Predicting Customer Churn: A Machine Learning Approach Using Logistic Regression and Random Forest

Table of Contents

1 Introduction

Overview of the project

Keeping existing customers is important for any business. For a company like Telco, having a customer churn is a big problem, because finding a new customer is much harder and more expensive than keeping a current customer happy. This project uses statistical learning techniques to predict which customers are most likely to churn. By analysing customer data, I aim to identify the key factors influencing churn. The findings from this research could help Telco create informed, targeted customer retention strategies, save money, and increase profits.

Objective of the Project

The main goals of this project are:

1. **Predict Churn:** To build a model that can accurately classify whether a customer will churn or stay with the company.
2. **Compare Models:** To train, test, and evaluate several different machine learning models to see which one performs the best for this specific task. The models I will build and compare include:
    - **Logistic Regression:** A model used for binary classification, meaning it predicts the probability of an event belonging to one of two categories (ex. churn/retained)
    - **Random Forest:** A model that combines the predictions of many decision trees to get a more accurate result.
3. **Identify Key Drivers:** To understand why customers leave by identifying the most important factors that influence their decision to churn.

Dataset Description

Source: This analysis uses the publicly available Telco Customer Churn dataset from IBM Sample Data Sets, sourced from Kaggle. The dataset contains 7,043 customer records (observations) with 21 attributes (variables) that detail customer demographics, account information, and the services signed up for.
Link: https://www.kaggle.com/datasets/blastchar/telco-customer-churn

Variables

- **Response:**
    - Churn: A binary indicator of whether the customer has churned or not (Yes or No)
- **Predictors:**
    - **Demographic Information**
        - Customer ID: A unique identifier for each customer.
        - Gender: The customer's gender (Male, Female).
        - Senior Citizen: Indicates if the customer is 65 years of age or older (1 = Yes, 0 = No).
        - Partner: Whether the customer has a life partner (Yes, No).
        - Dependents: Indicates if the customer has children or other dependents (Yes, No).
    - **Account Information**
        - Tenure: The number of consecutive months the customer has been with the company.
        - Contract: The duration of the customer's contract (Month-to-month, One year, Two year).
        - Paperless Billing: Whether the customer has opted for paperless billing statements (Yes, No).
        - Payment method: The method by which the customer pays their bill (Electronic check, Mailed check, Bank transfer (automatic), Credit card (automatic)).
        - Phone Service: Whether the customer subscribes to home phone service (Yes, No).

        - Monthly Charges: The customer's current total monthly charge.
        - Total Charges: The total amount charged to the customer over their entire tenure.
    - **Services Subscribed**
        - Multiple lines: Whether the customer has multiple phone lines (Yes, No, No phone service).
        - Internet Service: Customer’s internet service provider (DSL, Fiber optic, No)
        - Online Security: Whether the customer has online security or not (Yes, No, No internet service)
        - Online Backup: Whether the customer has online backup or not (Yes, No, No internet service)
        - Device Protection: Whether the customer has device protection or not (Yes, No, No internet service)
        - Tech Support: Whether the customer has tech support or not (Yes, No, No internet service)
        - Streaming TV: Whether the customer has streaming TV or not (Yes, No, No internet service)
        - Streaming Movies: Whether the customer has streaming movies or not (Yes, No, No internet service)

2 Data Preprocessing

Data Preparation and Feature Engineering

After understanding what each variable represents, it is important to convert them into the correct data type if not already. This will better represent the nature of the data while also making it suitable for modelling. I started by converting all categorical variables, such as demographic information, service subscriptions, and payment methods, into the proper factor data type.
Next, I cleaned the data to make it simpler and more consistent. I removed the customer ID column because it's unique to each person and will not be useful for prediction. Another important part of this cleaning was simplifying the service columns. The original data had categories like "No internet service" and "No phone service" , alongside a "No" category.

I simplified columns these for three key reasons:

1. **To Remove Redundant Information:** A value of "No internet service" technically just means the customer can't have the service because they lack the main package. For the model's purpose, the important fact isn't the reason, but simply that they do not have the service. This means that combining these cases with "No" strips away this noise.
2. **To Make the Model's Job Easier:** Giving the model three categories where two mean essentially the same thing can weaken its ability to find strong patterns. Simplifying the categories into a clear "Yes" or "No" provides a much stronger and clearer signal for the model to use, which would make its predictions more powerful.
3. **To Improve Interpretation:** The results become easier to explain. It's more impactful to say "Customers without Tech Support were more likely to churn" than to try to explain the difference between "No" and "No internet service."

I also shortened long category names, like changing "Bank transfer (automatic)" to "BankTransferAuto", to make the data easier to work with. Finally, I made sure that all the newly adjusted columns were properly formatted as categories. This entire process of cleaning and organizing the data is important because it directly improves the accuracy, reliability, and explainability of the models that will be used to predict customer churn.

