Predicting Customer Churn in the Telecom Sector Using Machine Learning

Abhay Pardeshi
Management Information Systems and Analytics Department
School of Business & Economics
State University of New York at Plattsburgh

Abstract

Telecom companies spend a lot to acquire new customers, making it crucial to retain these customers and ensure they continue using their plans. Churn refers to customers leaving their current service provider for another one. Once companies secure a long-term customer, they can profit significantly from them. Therefore, customer churn is a major concern for a company's revenue and business growth, especially in the competitive telecommunications industry. By analyzing customer behavior and the services they use, companies can predict churn. This predictive capability allows them to minimize losses and enhance their business performance.

1. Introduction

Businesses face a serious problem with loss of customers, particularly in the highly competitive telecom sector. Customers quitting their subscriptions or moving to competitors is referred to as customer churn. Since keeping current customers is typically less expensive than finding new ones, high churn rates result in significant revenue losses. This emphasizes how crucial churn prediction is to the long-term sustainability of businesses. Businesses can find churn predictors and behavioral patterns by utilizing robust machine learning models. This allows for focused interventions that improve long-term profitability and client retention.

Rich customer-level data, such as customer IDs, demographic information, billing details, usage patterns, and service-related data, are available in the Customer Churn Prediction Dataset used in this study. The dataset is appropriate for supervised learning tasks since it is labeled to show whether a client has churned. Building a predictive model that identifies clients at high risk of leaving requires key data including tenure, contract type, monthly expenses, and internet service type. Telecom businesses may ensure higher customer satisfaction and lower attrition by evaluating these factors and putting proactive initiatives like tailored marketing campaigns, loyalty plans, and improved customer service into place.

This project aims to create and assess a machine-learning model that can accurately predict customer churn. The project aims to provide actionable insights for commercial applications through model training, exploratory data analysis (EDA), data pretreatment, and performance evaluation. This strategy not only improves the business's capacity to predict employee attrition, but it also offers a framework for implementing predictive analytics in other sectors.

1.1. Project Background

The dataset comprises detailed records of customer interactions, including demographic attributes, service subscription details, billing patterns, and churn labels. These features allow a comprehensive analysis of customer behavior and churn determinants. As the telecom sector faces increased competition, predicting and mitigating churn has become a crucial factor in maintaining a competitive edge.

Advancements in machine learning provide an opportunity to analyze large-scale data and uncover trends that are not immediately apparent through manual analysis. Key factors, such as contract length, monthly charges, and tenure, are especially relevant to churn prediction, as they encapsulate the customer experience and relationship with the service provider.

This project leverages machine learning algorithms, including Logistic Regression, Decision Trees, Random Forests, and XGBoost, to build a predictive model. These models allow for the identification of patterns and trends, enabling companies to predict high-risk customers and tailor interventions to improve retention. For instance, customers with short tenures or those on month-to-month contracts may require customized loyalty programs, while those with high billing discrepancies may benefit from billing support. This predictive approach ensures not only improved churn prediction accuracy but also better resource allocation for customer retention efforts.

1.2. Purpose and Scope

The primary purpose of this project is to develop a robust machine-learning model capable of predicting customer churn based on key service-related and demographic features. The model aims to identify at-risk customers by analyzing features such as contract type, monthly and total charges, service tenure, and subscription details. The insights gained can inform strategies to enhance customer satisfaction and loyalty, ultimately reducing churn rates.

The project encompasses the following major steps:

➔ **Data Preprocessing:** This includes loading and cleaning the dataset, handling missing values, addressing outliers, encoding categorical variables into numerical formats, and scaling numerical features to ensure consistency and balance.

➔ **Exploratory Data Analysis (EDA)**: Identifying patterns, visualizing feature distributions, and understanding correlations between features and churn labels. This step highlights the key predictors of churn, such as contract type and monthly billing.

➔ **Model Building**: Developing and evaluating multiple machine learning models, including Logistic Regression, Decision Trees, Random Forests, and XGBoost. Models are trained and validated to ensure robustness.

➔ **Model Evaluation**: Assessing model performance using metrics like accuracy, precision, recall, F1-score, and the AUC-ROC curve to determine the most effective predictive model.

2. Data Analysis

2.1. Literature Review

The Customer Churn dataset includes important features such as Age, Gender, Tenure, MonthlyCharges, ContractType, InternetService, TechSupport, and Churn. Below is a table indicating the features and their meanings.

Key features in the dataset include:

Features	Explanation
CustomerID	Unique identifier for each customer.
Age	The age of the customer, reflecting their demographic profile.
Gender	Gender of the customer (Male or Female)
Tenure	Duration (in months) the customer has been with the service provider.
MonthlyCharges	The monthly fee charged to the customer.
ContractType	The type of contract the customer is on (Month-to-Month, One-Year, Two-Year).
InternetService	Type of internet service subscribed to (DSL, Fiber Optic, None).
TechSupport	Whether the customer has tech support (Yes or No).
TotalCharges	Total amount charged to the customer (calculated as MonthlyCharges * Tenure).
Churn	Target variable indicating whether the customer has churned (Yes or No).

