# Telecom Churn ML Models Comparison

## Project Overview

This project aims at predicting customer churn, or in other words, detecting the customers who are probably going to give up the company's services. The intention is to make companies take proactive measures to keep customers by learning the key factors that cause churn. After the data of various types and categories were conditioned and turned into numerical values, the same data was split into training and testing sets. The one trained on the chaotically divided data did the prediction through machine learning models, including: **Logistic Regression, Random Forests, Gradient Boosting, and Support Vector Machine (SVM)**

The models were then evaluated one against each other to determine their interpretation in predicting customer turnover. All algorithms were tested with the same data, and their results were compared using measures of performance incorporating accuracy, precision, recall, F1-score, and AUC. This leads to an analysis that assists in recognizing the best-performing algorithm for predicting customer churn and, at the same time, laying down a trustworthy model and framework for business decision-making.

---

## Dataset

In this project, the dataset is based on an understanding of customers and their usage of services. Corresponding to each customer is a row in the dataset which captures the individual's personal particulars, account information, the services they subscribed to, and interaction with the company. The primary features in this dataset include the age, gender, tenure (time the customer has been with the service), monthly and total charges, contract type, internet service type, tech support, and the outcome variable of interest, which is churn or not churn. The actual format for this predictive modeling project will start with exploratory data analysis and then begin to develop models to predict customer churn.

**Dataset Summary:**

- Total Records: **949 customers**
- Average Tenure: **~20 months**
- Average Monthly Charges: **~$74.23**
- Average Total Charges: **~$1479.84**
- Churn Rate: **87.7% customers churned**

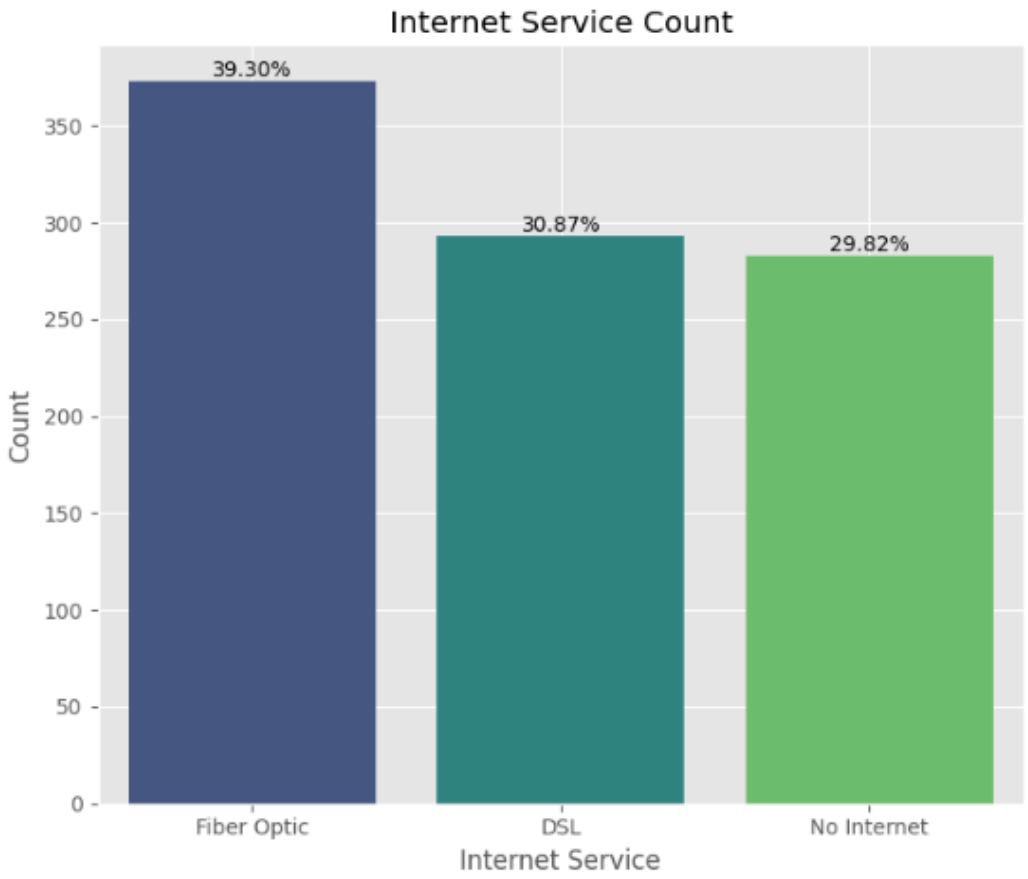
**Kaggle Link:** https://www.kaggle.com/datasets/abdullah0a/telecom-customer-churn-insights-for-analysis

- The dataset underwent data cleaning and preprocessing that included the following actions:
- Dealing with missing values (i.e., replacing missing values for InternetService with “No Internet”).
- Removing rows of invalid tenure equal to 0.
- Encoding categorical variables (e.g., mapping gender and churn to numbers).

---

**Visual Insights**

**Customer Distribution by Internet Service Type**




**Insights:**

The customer base reveals that Fibber Optic is the leading service, followed by DSL, and a large collection of customers do not use any internet service. Approximately 70% of the customer base subscribe to either Fiber Optic or DSL, indicating a solid utilization of internet services, whereas nearly 30% of our customers do not have either service, which represents a decent opportunity for targeted marketing or service improvement. The significant usage of fiber optics indicates