Handling missing values

```
   gender     SeniorCitizen Partner    Dependents     tenure      PhoneService MultipleLines   InternetService
Female:3488   0:5901        No :3641   No :4933   Min.   : 0.00   No : 682     No :4072      DSL       :2421
Male  :3555   1:1142        Yes:3402   Yes:2110   1st Qu.: 9.00   Yes:6361     Yes:2971      FiberOptic:3096
                                                  Median :29.00                              No        :1526
                                                  Mean   :32.37
                                                  3rd Qu.:55.00
                                                  Max.   :72.00

OnlineSecurity OnlineBackup DeviceProtection TechSupport StreamingTV StreamingMovies          Contract    PaperlessBilling
No :5024       No :4614     No :4621         No :4999    No :4336    No :4311        Month-to-month:3875   No :2872
Yes:2019       Yes:2429     Yes:2422         Yes:2044    Yes:2707    Yes:2732        One year      :1473   Yes:4171
                                                                                     Two year      :1695


         PaymentMethod  MonthlyCharges    TotalCharges    Churn
BankTransferAuto:1544   Min.   : 18.25   Min.   :  18.8   No :5174
CreditCardAuto  :1522   1st Qu.: 35.50   1st Qu.: 401.4   Yes:1869
ECheck          :2365   Median : 70.35   Median :1397.5
MailedCheck     :1612   Mean   : 64.76   Mean   :2283.3
                        3rd Qu.: 89.85   3rd Qu.:3794.7
                        Max.   :118.75   Max.   :8684.8
                                         NA's   :11
```

Figure 1 Summary Statistics and Missing Data Assessment for the Telco Customer Churn Dataset: A table displaying summary statistics and the count of missing values (NAs) for each variable in the initial dataset. The TotalCharges variable contains 11 missing entries, representing less than 0.2% of the total records. All other variables are complete.

As seen in Figure 1 above, there are 11 missing values in the TotalCharges column. Since this is only a very small part of the whole dataset (0.156%), I decided to remove those 11 rows from the dataset. This way, the information used to build the model is complete for every customer, which makes the results more trustworthy.

Preprocessed Dataset

After the data preprocessing the dimensions of our dataset are: 7,032 rows (customers) by 20 columns (variables).

Below are the variables of the preprocessed dataset:

- **Response Variable**
  - Churn: Whether the customer left the company (Yes or No). This is what we want to predict.
- **Demographic Information**
  - gender: Whether the customer is male or female.
  - SeniorCitizen: Whether the customer is a senior citizen (Yes or No).
  - Partner: Whether the customer has a partner (Yes or No).
  - Dependents: Whether the customer has dependents (Yes or No).
- **Account Information**
  - tenure: Number of months the customer has been with the company.
  - Contract: The type of contract (Month-to-month, One year, Two year).
  - PaperlessBilling: Whether the customer uses paperless billing (Yes or No).
  - PaymentMethod: The customer's payment method (BankTransferAuto, CreditCardAuto, ECheck, MailedCheck).
  - MonthlyCharges: The amount charged to the customer monthly.
  - TotalCharges: The total amount charged to the customer.
- **Service Information**
  - PhoneService: Whether the customer has phone service (Yes or No).
  - MultipleLines: Whether the customer has multiple phone lines (Yes or No).

- InternetService: The type of internet service (DSL, FiberOptic, No).
- OnlineSecurity: Whether the customer has online security (Yes or No).
- OnlineBackup: Whether the customer has online backup (Yes or No).
- DeviceProtection: Whether the customer has device protection (Yes or No).
- TechSupport: Whether the customer has tech support (Yes or No).
- StreamingTV: Whether the customer has streaming TV (Yes or No).
- StreamingMovies: Whether the customer has streaming movies (Yes or No).