Table 2.1: key features

Customer churn represents a persistent and growing challenge in the telecommunications industry, where retaining subscribers is far more cost-effective than acquiring new ones. The cost of marketing campaigns aimed at attracting new customers can be astronomical, and despite these efforts, many companies still face the inevitable problem of losing customers. According to a 2024 report, the average cost to acquire a single customer in the telecommunications industry is $694. This figure highlights just how expensive it is to bring in new subscribers, making it all the more critical for companies to focus on retention. The real goldmine, however, lies in improving retention rates among existing subscribers. It's far more cost-effective to retain a customer than to acquire a new one, and small improvements in retention can lead to significant financial gains. Reichheld and Schefter's research famously showed that even a modest 5% boost in retention could drive profits up by anywhere from 25% to 95%. While these figures were groundbreaking when they were first published, the dynamics of customer expectations and competition in today's market suggest that these numbers might need revisiting, especially as customer behavior evolves and new technologies disrupt traditional models.

The question of why customers leave, however, is not straightforward, and solving it feels a lot like piecing together a complex puzzle. A few years back, Choudhury and their team delved into this very issue, exploring the core reasons why people churn. Their findings pointed to frustrations over poor service quality, high prices, and unhelpful or ineffective customer support as major factors driving customers away. While these results align with conventional wisdom, what really caught my eye was how Choudhury's team used logistic regression to quantify the relationships between these factors and churn. It wasn't just about knowing what causes churn, but understanding the relative impact of each factor. This type of analysis was crucial in refining retention strategies, as companies could prioritize interventions that would have the greatest effect.

The landscape of churn prediction and customer retention strategies underwent a significant transformation in 2020 with the work done by Tzeng's group. Moving away from traditional statistical methods, Tzeng and their colleagues embraced machine learning techniques, particularly decision trees and random forests, to identify patterns in customer behavior. Their approach was more than just an academic exercise; it provided telecom companies with practical tools to pinpoint customers who were at a higher risk of leaving. This shift was like introducing an early warning system – a way to flag at-risk customers well before they made the decision to churn. By identifying these customers in advance,

companies could take proactive measures, such as offering targeted promotions or improving service quality, to retain them.

Perhaps the most exciting shift in recent years comes from the research of Chen in 2022. The industry has evolved beyond simple predictive models into more sophisticated applications of artificial intelligence (AI) to analyze customer behavior in real time. Rather than just predicting who might leave, modern telecom companies are using AI to understand the underlying reasons for customer dissatisfaction. AI can analyze a range of data sources, from call logs to customer service interactions, to identify patterns of behavior that suggest a customer is becoming discontented. What's more, these technologies enable companies to intervene before a customer decides to leave, offering tailored solutions or incentives that can sway them back toward loyalty. This evolution has shifted churn prediction from a reactive to a proactive strategy, making it possible not only to anticipate customer behavior but to influence it directly.

The evolution of churn analysis reflects a broader trend in the telecom industry: a move from one-size-fits-all approaches to more personalized and dynamic strategies. In the past, companies might have employed broad discounting strategies to retain customers, but today's advanced methods allow for more precise, data-driven interventions. These strategies are informed by a deeper understanding of the customer journey, enabled by a combination of machine learning, AI, and behavioral data analysis. The future of churn prediction is not just about forecasting but about taking immediate, targeted actions to ensure customers stay satisfied and engaged. Given the high costs of acquiring new customers, companies that prioritize retention are not only preserving their bottom line but also building long-term customer loyalty, making retention efforts even more crucial in a highly competitive and expensive market.

2.2. Data Challenges and Abnormalities

Several challenges arise when working with the Telecom Customer Churn Insights dataset, which must be addressed before building the prediction model:

➔ **Missing Data:** Some columns like Tenure, totalCharges, and InternetService have some missing data. The reason for missing data and how can we eliminate them or solve them will be applied.