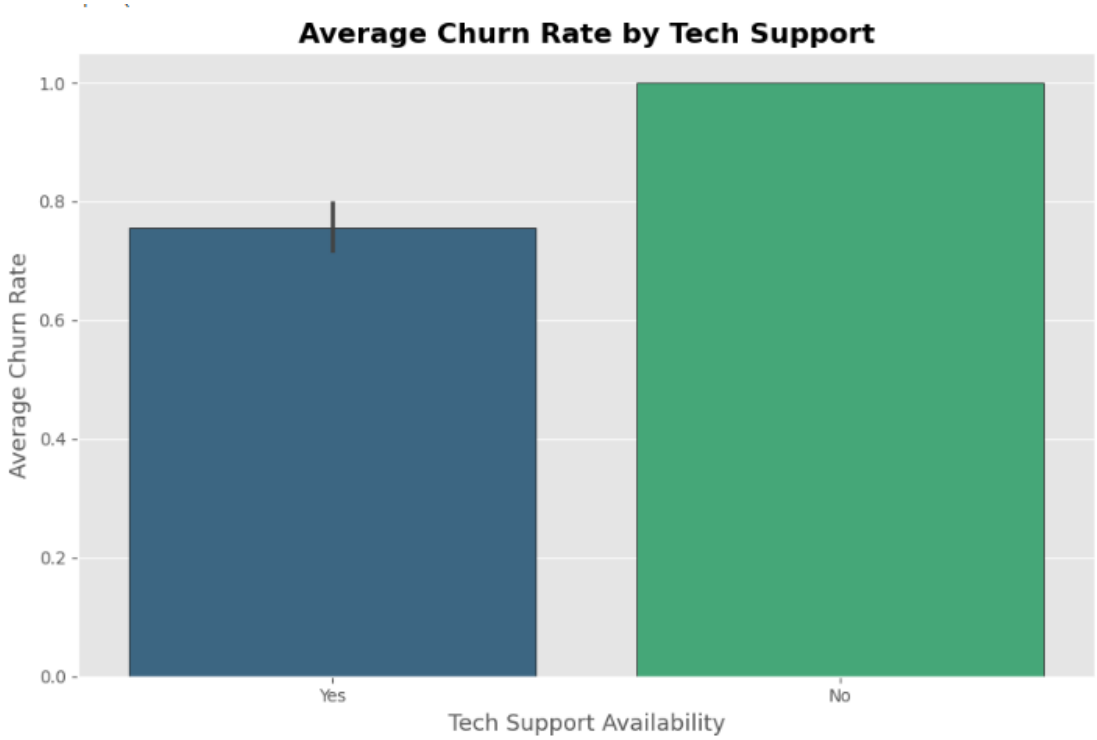
customer value of high-speed internet service; however, it could indicate dissatisfaction with pricing or service performance if they decide to churn.

**Key Highlights include:**

- Fiber Optic is the leading service (≈39.3% of customers)
- DSL has moderate utilization (≈30.9% of customers)
- Approximately 30% of customers do not use any internet service
- In total, approximately 70% of customers use an internet service, indicating market strength
- The No Internet segment may represent a growth opportunity

---

**Average Churn Rate by Tech Support Availability**




**Insights:**

The bar chart represents the average customer churn rate in connection with Tech Support availability. Customers who do not have Tech Support have a churn rate close to 100%. Customers who have access to Tech Support only have a churn rate of around 75%. Therefore, it can be inferred that Tech Support is a factor that adds a protective shield against churn and satisfactory customer ratings. Customers without Tech Support rated a high degree of dissatisfaction and did

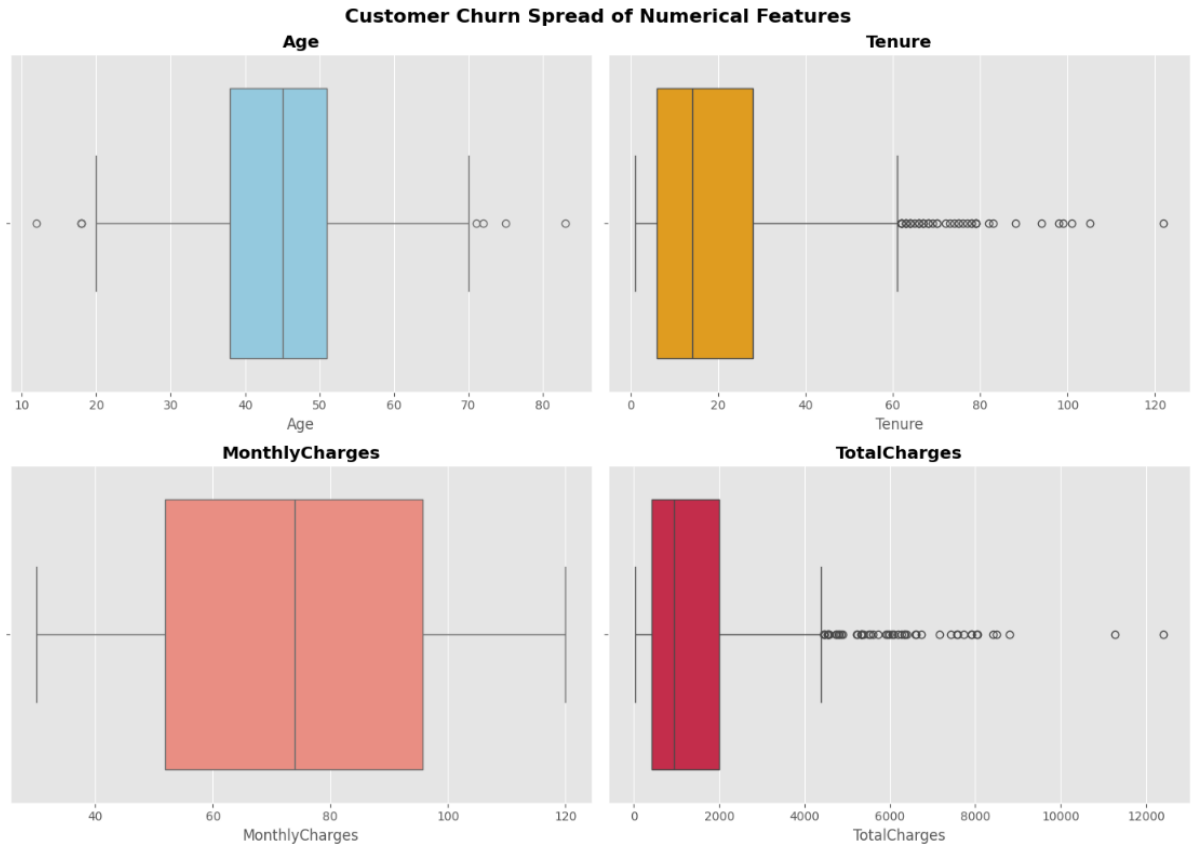
not resubscribe to their premium subscriptions after trial. Overall, and for the reduction of churn rates, having Tech Support easily accessible (and making Tech Support effective in delivery) is important for overall ratings of customer satisfaction.

**Key Highlights include:**

- Customers without Tech Support will always churn.
- Customers with Tech Support have a much lower churn.
- Tech Support availability is an important determinant of customer satisfaction and retention.