3 Exploratory Data Analysis (EDA)

Demographic Information

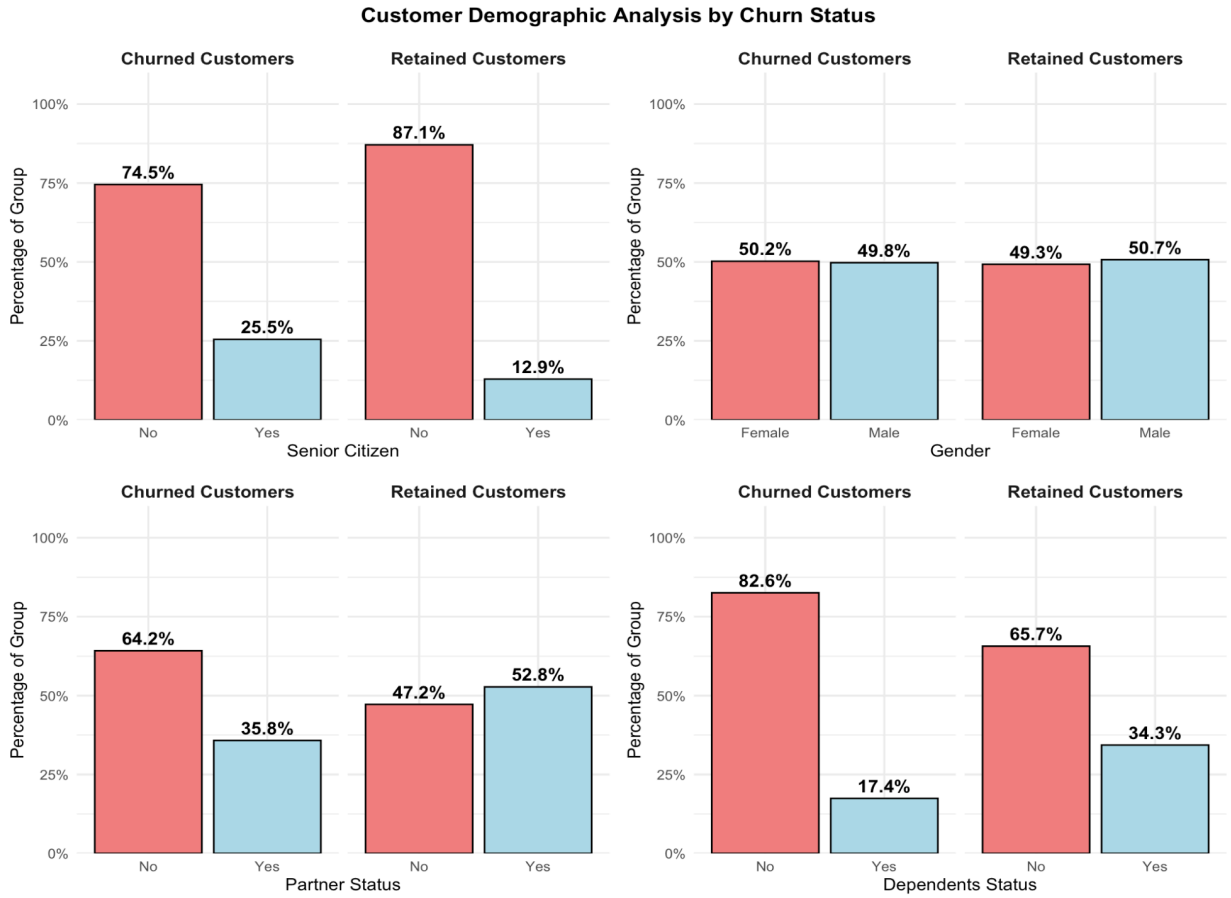

Figure 2 Churn Rates by Key Demographic Factors: A preliminary comparative analysis of customer churn across demographic segments. The plots reveal that customers who are Senior Citizens, lack a Partner, or have no Dependents exhibit significantly higher churn rates. Gender appears to have a negligible effect on churn.

Figure 2 highlights several key demographic trends:

- **Senior Citizen Status:** Seniors constitute 25.5% of churned customers versus 12.9% of retained customers, suggesting that older customers are disproportionately likely to churn. This may reflect affordability concerns or difficulties with technology adoption.
- **Gender:** The churn distribution is nearly identical across genders (≈50% male/female in both churned and retained groups). This implies that gender is not a significant predictor of churn in this dataset.

- **Partner Status:** 64.2% of churners have no partner, while retention is stronger among partnered individuals (52.8%). This suggests that household stability and shared service needs may mitigate churn risk.
- **Dependents:** A substantial 82.6% of churners have no dependents, compared to 65.7% among retained customers. This may suggest that customers with dependents may perceive greater reliance on stable services, which would lead to reduced churn likelihood.