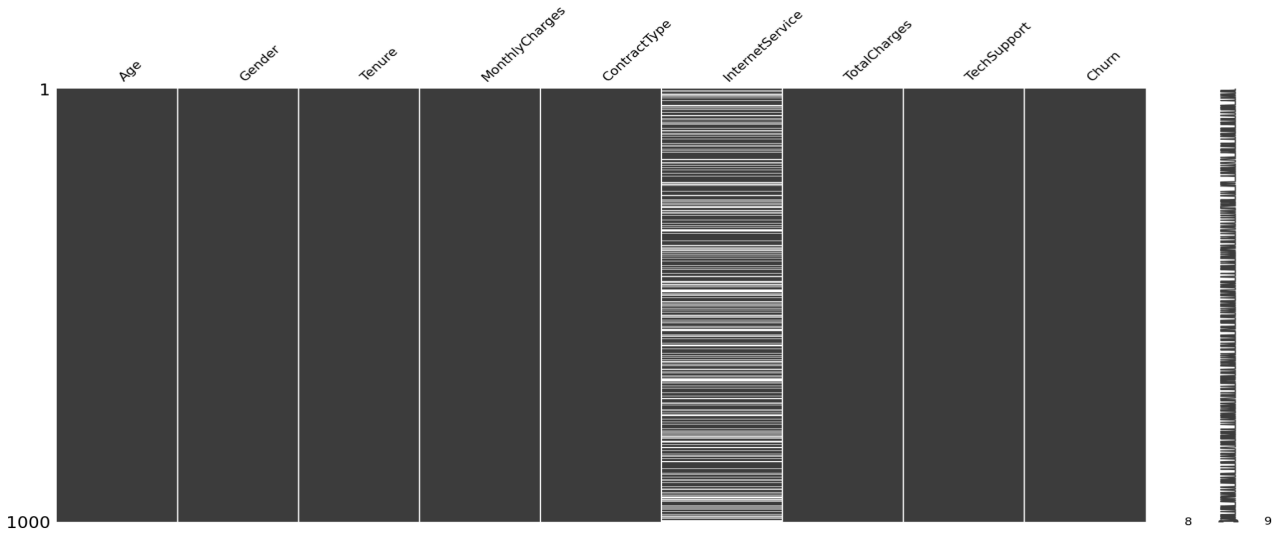

Fig 1 : Missing Values

➔ **Imbalanced Data:** The dataset has a class imbalance with a small proportion of customers who have churned compared to those who have not. This imbalance can negatively impact the performance of machine learning models, especially those sensitive to skewed distributions, like logistic regression. Techniques such as oversampling the minority class (churners), undersampling the majority class (non-churners), or applying SMOTE (Synthetic Minority Over-sampling Technique) may be used to address this issue.

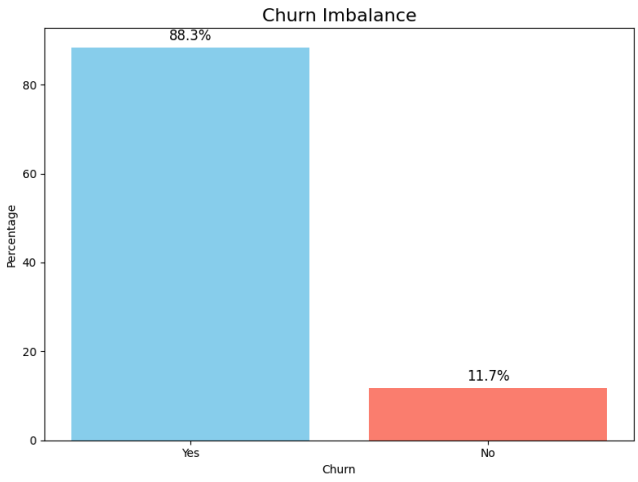

Fig 2: imbalanced data - churners vs non-churners

➔ **Multicollinearity:** Some features, like Monthly Charges and Total Charges, might be correlated, which could affect model performance. Detecting and addressing multicollinearity through techniques such as variance inflation factor (VIF) analysis or feature elimination will be necessary.

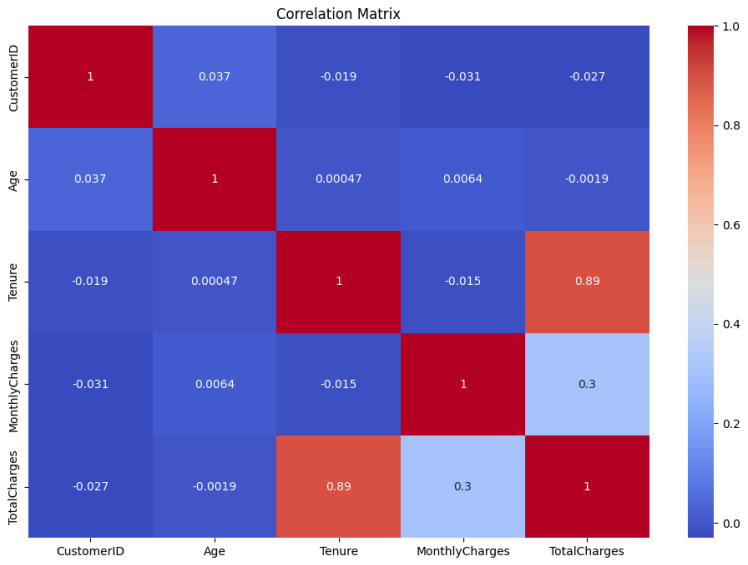

Fig 3: Multicollinearity - correlation matrix

➔ **Categorical Variables:** Categorical variables, such as TechnicalSupport, and ServiceType, must be converted into numerical values or boolean before feeding them into machine learning algorithms. Techniques such as one-hot encoding will be applied to properly handle these features.