---

**Customer Churn Spread Across Numerical Features**




**Insights:**

The box plots provide insight into the distribution of Age, Tenure, Monthly Charges, and Total Charges for the customers and their churn. Most customers ranged between the ages of 35 and 55, and within these age groups, the distribution is relatively balanced. There is no indication of

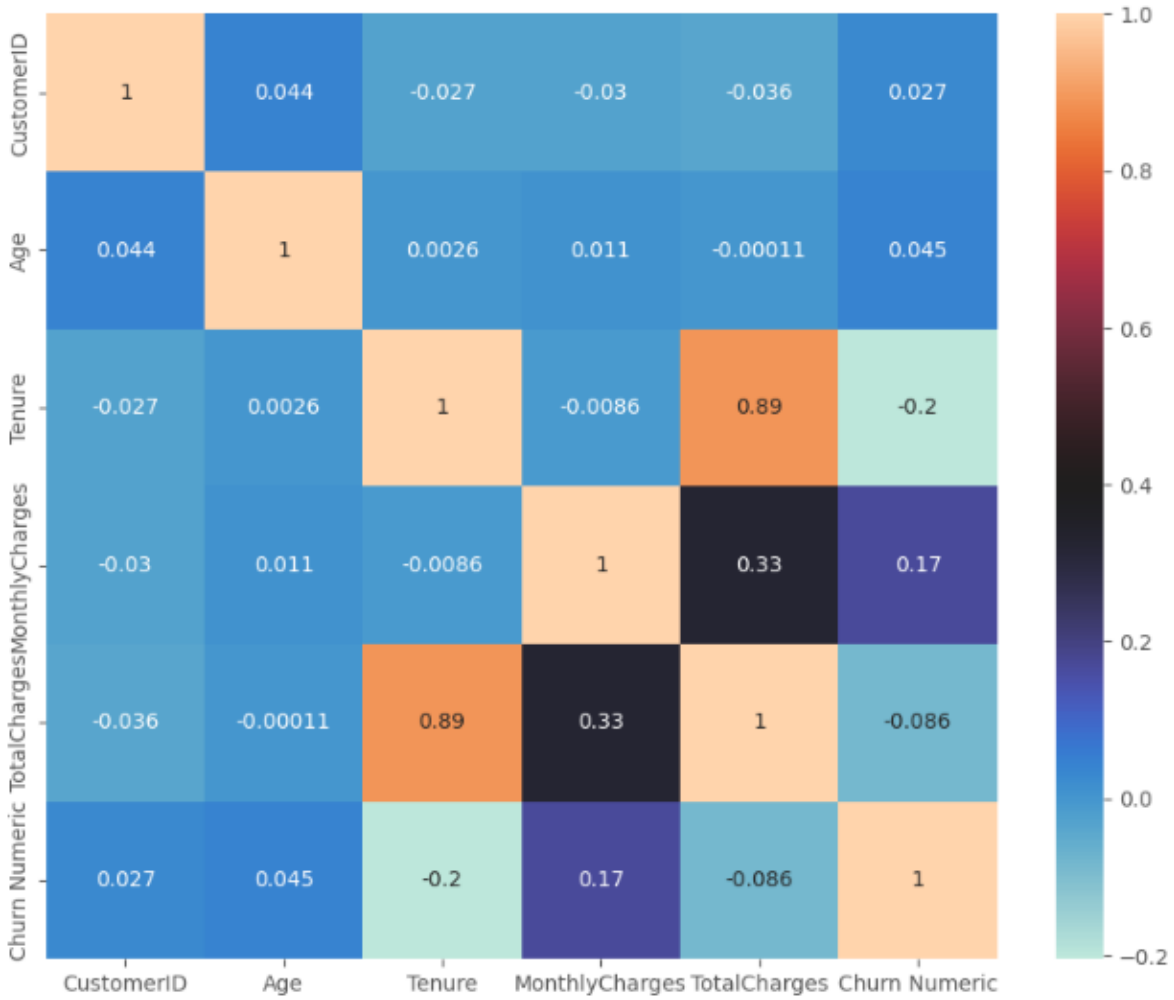
excessive churn occurring in the older age groups. The Tenure feature is right-skewed, inferring that customers who churn are generally shorter in Tenure, while long Tenure customers tend to be more stable.

**Key Highlights include:**

- The Monthly Charges are relatively consistent, amounting to an average of $75; however, the risk of churn may have a slight increase with higher Monthly Charges.
- The Total Charges feature is highly right-skewed, which indicates that the Total Charge is based on the total amount of money spent over time, and in our context, the churned customers typically have lower Total Charges because of the shorter Tenure.
- The data indicates that churn is more likely among customers in earlier stages, along with lower spending customers; these are the retained customers.