Account Information

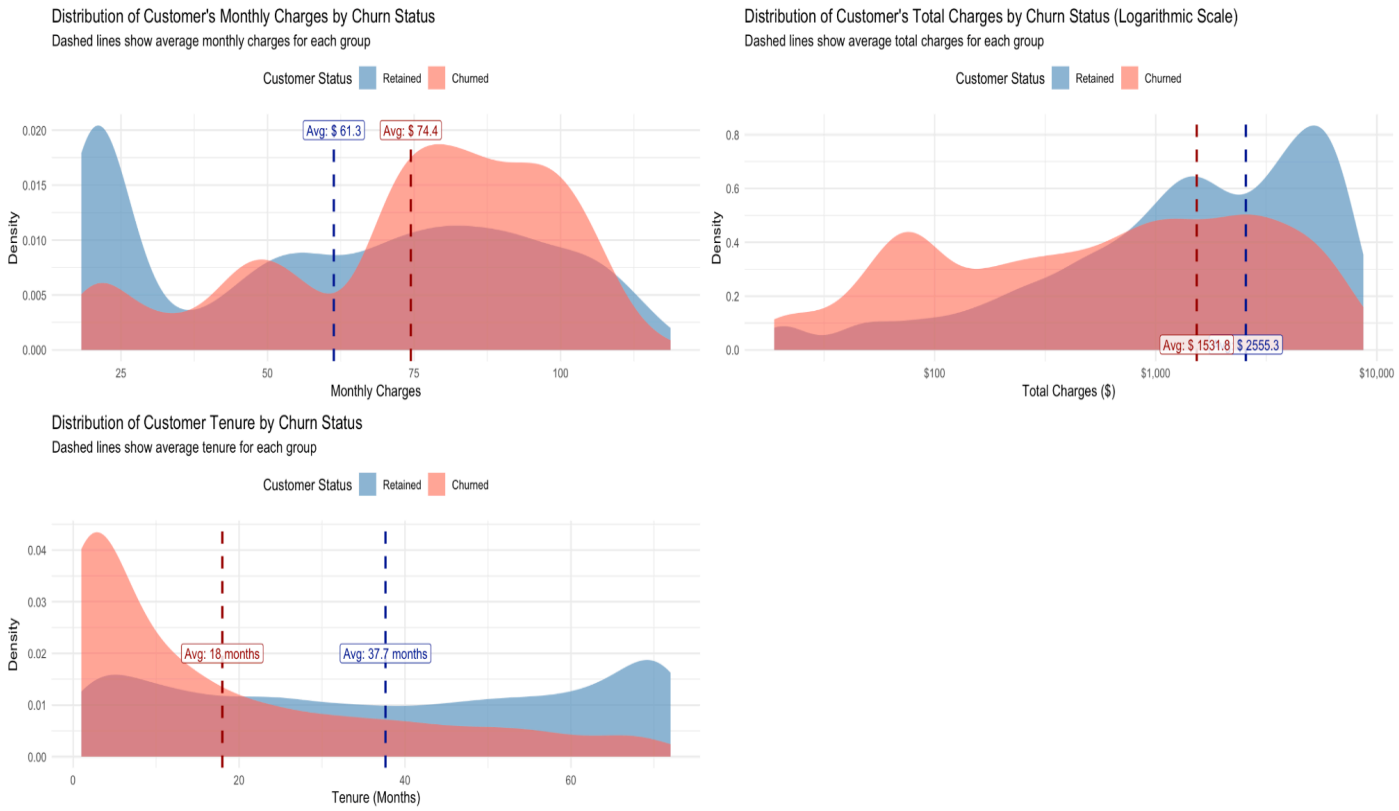

Figure 3 Distributions Densities of Continuous Variables by Churn Status: Density plots comparing the distribution of Monthly Charges, Total Charges, and Tenure for churned versus retained customers. Vertical dashed lines show group averages.

Figure 3 highlights several key billing and tenure trends:

- **Monthly Charges:** Churned customers have a higher average monthly charge ($74.4) compared to retained customers ($61.3). The distribution shows that higher service costs are associated with increased churn likelihood, suggesting that pricing sensitivity plays a role in attrition.
- **Total Charges:** Retained customers accumulate higher average total charges ($2555.3) compared to churned customers ($1531.8). This reflects tenure effects: long-term customers naturally accumulate more total charges, while churned customers exit earlier, limiting their total spend.
- **Tenure:** Average tenure for retained customers is 37.7 months, nearly double that of churned customers (18 months). The distribution highlights that short-tenure customers are disproportionately represented in the churned group, which shows that customer retention risk is highest early in the lifecycle.