2.3. Exploratory Data Analysis (EDA)

The customer churn prediction project begins with a comprehensive exploration of a dataset containing 1,000 customer records. The dataset encompasses a rich array of features including demographic information, service details, and customer behavior metrics. Initial data review reveals a diverse customer base with ages ranging from 12 to 83 years, and a relatively balanced gender distribution, with females representing 53.8% and males 46.2% of the total customers.

One of the most striking observations from the exploratory data analysis is the significant class imbalance in the churn variable. An overwhelming 88.3% of customers have churned, compared to only 11.7% who have remained with the service provider. This dramatic imbalance immediately signals a critical business challenge and suggests the need for sophisticated modeling techniques to accurately predict and understand customer attrition.

The analysis delves deep into various dimensions of customer characteristics. Contract types show interesting patterns, with month-to-month contracts dominating at 51.1%, followed by one-year contracts at 28.9%, and two-year contracts at 20%. Internet service distribution reveals a preference for fiber optic services, with a notable number of customers having undefined or unknown service types. This nuanced breakdown provides insights into customer preferences and potential churn drivers.

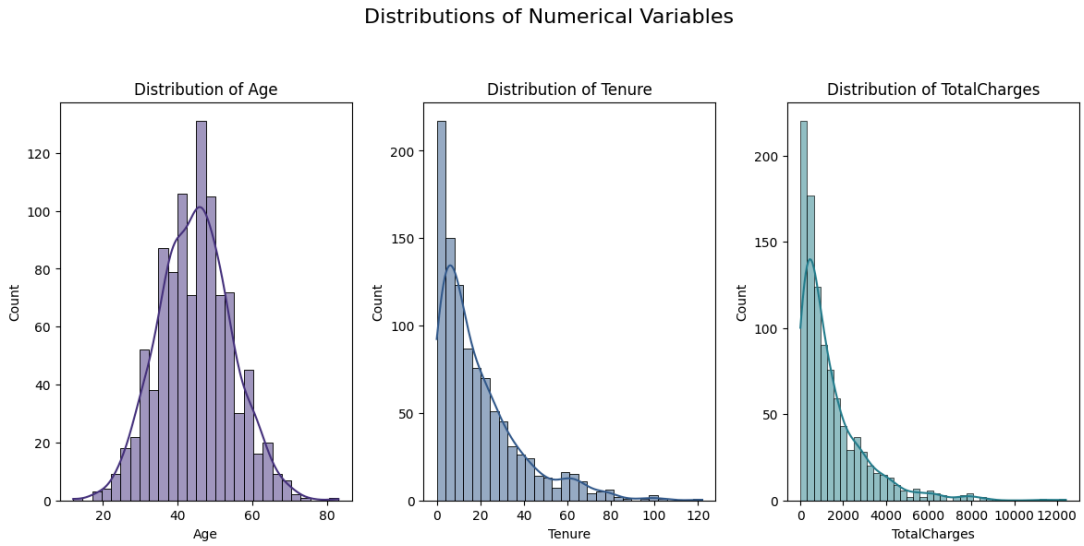

Fig 4 : Distributions of Numerical Variables

2.4. Feature Engineering

Feature engineering played a crucial role in transforming raw data into meaningful predictive inputs. The researchers employed several sophisticated techniques to enhance the dataset's predictive power. Age was transformed into categorical groups, allowing for more nuanced analysis of how different age segments interact with the service. Categorical variables were one-hot encoded, and missing values in the internet service column were strategically handled by introducing an "Unknown" category rather than simply dropping or imputing the data.

The visualization techniques used in the analysis were particularly illuminating. Histogram plots explored the distributions of key numerical variables like age, tenure, and monthly charges. A correlation matrix revealed the interconnections between different numerical features, helping to identify potential predictive relationships. Additional visualizations broke down contract types, tech support usage, and churn rates across different demographic segments, providing a multifaceted view of customer behavior.

Statistical summaries revealed additional insights into customer characteristics. The average customer age is 44.7 years, with an average monthly charge of $74.39 and an average tenure of 18.97 months. These baseline metrics provide context for understanding individual customer variations and overall business performance. The comprehensive approach to data exploration not only prepares the data for predictive modeling but also generates actionable insights for business strategy.

The most critical recommendation emerging from this analysis is the need to address the extreme class imbalance. Potential strategies include using techniques like SMOTE (Synthetic Minority Over-sampling Technique), implementing class weights in model training, and exploring ensemble methods that are robust to imbalanced datasets. Moreover, the analysis suggests developing targeted retention strategies based on age groups, contract types, and other distinguishing customer characteristics.