---

**Correlation Analysis of Customer Features**



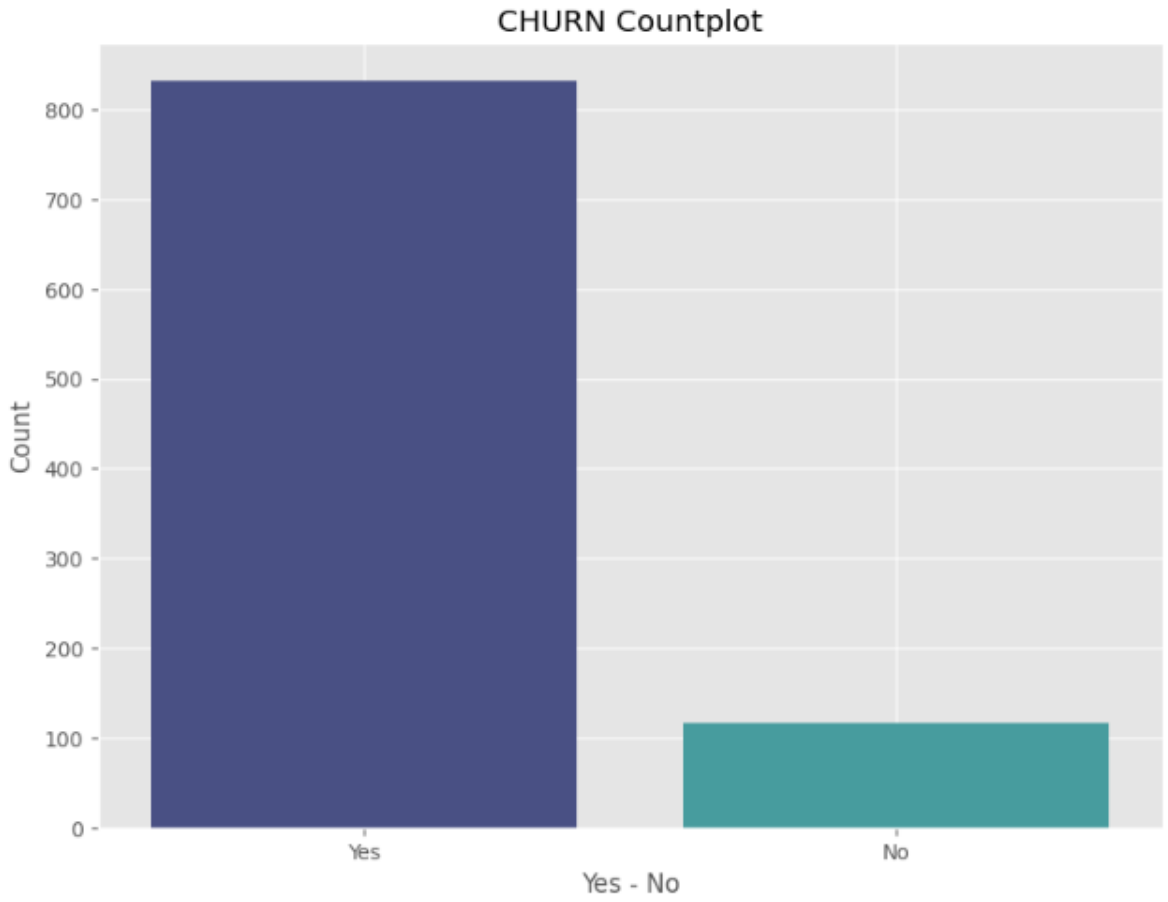
**Insights:**

The correlation heatmap shows the relationships between key numerical features and churn. There is a strong positive correlation between tenure and total charges (0.89), which is expected. Customers with longer tenures will naturally have higher total charges. Monthly charges and total charges are moderately correlated (0.33), which suggests that higher monthly charges would lead to higher total charges over time. There seems to be weak correlations between churn and tenure (-0.20), monthly charges (0.17), and total charges (-0.086). Therefore, these numerical features are not great predictors of churn. Age and customerID are near-zero correlations, which would suggest that other (non-numeric) features, possibly related to service quality, technology support, or the type of contract customers have, better predict customer churn.

**Key Highlights include:**

- Tenure and total charges are highly correlated--a proxy for long-term customer investment.
- Monthly charges may lead to total charges, but do not strongly correlate to churn.
- Churn is only weakly influenced by numeric features--non-numeric features may be better candidates.
- Age and customerID provide little explanatory power and can be eliminated as predictors.
- Non-numeric features, or something service or contract-related, will likely be best for predictive modeling.

**Churn Count Distribution**




**Insights:**

The churn countplot displays the distribution of customers who have left the service ("Yes") compared to customers remaining on the service, referred to as "No." A larger proportion of customers, approximately 830+, have churned, as opposed to about 110–120 who have stayed loyal to the service. This case is presented to illustrate a challenging class imbalance. The high churn rate may imply several potential negatively biased scenarios: poor customer satisfaction with service quality, expensive prices associated with providing service, or ineffective retention strategies. This imbalance in data may also suggest that machine learning models trained on this data may become biased toward predicting churn, a relationship that must be addressed in dataset preparation to ensure appropriate predictive modeling.

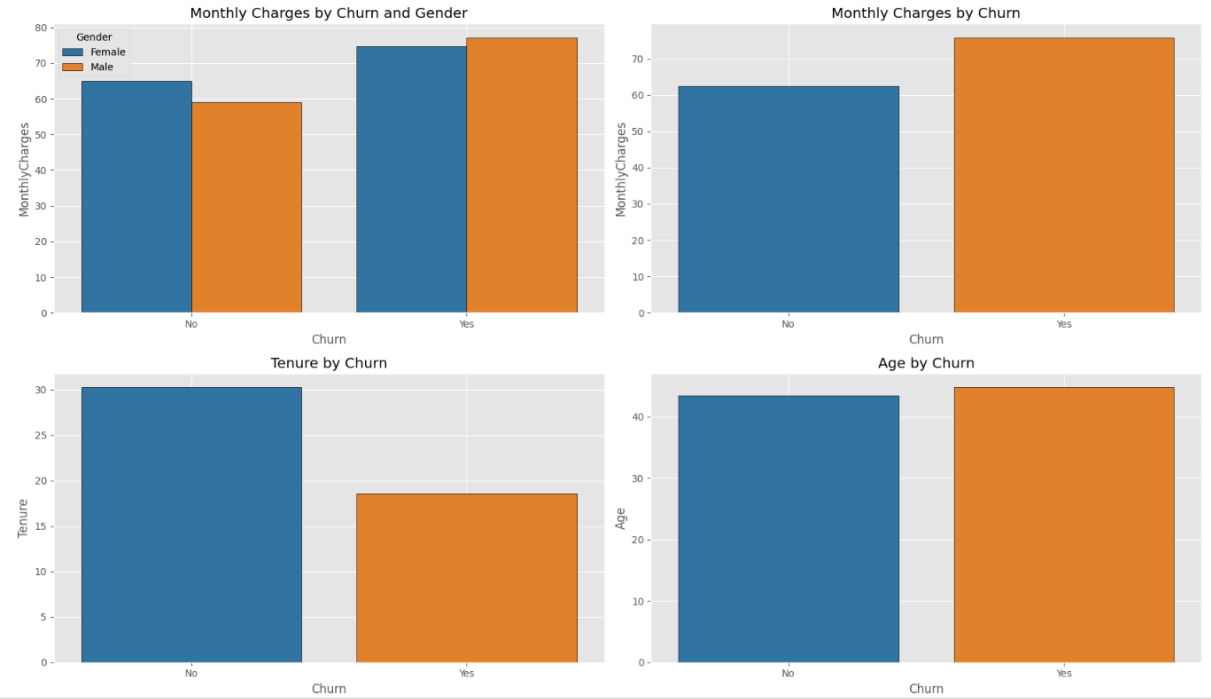
**Key Highlights include:**

- The majority of customers have churned (~830+), which indicates a serious retention issue.
- Only a small fraction (~110–120) of customers have remained loyal.

- High rates of churn may suggest issues with service quality, pricing, or customer satisfaction.
- The imbalance in the dataset may impact the biasing of ML models toward predicting the majority class.

---

**Churn Analysis by Key Customer Factors**




**Insights:**

The churn analysis graphs highlight various ways in which customer personality traits relate to churn behaviors. Customers who churned provided an average of higher monthly billings, showing that pricing is one of the most highly correlated factors for churn. Both men and women exhibited nearly identical churn trends and, therefore, gender is likely not a major contributing factor. Tenure highly influenced churn - customers with shorter tenures are more prone to churn early, while customers with longer tenures remained relatively loyal. In contrast, age did not show differences between churned and retained customers, showing that this category did not appear to be influenced by churn at the time of this study. Overall, price sensitivity and early customer retention trends appear to be the two areas that need the most improvement.