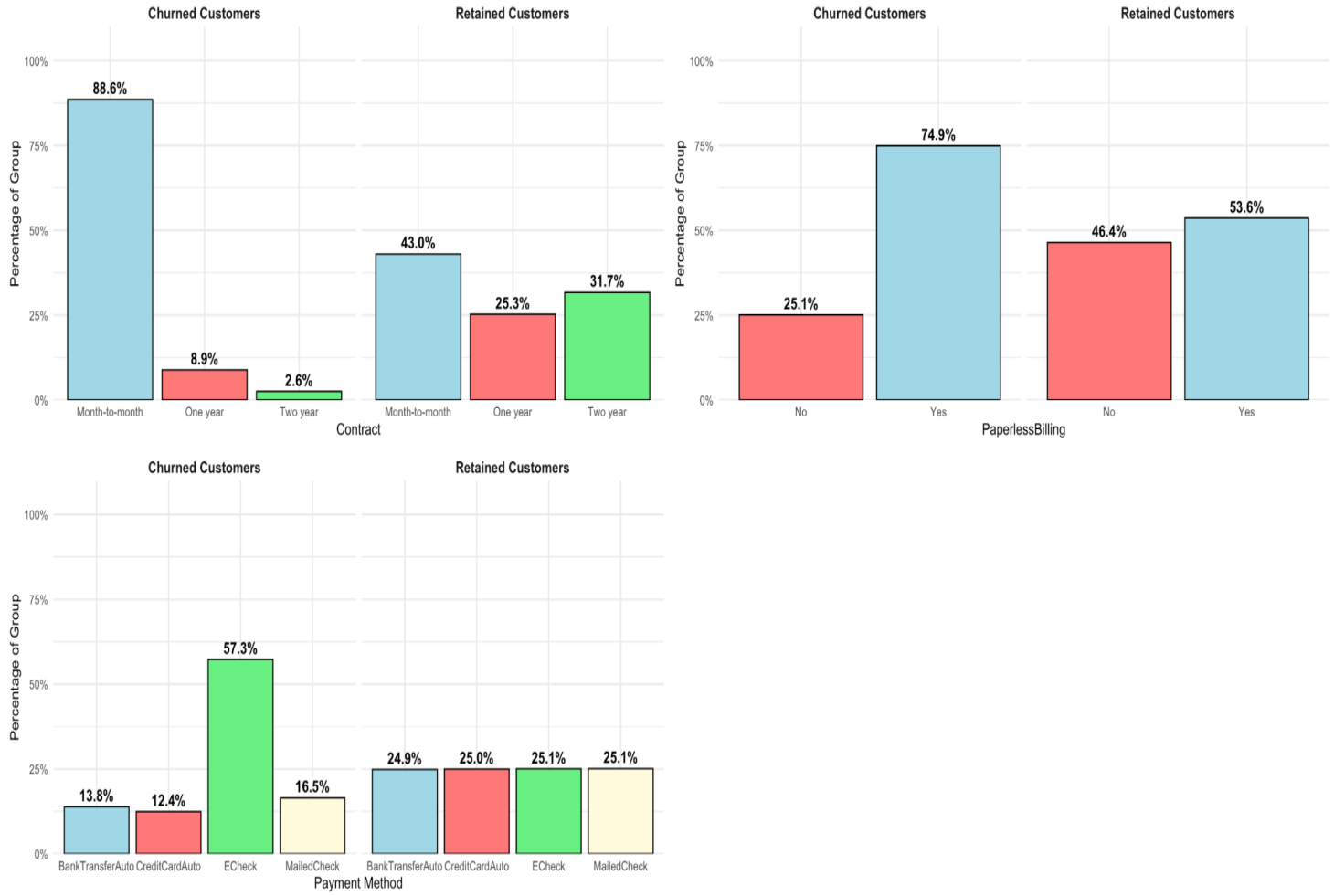

Figure 4 Churn And Retention Prevalence by Contract, Billing, and Payment Method: Bar plots showing the proportion of customers who churned within each category of account information.

Figure 4 highlights several key trends among contracts, billings, and payment methods:

- **Contract Type:** 88.6% of churners are on month-to-month contracts, with long-term contracts (1-2 years) show much lower churn rates. This suggests that locking in customers with longer contracts reduces churn significantly.
- **Paperless Billing:** 75% of churners use paperless billing, compared to ~54% of retained customers. This could reflect that paperless billing is more common among short-term/month-to-month customers, which overlaps with higher churn risk.
- **Payment Method:** E-Check customers churn the most (57.3%), while automatic bank/credit card transfers have much lower churn. This may be because automatic payments are easy and convenient, making people less likely to switch.

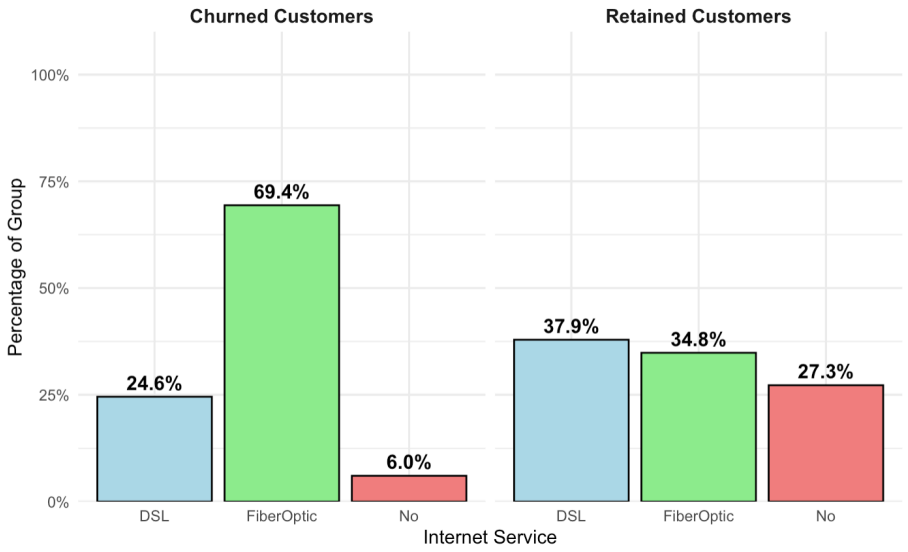

Figure 5 Internet Service Profile by Churn Status: Bar plots showing the percentage of customers with DSL, Fiber Optic, or No internet service. The churned group is characterized by a high prevalence of Fiber Optic subscriptions.