By combining rigorous statistical analysis, advanced feature engineering, and insightful visualization techniques, this churn prediction project transforms raw customer data into a powerful tool for understanding and predicting customer attrition. The approach goes beyond simple predictive modeling, offering a comprehensive framework for

understanding the complex dynamics of customer retention in a competitive service environment.

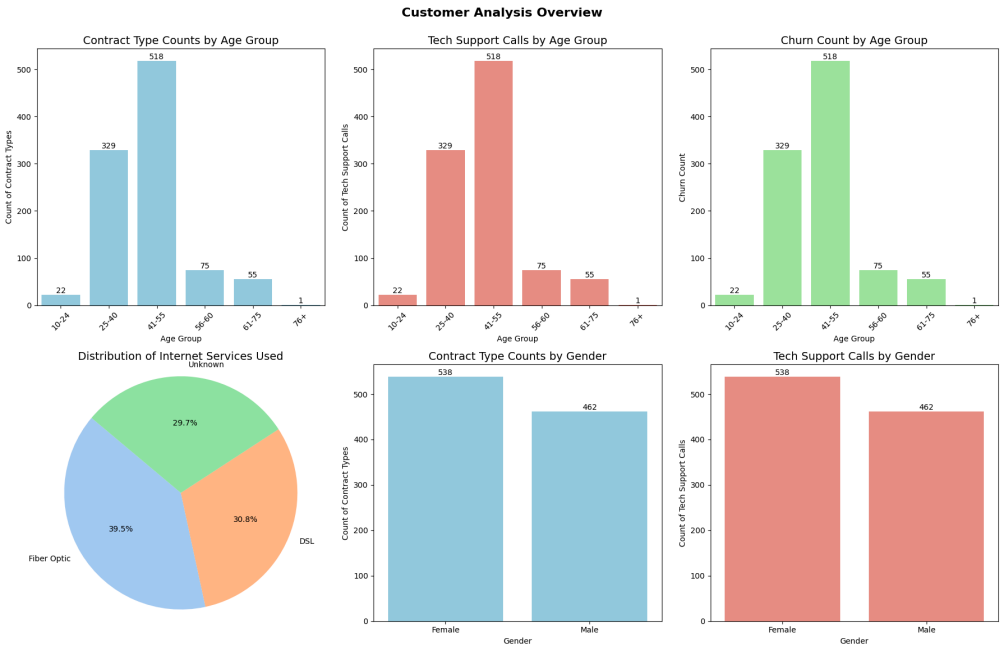

Fig 5: customer analysis overview

3. Methodology

3.1. Logistic Regression

Logistic regression was employed to predict customer churn based on various customer features. The model aimed to estimate the probability of a customer churning (leaving the service) based on characteristics such as tenure, monthly charges, contract type, and tech support.

Data Preprocessing

Data preprocessing involved scaling continuous features using StandardScaler to ensure equal contribution of each feature, and one-hot encoding for categorical variables such as ContractType_One-Year, ContractType_Two-Year, and TechSupport_Yes. This step standardized the dataset for the logistic regression model.

Feature Selection

Initially, features like Tenure, MonthlyCharges, and TotalCharges were included. However, due to the high correlation between Tenure and TotalCharges (0.87), TotalCharges was excluded after analysis, as it was redundant. The final selected features were: Age, Tenure, ContractType_One-Year, ContractType_Two-Year, InternetService_Unknown, TechSupport_Yes, and age group categories.

Model Building and Fitting

The logistic regression model was built using Scikit-learn's LogisticRegression class, with a balanced class weight strategy to handle class imbalance. The model was trained on the training set and evaluated on both the training and validation sets. Metrics like accuracy, precision, recall, F1-score, and confusion matrix were used to assess performance.

Model Evaluation

The model achieved an accuracy of 98.5% on the training set, with excellent precision and recall for predicting churn. However, the validation set showed slightly reduced performance, with an accuracy of 94%, suggesting some overfitting to the training data.

Model Interpretability

The coefficients of the logistic regression model revealed that Tenure and TechSupport_Yes had the most significant influence on churn predictions, with negative coefficients indicating lower churn for customers with longer tenure or who received tech support. The VIF analysis identified multicollinearity between Tenure and TotalCharges, which led to the exclusion of TotalCharges from the final model.

Conclusion

The logistic regression model, after feature selection and refinement, provided accurate predictions of customer churn. The model highlighted key factors influencing churn, including contract type, tenure, and tech support, which can guide retention strategies for telecom companies.

3.2. K-Nearest Neighbors (KNN)

The K-Nearest Neighbors (KNN) algorithm was applied to predict customer churn, using the features selected from the logistic regression model. KNN is a non-parametric algorithm that classifies a data point based on the majority class of its nearest neighbors in the feature space.

Model Building

The KNN model was built with k=1 and balanced class weights to address the class imbalance. Various values of k were tested, with the best performance achieved at k=5.