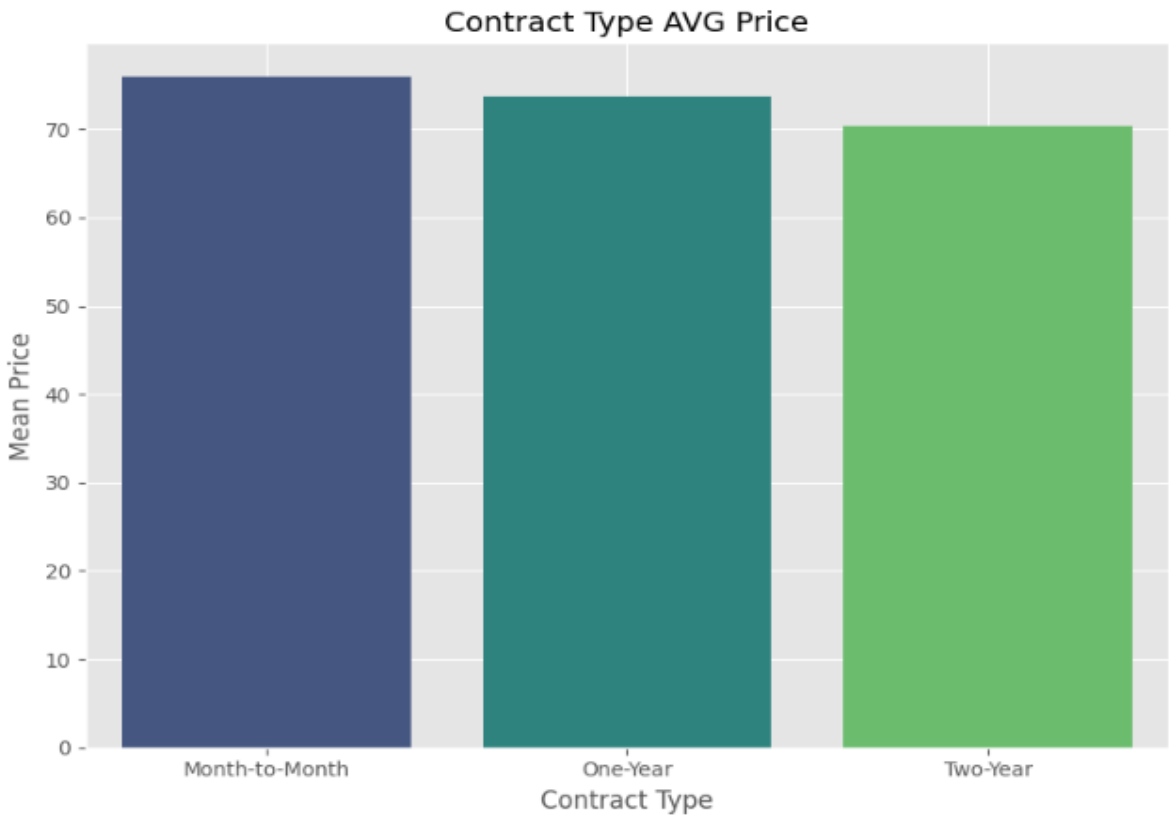
**Key Highlights include:**

- Higher monthly charges are highly correlated with churn.
- Tenure is a significant factor (i.e., newer customers are more actively disengaging!)
- Gender is a less influential factor.

- Age differences do not show a distinction in churned behaviors.
- This suggests that stakeholder organizations should focus on quantifying customers.
- Interventions focused on price adjustments and shortening customer event time and touch points, since those were the contributing factors to churned customers.

---

**Average Monthly Charges by Contract Type**



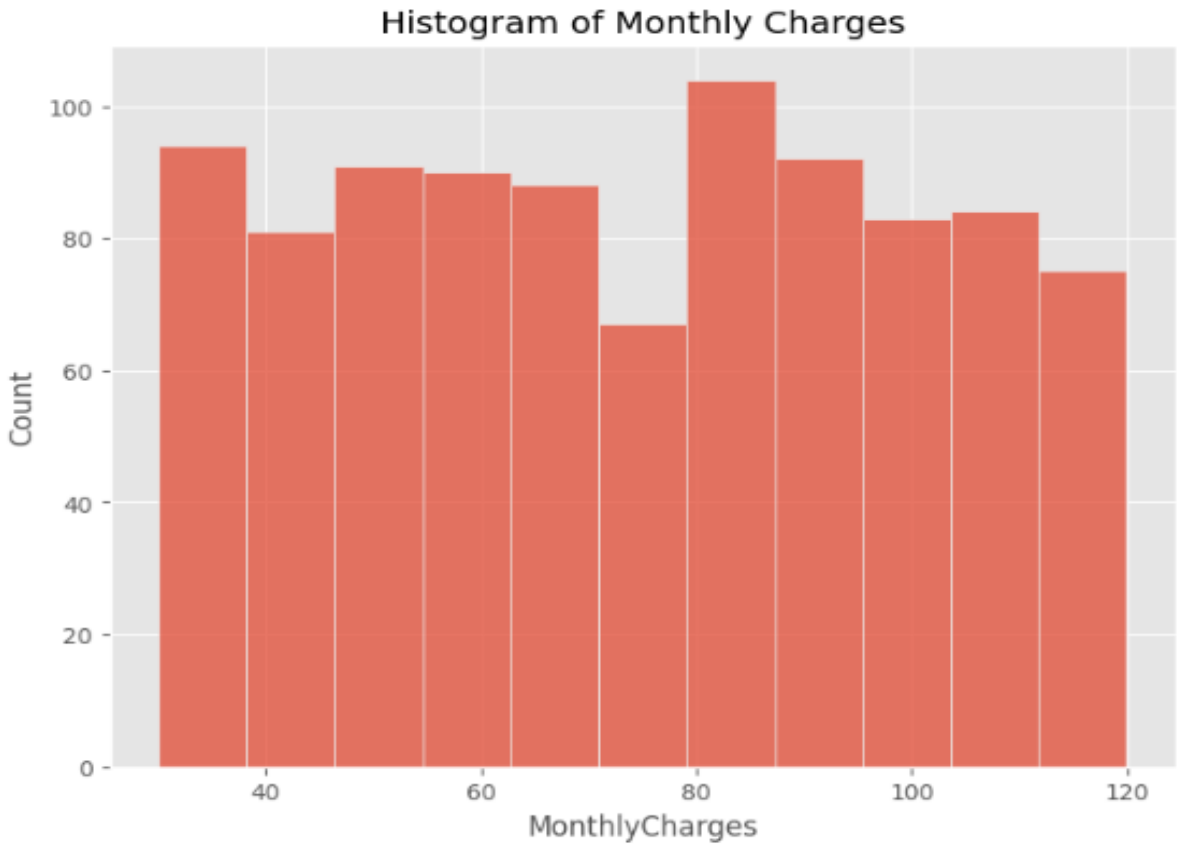

The chart demonstrates that customers on Month-to-Month contracts pay, on average, the most (~$76) per month, compared to One-Year (~$73) and Two-Year (~$70) contracts, which provide continuously less expensive costs. In this way, customers entering longer contracts are provided incentives to do so through pricing, while also retaining customers through longer contracts. Month-to-month customers are given flexibility, but it also may expose them to higher churn risk. Being able to target Month-to-Month customers with an incentive to enter a longer contract could help stabilize or reduce churn, while the customer is generally still satisfied with their change in options.