Figure 5 suggests a strong link between internet service type and customer churn. A large majority (69.4%) of customers who left used fiber optic internet, which is far higher than those with DSL (24.6%) or no internet (6.0%). In contrast, the group of customers retained is much more balanced; similar percentages use DSL (37.9%) and fiber optic (34.8%), with a notable portion (27.3%) not using internet service at all. This indicates that fiber optic users are disproportionately likely to churn.

Service Information

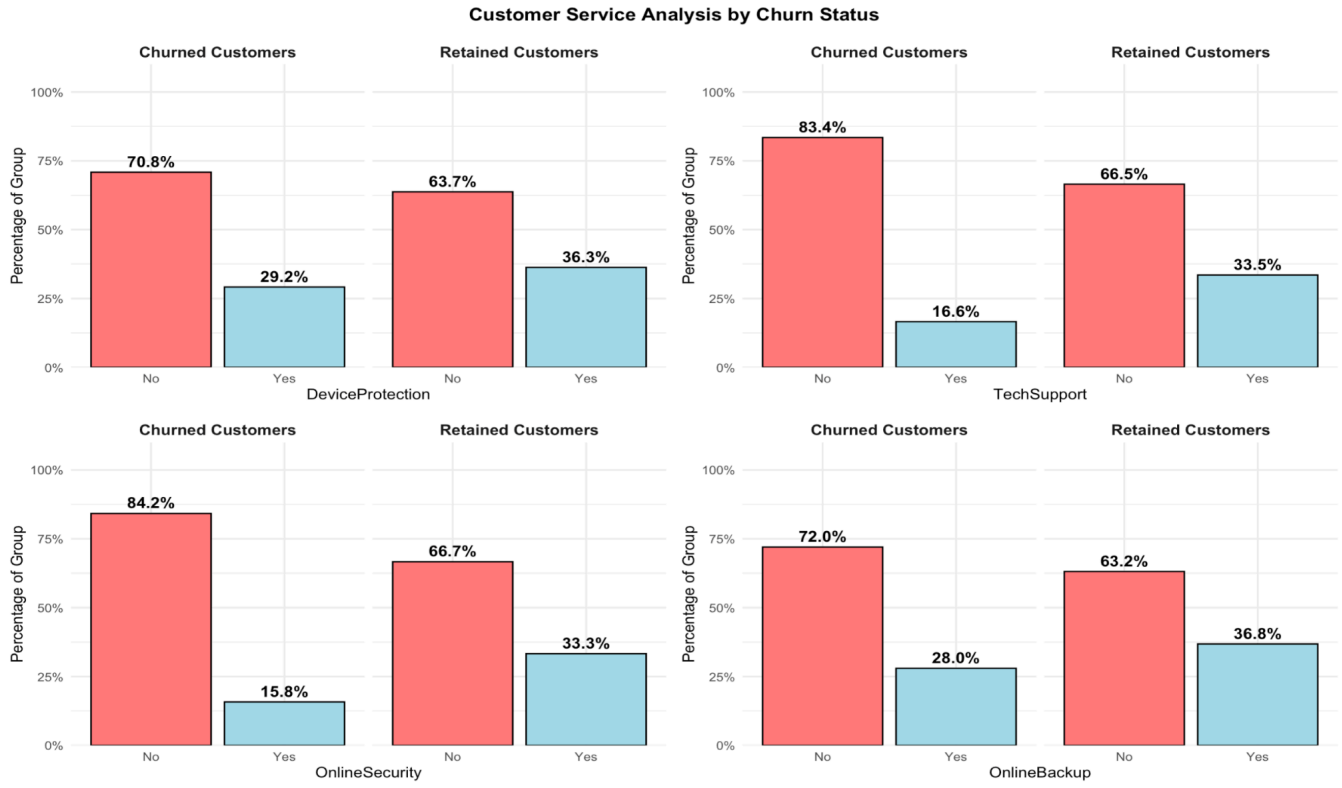

Figure 6 Impact of Support Services on Churn: Bar plots showing the percentage of customers with and without key service add-ons (Online Security, Tech Support, Online Backup, and Device Protection) in churned and retained customers.