Model Evaluation

The KNN model achieved an accuracy of 91% on the test set, with precision for churn (class 1) at 0.94 and recall at 0.96. However, it had lower performance for non-churn (class 0), with precision of 0.63 and recall of 0.52. The confusion matrix showed 170 true positives, 12 true negatives, 7 false positives, and 11 false negatives.

ROC-AUC Score

The initial ROC-AUC score was 0.7411, which improved to 0.9156 after fine-tuning k and adjusting class weights.

Feature Importance

Although KNN doesn't directly provide feature importance, certain features significantly impacted predictions. The most important features were:

- ContractType_One-Year (0.0645)
- Tenure (0.0540)
- Age (0.0360)

- age_group_41-55 (0.0360)

Features like age_group_56-60 and age_group_76+ had minimal impact.

Conclusion

The KNN model performed well with an accuracy of 91%, excelling at predicting churn but struggling with non-churn predictions. The ROC-AUC score of 0.9156 and the analysis of feature importance suggest that contract type and tenure are key drivers for churn prediction. Further refinement could focus on improving the model's recall for non-churn cases.

3.3. Decision Tree

The Decision Tree model was used to predict customer churn with a maximum depth of 4 to prevent overfitting, using the Gini impurity criterion.

Model Evaluation

The model achieved a training accuracy of 97.17% and a test accuracy of 96.50%. The confusion matrix showed 23 true negatives, 170 true positives, 7 false negatives, and no false positives. The classification report indicated high precision and recall for churn, with a weighted average F1-score of 0.97.

Feature Importance

Key features influencing churn prediction were:

- Tenure (37.74%)
- ContractType_One-Year (19.55%)
- ContractType_Two-Year (16.10%)
- InternetService_Unknown (12.31%)
- MonthlyCharges (11.75%)

Overfitting and Best Parameters

Increasing max_depth beyond 4 did not improve test accuracy, indicating overfitting. The optimal parameters were max_depth=4, ensuring good generalization.

Conclusion

The Decision Tree model demonstrated strong performance, particularly for predicting churn, with high precision and recall. It balanced accuracy and generalization, with the optimal max_depth of 4.

3.4 Random Forest

The Random Forest model was used to predict customer churn, with hyperparameters optimized for performance. The best parameters identified were: max_depth=5, max_features='sqrt', min_samples_leaf=1, min_samples_split=200, and n_estimators=25.

Model Evaluation

The model achieved a training accuracy of 97.24% and a test accuracy of 95.50%. The performance was stable across different values of estimators, with slight variations in test accuracy ranging from 95.00% to 96.50%, showing that the model generalizes well.

Feature Importance

Key features that contributed the most to the model's predictions included:

- Tenure (Importance: 0.3148)
- ContractType_One-Year (Importance: 0.2183)
- MonthlyCharges (Importance: 0.1107)
- InternetService_Unknown (Importance: 0.1105)
- ContractType_Two-Year (Importance: 0.0884)

Age and Gender_Male had relatively low importance, with some age groups such as age_group_56-60 and age_group_76+ showing zero importance, indicating minimal contribution to the model's prediction.

Overfitting and Conclusion

The model exhibited stable performance across a range of estimators, without significant overfitting. The Random Forest model demonstrated strong predictive power with excellent generalization on the test set, especially when using the best hyperparameters. It highlighted the importance of Tenure, ContractType_One-Year, and MonthlyCharges in predicting churn.

3.5 XGBoost

The XGBoost model was used to predict customer churn, achieving a test accuracy of 95%. The best hyperparameters identified were:

- colsample_bytree: 0.7
- gamma: 0.1
- learning_rate: 0.3
- max_depth: 5
- min_child_weight: 1
- n_estimators: 25
- subsample: 0.8

Model Evaluation

The model achieved a training accuracy of 98%, and as the number of estimators increased, the test accuracy stabilized at 95%, with no significant improvement beyond 40 estimators.

Feature Importance

The most important features identified by the model were:

- ContractType_One-Year (Importance: 0.302)
- ContractType_Two-Year (Importance: 0.142)
- InternetService_Unknown (Importance: 0.119)
- Tenure (Importance: 0.093)
- age_group_41-55 (Importance: 0.062)

These results show that contract type and internet service status are the primary factors in predicting churn, while demographic features like age and tenure also play significant roles.

Conclusion

The XGBoost model performed well with strong predictive power, particularly when predicting churn based on contract type and internet service status. The model demonstrated high accuracy and stability with increasing estimators, achieving optimal performance with the specified hyperparameters.