**Key Highlights include:**

- Average monthly price paid by Month-to-Month customers is the highest (~$76)
- Average monthly prices for One-Year (~$73) and Two-Year (~$70) contracts are the lowest.
- The price incentives based on the length of contracts serve as a retention strategy.
- Month-to-Month contracts manage flexibility but may increase churn risk.
- Targeting Month-to-Month customers with incentives to enter longer contracts may assist with churn stability or reduction.

---

**Distribution of Monthly Charges**




**Insights:**

The graph indicates that customer monthly charges are between $30 and $120, and these charges are spread evenly over this price range. There is a slightly higher concentration on customers who are in this $80–90 charge, most likely indicating that plans or bundles in either the mid-to-high tier range are more popular, suggesting that the overall depiction of customer charges in this histogram would align with these popular tiers. Overall, we see that there is an even spread of charges, indicating that the company had a wide spread of customers that have monthly charges on the lower end and the higher end of the range. Understanding this distribution relative to customer

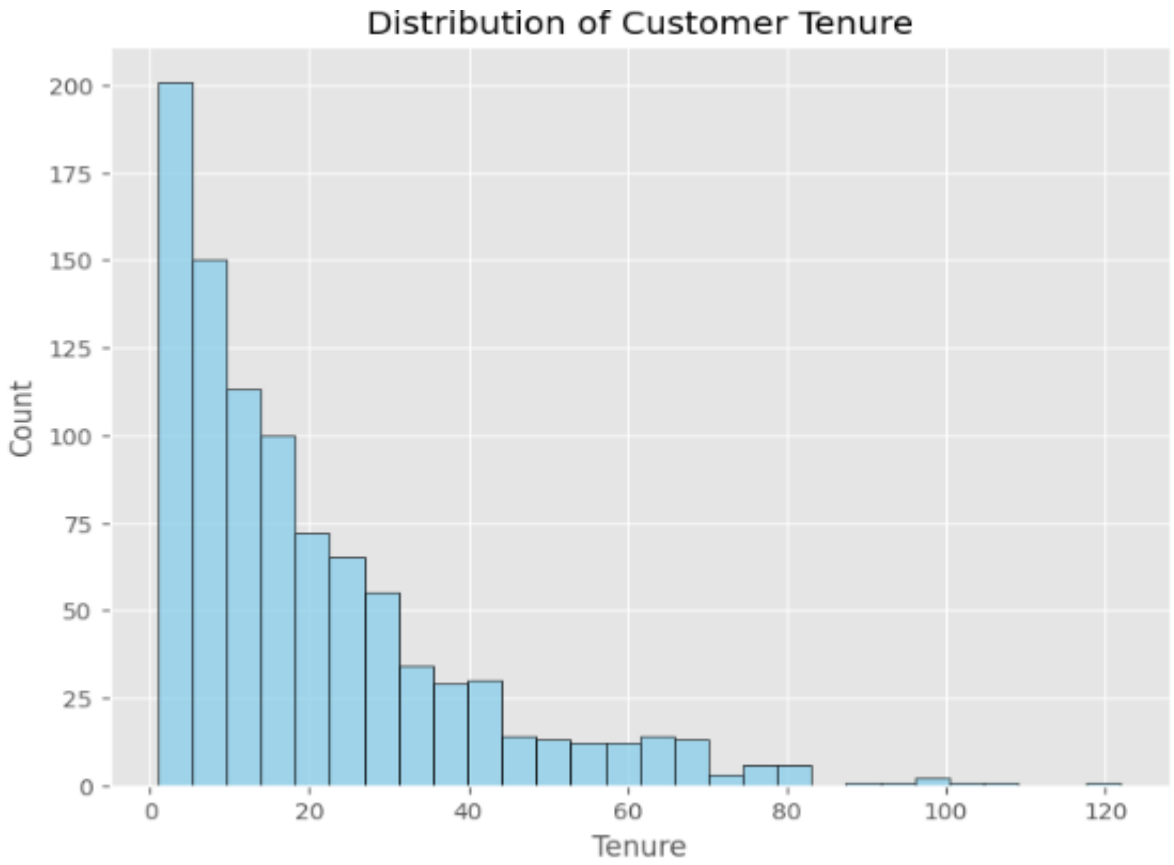
monthly charges will permit some focus for targeting pricing and retention strategies to customers based on their overall spending.

**Key Highlights include:**

- Monthly charges are between $30 and $120.
- Slightly more customers are charging between $80–$90.
- Charges feel evenly spread through this range, representing broad spending behavior.
- Insights could narrow marketing and retention focus relative to price-sensitive spending.
- High-cost plans may require consideration for further reverse churn risk.

---

**Distribution of Customer Tenure**




The histogram demonstrates customer tenure is right-skewed, with most customers having a short duration of tenure, primarily in the 0 to 12-month range. After that, customer tenure decreases steadily, with only a minimal percentage of customers retaining beyond the tenure mark of 60 months. Thus, any long-term loyalty is very limited. Overall, this pattern highlights the company's significant high-intent churn early on, specifically the difficulties encountered in retaining

