracts are associated with reduced churn, underscoring the importance of customer support quality and contract stability.





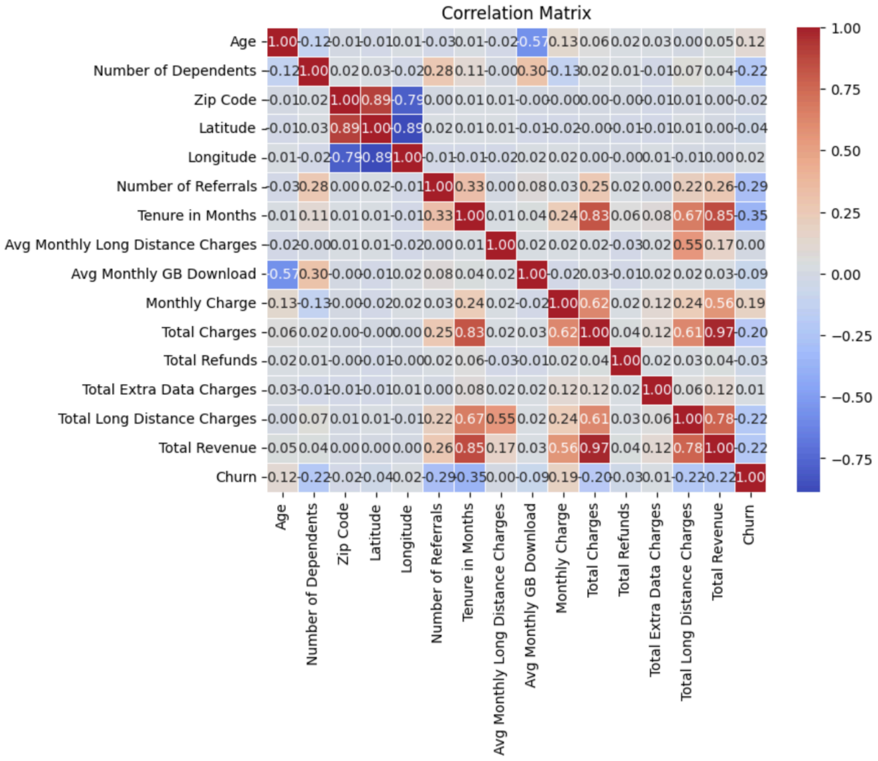
3. Price Sensitivity plays a pivotal role, with higher monthly charges leading to increased churn rates.

From EDA, we eliminated non-informative features such as Customer ID and location-based data (City, Zip, Code) to concentrate on variables directly influencing churn. Certain services (eg., Internet Service, Online Security) emerged as pivotal in predicting churn, suggesting tailored customer experience improvements. Financial attributes, including Monthly Charge, Total Charges, and Total Revenue, were highlighted as critical, underscoring the need to focus on pricing strategies and service value to enhance customer loyalty. The length of the customer's contract significantly influences churn, with Month-to-Month contracts showing higher churn rates compared to longer-term agreements.

Feature Engineering

To prepare the data for feature engineering, thorough data cleaning and preprocessing was done as follows:

1. Eliminated irrelevant columns to focus on significant predictors of churn.
2. Addressed missing values and standardized numerical features to improve model performance.
3. Encoded Categorical Variables: Categorical variables like Gender, Internet Service, and Contract were encoded using one-hot or binary encoding to make them usable for machine learning models.
4. Standardization of Numerical Values: Numerical features were standardized using techniques like log transformation for right-skewed distributions (Monthly Charge, Total Charges) to normalize their scales and improve model training efficiency.
5. Correlation Analysis: Correlation between features was evaluated to avoid multicollinearity, which could skew the model's predictions. Highly correlated features such as Total charges and monthly charges were identified for potential exclusion to ensure model robustness.