4. Results

4.1. Logistic Regression Results

The logistic regression model demonstrated strong performance, achieving an accuracy of **96.25%** on the training set and **96.00%** on the validation set. The confusion matrix for the training set indicated **687 true negatives**, **19 false positives**, **672 true positives**, and **34 false negatives**, reflecting a robust classification capability. For the validation set, the matrix showed **19 true negatives**, **4 false positives**, **173 true positives**, and **4 false negatives**, indicating consistent performance across datasets. The model's Akaike Information Criterion (AIC) value was **1927.80**, signifying the trade-off between model fit and complexity. Comparatively, the K-Nearest Neighbors (KNN) classifier, using balanced class weights, achieved an accuracy of **92.00%** with a precision-recall balance leaning towards higher recall for class "1". The KNN ROC-AUC score was **0.9156**, highlighting its ability to distinguish between classes.

4.2. Decision Tree Results

The decision tree classifier excelled with a training accuracy of **97.17%** and a test accuracy of **96.50%**. The confusion matrix on the test data demonstrated perfect classification of the minority class, with **23 true negatives** and **170 true positives**, while producing **7 false negatives** and no false positives. The classification report indicated a **precision of 0.77** and a **recall of 1.00** for the minority class, suggesting the model effectively identified this class without over-predicting it. For the majority class, precision and recall were both excellent, contributing to an overall weighted accuracy of **96%**.

4.3. Random Forest Results

The random forest classifier achieved a test accuracy of **95.50%** with optimal parameters, including a maximum depth of **5**, using the square root of features per split, and a minimum sample split of **200**. The best training accuracy was **97.24%**, demonstrating the model's ability to generalize effectively. The relatively shallow depth and careful parameter tuning reduced overfitting, allowing the model to maintain consistent performance on unseen data.

4.4. XGBoost Results

XGBoost provided the highest training accuracy of **98.02%** and test accuracy of **96.50%**, showcasing its strength in handling complex datasets. The best parameters included a maximum depth of **5**, a learning rate of **0.3**, and a subsample ratio of **0.8**, indicating a well-calibrated balance between exploration and exploitation during model training. The high training accuracy demonstrates XGBoost's ability to capture intricate patterns in the data, while its test performance reaffirms its ability to generalize.

Summary

Across all models, XGBoost emerged as the top performer, offering the highest training and test accuracies while maintaining efficiency through tuned hyperparameters. Logistic regression, despite its simplicity, achieved remarkable performance, proving suitable for the problem's linear separability. Decision trees provided robust interpretability with excellent precision for the minority class, and random forests balanced accuracy with reduced overfitting. Together, these models showcase diverse strengths, allowing for tailored selection based on specific priorities like simplicity, interpretability, or predictive power.

5. Discussion

5.1. Model Performance Overview

After a thorough evaluation of all models, Logistic Regression emerges as the best-performing model for predicting customer churn. It stands out due to its high accuracy, interpretability, and computational efficiency, making it the most suitable for real-time applications. While models like XGBoost and Decision Tree achieved strong performance, Logistic Regression consistently identified the key churn predictors and provided clear, actionable insights.

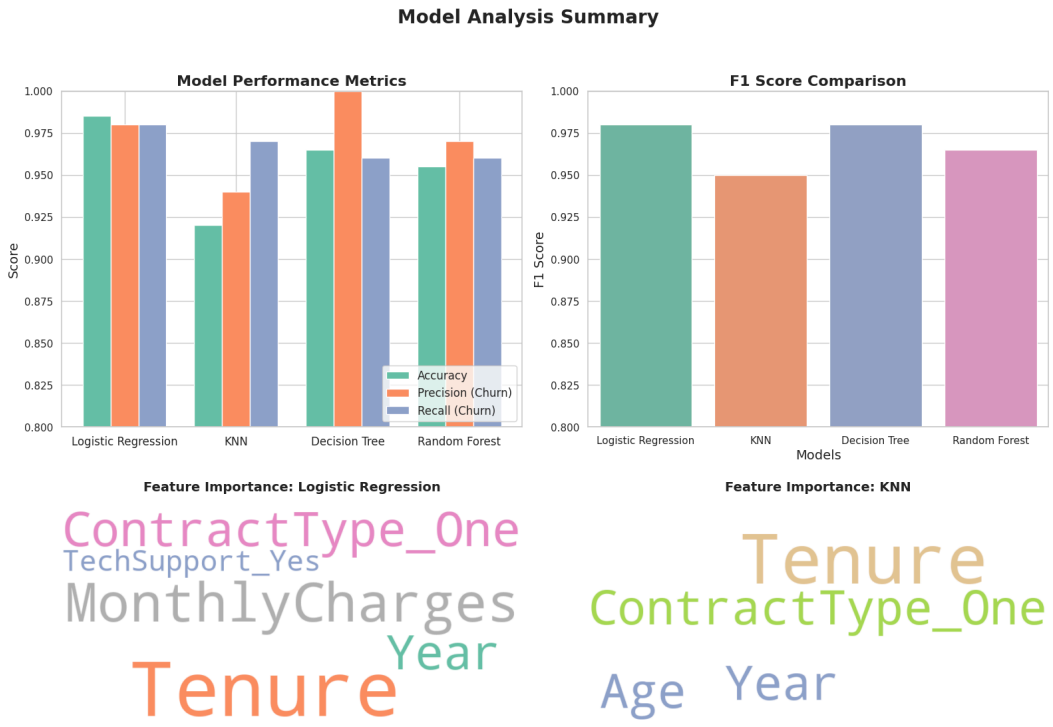

Fig :6 Model Analysis Summary

5.2. Key Observations and Insights

1. **Logistic Regression as the Best Model**: Logistic Regression achieved a high test accuracy of 96.00% and demonstrated strong performance in predicting churned