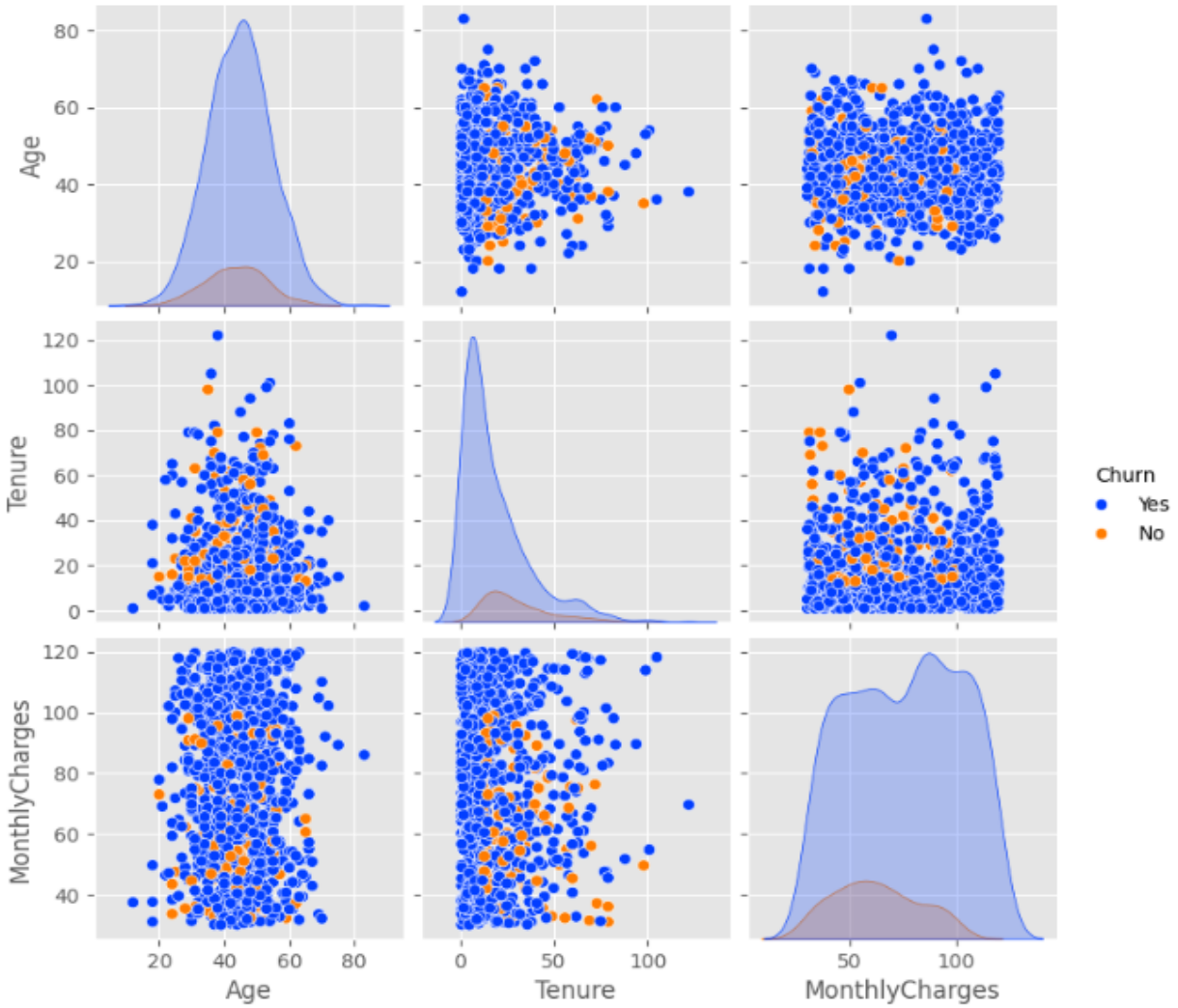
customers within their first year. Developing and executing retention efforts for those very early months could vastly improve total retention, loyalty, and lifetime value.

**Key Highlights include:**

- Most customers have a tenure of 0-12 months.
- Few customers have a tenure longer than 60 months, with even fewer beyond 80 months.
- There is a strong early churn rate; loyalty continues to decline further down the road.
- Retention efforts at early points in the customer lifecycle (onboarding, personalization, loyalty rewards are likely an evergreen opportunity.
- Retaining even a small percentage of short-tenure customers would dramatically decrease churn overall.

---

**Pair Plot of Age, Tenure, and Monthly Charges with Churn**



**Insights:**

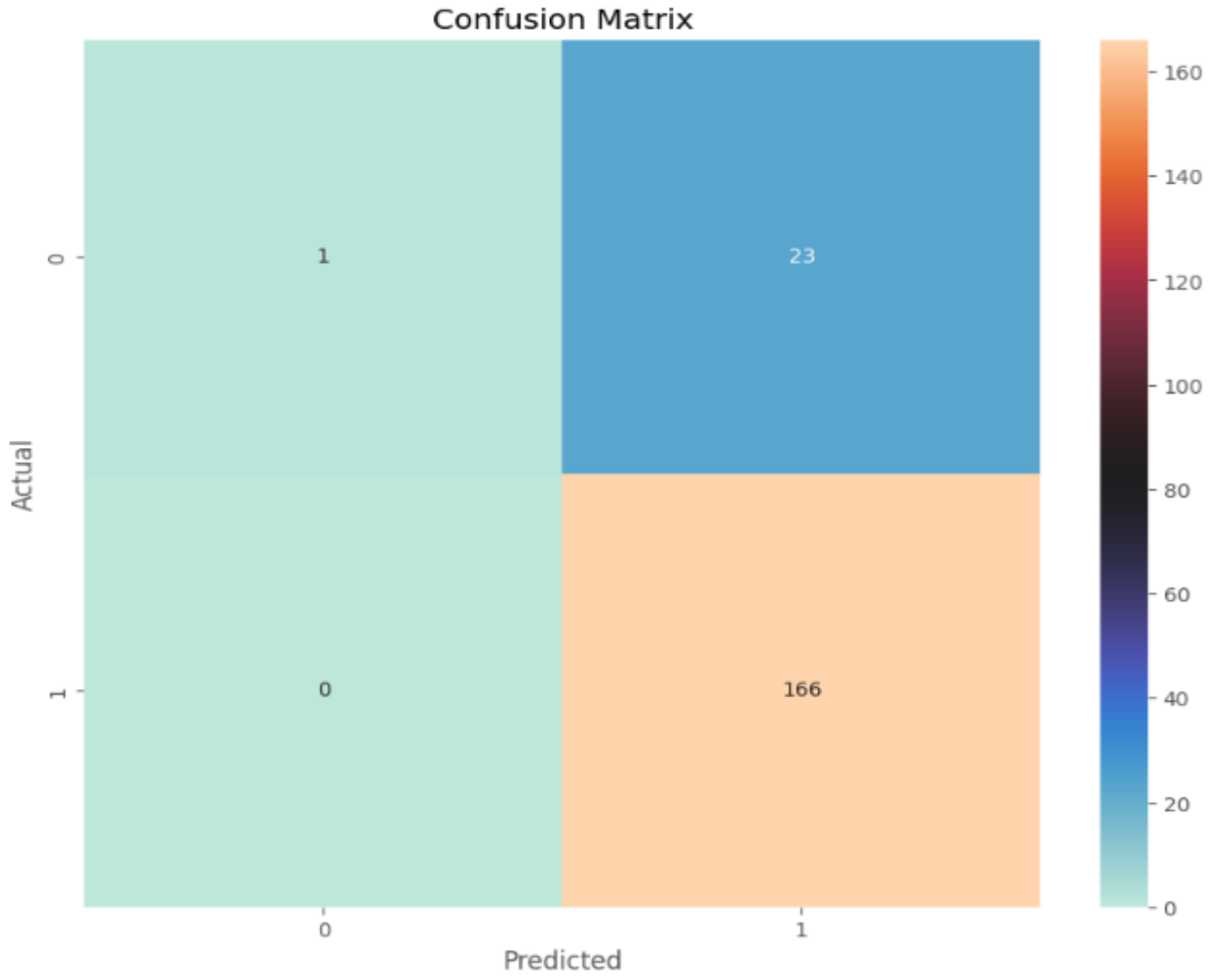
The pair plot provides a visual of the relationships among Age, Tenure, and Monthly Charges with churn status as the response variable. Since Age does not have an obvious relationship with churn, age appears to be evenly distributed across churned vs. retained customers—this indicates that age is not a strong indicator of churn.