Figure 6 highlights some core drivers of churn:

- **Lack of Online Security, Tech Support, Device Protection, and Online Backup:** Churned customers overwhelmingly did not have these services: 84.2% (Online Security), 83.4% (Tech Support), 70.8% (Device Protection), and 72.0% (Online Backup) of churners did not subscribe to these. The percentages are significantly lower among retained customers, suggesting these services are major retention factors.

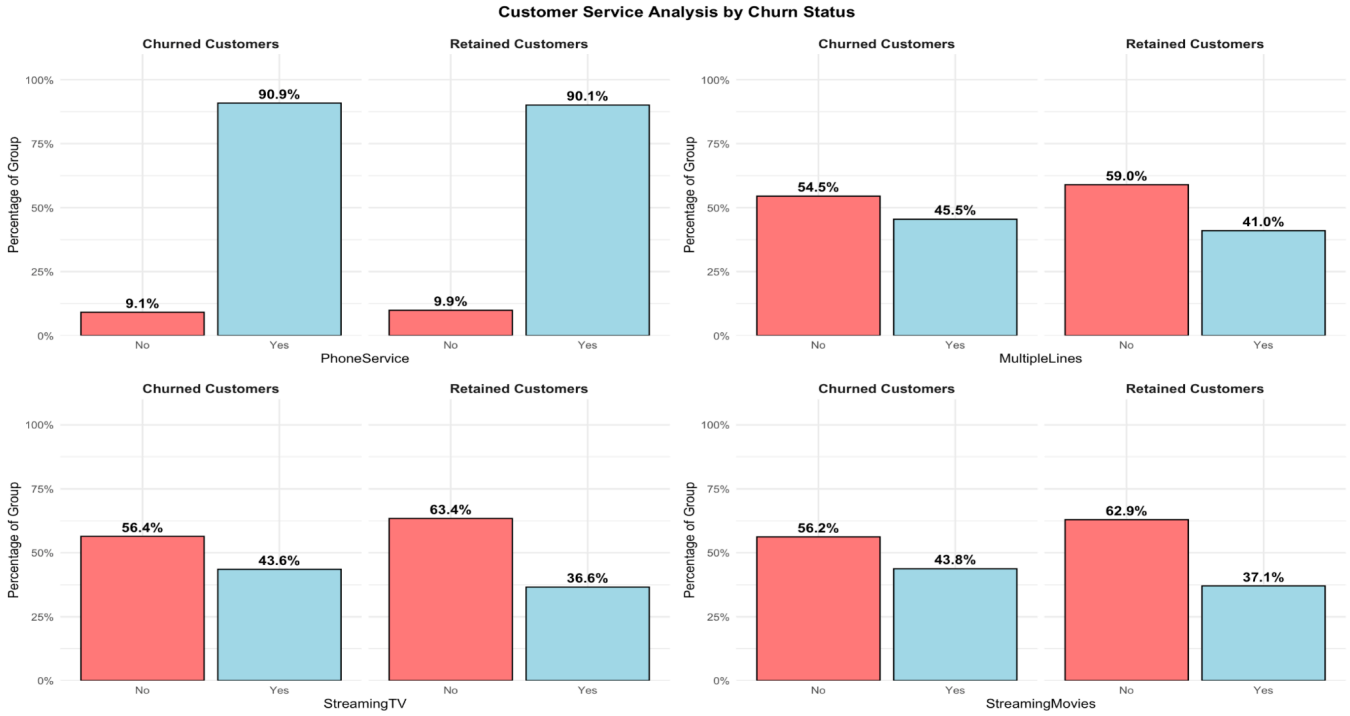

Figure 7 Entertainment and Phone Service Profile by Churn Status: Bar plots showing the proportion of customers with various entertainment and phone services. The distributions are relatively similar between groups, indicating a weaker association with churn.

Figure 7 suggests that the following service in a weak or moderate indicator of churn:

- **Phone Service:** Nearly all customers (churned and retained) had phone service, so its presence alone is not a strong churn predictor.
- **StreamingTV and StreamingMovies:** Churn rates are higher among those without streaming services, but the difference is much smaller compared to security/support services. 56–63% of churners lacked StreamingTV or StreamingMovies, compared to higher retention among those subscribed, but the gap is not as large.
- **Multiple Lines:** More churned and retained customers have no multiple lines, but the difference between churned (54.5%) and retained (59.0%) is modest, which may suggest only a minor association.