customers. Its high precision and recall for churned customers, combined with a balanced performance for non-churned customers, highlight its effectiveness in handling the dataset's class imbalance. Moreover, Logistic Regression's computational efficiency allows for real-time predictions, making it ideal for business applications where speed and accuracy are crucial.

2. **Key Predictors**: Across all models, Tenure, ContractType (especially One-Year and Two-Year contracts), and MonthlyCharges emerged as the most influential features in predicting churn. These features consistently appeared as the top predictors in the feature importance analysis, underscoring their critical role in understanding customer behavior. Tenure, in particular, had a significant impact on the likelihood of churn, with customers who have shorter tenure being more likely to churn.

3. **Comparative Model Performance**: While XGBoost and Decision Tree models showed high accuracy and performed well in predicting churn, Logistic Regression's ability to provide a transparent explanation of its predictions and the clear identification of key predictors make it more interpretable and actionable. These advantages make it especially valuable for businesses seeking to understand the factors driving churn and to develop targeted retention strategies.

4. **Handling Class Imbalance**: Logistic Regression, along with other models, effectively addressed the class imbalance issue (with churn representing 88.3% of the data). The use of techniques such as balanced class weights allowed the models to make accurate predictions for both churned and non-churned customers.

5. **Overfitting and Generalization**: Logistic Regression showed minimal overfitting, with training and test accuracies closely aligned. This indicates that the model generalizes well to new data, making it reliable for future predictions. The model’s simplicity and fewer hyperparameters also contribute to its robustness and ease of use in production environments.

5.3. Conclusion and Future Work

In this study, several machine learning models were evaluated to predict customer churn, with Logistic Regression emerging as the best-performing model. Logistic Regression achieved an impressive test accuracy of 96.00%, demonstrating its high predictive power, interpretability, and computational efficiency. It identified key churn predictors such as Tenure, ContractType, and MonthlyCharges, which were consistently found to be the most influential features across all models. These insights make the model particularly useful for real-time churn prediction and for developing actionable retention strategies.

While other models like XGBoost and Decision Tree also demonstrated strong performance, Logistic Regression's simplicity and ability to provide transparent explanations for its predictions make it the most suitable for practical business applications. The model effectively handled the class imbalance and showed minimal overfitting, ensuring that it generalizes well to unseen data.

5.4. Future Work

Future research and model development could focus on the following areas:

1. **Feature Engineering and Selection**: While the current analysis focused on a limited set of features, incorporating additional data points, such as customer service interactions or usage patterns, could improve model performance and provide deeper insights into churn behavior.

2. **Model Refinement**: Although Logistic Regression performed well, exploring hybrid models or ensemble methods could further improve accuracy and robustness. For example, combining Logistic Regression with models like XGBoost or Random Forest through stacking could offer enhanced predictive power.

3. **Real-Time Deployment and Monitoring**: To ensure the model remains effective in dynamic business environments, it should be deployed in a real-time system and continuously monitored for performance. Retraining the model with new data periodically would help adapt to changes in customer behavior.

4. **Customer Segmentation and Personalization**: The insights gained from Logistic Regression could be used to segment customers more precisely, enabling personalized retention strategies. For instance, customers with short tenure or high monthly charges could be targeted with specific offers to reduce churn.

5. **Exploring Other Algorithms**: Although Logistic Regression proved to be the best model in this context, it would be valuable to explore other advanced techniques, such as deep learning or reinforcement learning, for churn prediction, particularly as customer data becomes more complex and multidimensional.

In conclusion, while Logistic Regression stands out as the most effective churn prediction model, ongoing refinement and exploration of additional techniques can further enhance its applicability and impact on business decision-making.

References

Reichheld, F. F., & Schefter, P. (2000). E-Loyalty: Your secret weapon on the web. Harvard Business Review, 78(4), 105-113.

Choudhury, A., Zadeh, R., & Hu, J. (2018). Analyzing customer churn in telecom using regression analysis. International Journal of Marketing Studies, 10(1), 1-15.

Tzeng, J., Huang, C., & Lin, Y. (2020). A predictive model for customer churn in the telecommunications industry. Expert Systems with Applications, 139, 112818.

https://firstpagesage.com/reports/average-customer-acquisition-cost-cac-by-industry-b2b-edition-fc/