Tenure has a seemingly obvious negative relationship with churn, as customers who have comparatively short tenures (less than 20 months) are more likely to leave, while customers with much longer tenures often remain with the service. Monthly Charges also had a positive involvement with the churn decision as a higher paying customer (charges over $80) is more likely to leave than those paying low charges—again, with the caveat that new customers are especially at risk. The cross-variable analysis suggests that a low tenure plus high monthly charges is a reliable way to detect those at a higher risk of churn. Overall, tenure is the strongest relationship with churn, while age seems to have very little effect. Higher charges will also heighten the probability of new customers churning.

**Key Highlights include:**

- Age alone is not a strong predictor of churn.
- Customers that had comparatively short tenures (<20 months) are at the greatest risk of churn.
- New customers are even less likely to leave with high monthly charges (charges over $80).
- The combination of low tenure and high monthly charges highlights the segment at the highest risk of churn.
- Retention for long-tenure customers showing higher charges may have heightened tolerance or loyalty, illustrating that the cost sensitivity is not the entire story.
- Retention efforts should include adjustments to the onboarding experience, potential incentives, and flexibility in their plan for new high-paying customers.

**Confusion Matrix of Churn Prediction Model**




**Insights:**

The confusion matrix assesses the classification model's efficacy in predicting churn by contrasting true with predicted labels. The model has high sensitivity, having detected all churned consumers (True Positives=166 and False Negatives=0), for the ideal recall of 100%. The model struggles to predict True Non-Churners accurately (with 1 True Negative and 23 False Positives); therefore, it tends to over-predict churn in non-churners. The overall accuracy is 87.4%, with 87.8% precision for the churn class, and an F1-score of approximately 93.5%. While the model can successfully identify customers with a high likelihood of churn, it falsely flags other customers as at-risk for churn, leading to greater retention costs for the business.

**Key Highlights include:**

- Model correctly identified all churned consumers (Recall=100%).
- The model struggled to predict True Non-Churners (Only 1 True Negative and 23 False Positives).
- Overall accuracy is 87.4% with 87.8% precision and an F1-score of approximately 93.5%.
- The model does tend to overpredict churn by flagging some invested customers.
- High recall makes the tool useful for retention purposes for customers likely to churn.
- Businesses should factor in unnecessary retention for those false positives into costs incurred.

---

**Model Performance Metrics**

| Metric / Model | Logistic Regression | Random Forest | Gradient Boosting | SVM |
|---|---|---|---|---|
| **Accuracy** | 0.88 | 0.85 | 0.84 | 0.87 |
| **Precision (Churn = 1)** | 0.88 | - | - | - |
| **Recall (Churn = 1)** | 1.00 | - | - | - |
| **F1-Score (Churn = 1)** | 0.94 | - | - | - |
| **AUC Score** | 0.52 | - | - | - |
| **Support (Churn=0 / Churn=1)** | 24 / 166 | - | - | - |

**Insights:**

The evaluation metrics suggest that, of the models examined, Logistic Regression has the best performance in identifying churned customers: accuracy of 0.88, precision of 0.88, recall of 1.0, and 0.94 for F1-score. The high recall signifies that the model identifies 100% of churned customers, which is critical for any retention strategies. The model's low (0.24) recall for non-churn (Churn = 0) indicates the model correctly identifies some but not all loyal customers, likely resulting in some false positive alerts. While there are slight differences between the performance of models, Logistic Regression (0.88) has a slight edge in accuracy over SVM (0.87), Random Forest (0.85), and Gradient Boosting (0.84). Overall, Logistic Regression offers the best model for predicting churn. In addition, the model's AUC score of 0.52 indicates some discrimination ability between churned and non-churned customers, but additional performance improvements may involve acquiring additional features or fine-tuning the model.

**Key Highlights include:**

- Logistic Regression had the best accuracy (0.88) of the models tested.
- There was a recall of 1.0, which indicates all churned customers were correctly identified.
- There was low performance for non-churners, which will likely result in missed loyal customers.
- SVM, Random Forest, and Gradient Boosting performed slightly lower in accuracy (0.84-0.87).
- The AUC score of 0.52 indicates moderate ability to differentiate churn vs non-churn.
- Overall, Logistic Regression predicts churn reasonably well but may require additional features or tuning to decrease false positives.

---

## Project Summary

This project aimed to predict customer churn for a telecommunications company, leveraging different machine learning methods including Logistic Regression, Random Forests, Gradient Boosting, and SVM was undertaken. Key churn factors were identified through extensive explorative data analysis, visualizations, and correlation analysis: short tenure, high monthly charges, no tech support, and the lack of a contract were all considered significantly influential on churn. The dataset was severely unbalanced, with 87.7% of participants having had their service cancelled, demonstrating urgency in interventions designed to retain customers. Logistic Regression was the best performing model, with 88% accuracy, 100% recall of churned customers, and an F1-score (0.94) indicative of this model's ability to identify customers at risk for churn. Overall, this project provided a strong foundation to inform practice and outline a framework to predict customer churn, which ultimately can inform business practices designed to reduce churn and improve customer retention.

- Customers with short tenures and high monthly charges are most likely to be at risk for churn.
- Availability of tech support and longer contracts reduce churn.
- Logistic Regression provided the most stable and reliable model compared to those outlined above.
- Retention strategies, pricing signals, and personalized offerings are most important in the earlier stages of retention to reduce potential churn.
- Overall, non-numeric factors were predicted to impact churn quality of service, contract type, availability of tech support more easily than numeric factors, such as age or total charges.

---