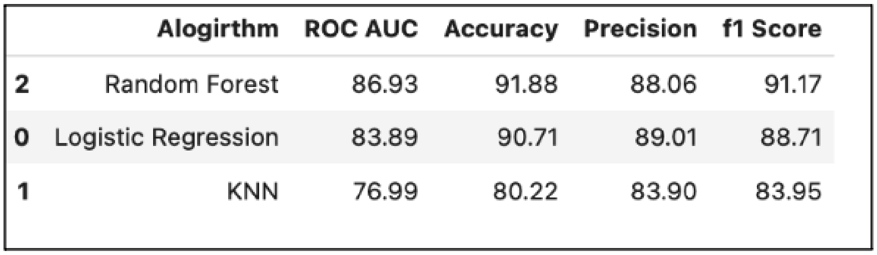


Modeling Results

The model utilized a 70/30 train-test-split with stratified sampling to maintain a balanced representation of churned and non-churned customers.

Training was done on the cleaned and processed data on 3 models to identify the model that best suits the data and can be used to better predict customer churn rate.

The following are the results from the 3 models:



	Alogirthm	ROC AUC	Accuracy	Precision	f1 Score
2	Random Forest	86.93	91.88	88.06	91.17
0	Logistic Regression	83.89	90.71	89.01	88.71
1	KNN	76.99	80.22	83.90	83.95

The Random Forest model demonstrated the highest accuracy, AUC score, and F1 score, indicating its superior ability to differentiate between customers who would churn and those who would not.

Logistic Regression, noted for its high precision, proved especially valuable in correctly identifying churned customers with minimal false positives. Despite its lower precision, the K Nearest Neighbors model also performed reasonably well and was recognized as a viable option for training the dataset.

Given the models' performances, Logistic Regression seems like the best option as the preferred model due to its high precision, ease of interpretability, and simplicity, which are crucial for dynamic adaptation to customer behavior changes. Moreover, the model's interpretability provides significant insights into the impact of each feature on churn likelihood, aiding in strategic decision-making.

Deployment Considerations

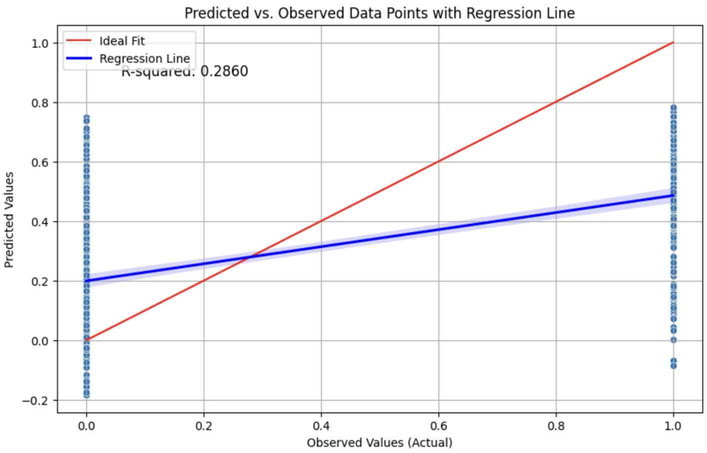
The telecommunication customer churn prediction model employs a strategic deployment on cloud platforms like Google Cloud Platform (GCP) or Amazon Web Services (AWS) SageMaker, leveraging their scalability for real-time analytics. This approach includes containerization with Docker for consistent environments across different platforms and deploying the model as a microservice with a REST API, enabling seamless integration and providing immediate insights to customer relationship managers.

To ensure the model's effectiveness and integrity, continuous monitoring with tools like Grafana or Kibana, implementation of CI/CD pipelines for regular updates, version control for datasets and models, and a feedback loop for ongoing refinement are integral.

Equally important are stringent security measures including data encryption, strict access control, regular security audits, and adherence to data protection guidelines such as GDPR, establishing a zero-trust architecture to safeguard sensitive customer data and maintain customer trust.

Appendix

1. To enhance model performance and address dataset imbalances, strategies such as employing oversampling techniques like SMOTE and hyperparameter tuning can be considered.
2. **Linear Regression** performance:



R squared value is only 0.286. Linear regression models are based on the assumptions of linearity and independence of features.