4 Logistic Regression Modeling Approach

4.1 Model Rationale

The first analytical technique used is LASSO (Least Absolute Shrinkage and Selection Operator) Logistic Regression. This method is particularly well-suited for this task because it will predict customer churn and automatically select the most relevant features by shrinking less important coefficients to zero using L1

regularization. This produces a simpler, more interpretable model that reduces overfitting and improves generalizability to new data. The process involves:

1. **Shrinking coefficients:** Reduces the model's complexity.
2. **Performing feature selection:** Forces the coefficients of less important variables to exactly zero, which removes them from the model.
3. **Improving generalizability:** Helps create a model that is more likely to perform well on new, unseen data.

4.2 Meeting Model Assumptions

The following key assumptions for logistic regression were assessed and addressed:

1. **Linearity of Continuous Variables:** The relationship between continuous predictors and the log odds of churn was assessed. It was found that the variable MonthlyCharges had a non-linear relationship with the log odds of churn. As shown in Figure 8 below:

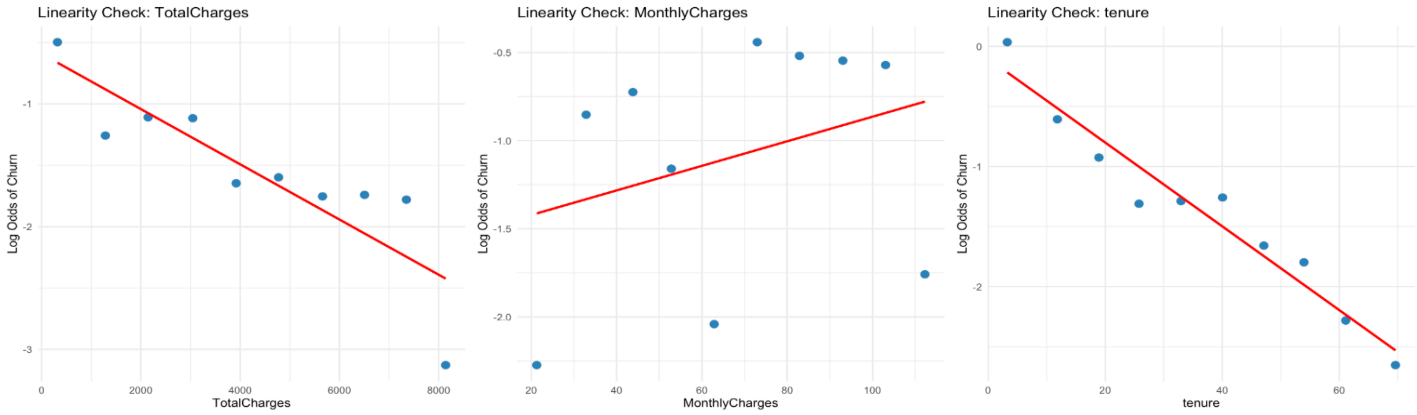

Figure 8 Assessing the Linear Log-Odds Assumption for Continuous Predictors: Scatterplots of continuous variables against the log-odds of churn, used to check the linearity assumption for logistic regression. While Tenure and Total Charges show a reasonably linear relationship, Monthly Charges shows a clear non-linear pattern. This violation justified the transformation of Monthly Charges into a categorical variable for the final model.

To solve this, the MonthlyCharges var was binned into a categorical variable (MonthlyCharges_cat: Low, Medium, High, Very High), which showed a more linear relationship with the log odds of churn and improved model interpretability. The linearity is now shown in Figure 9:

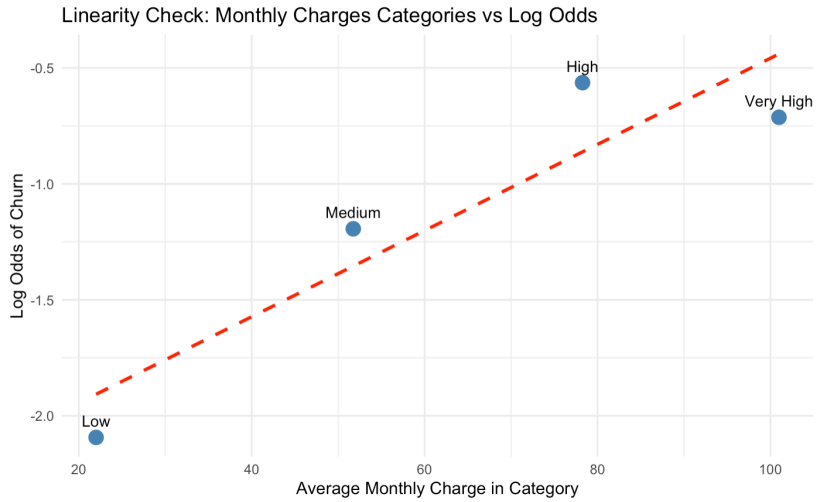

Figure 9 Log-Odds after Categorization of Monthly Charges: The relationship between the binned average of the new MonthlyCharges_cat variable and the log-odds of churn. Converting the continuous Monthly Charges variable into a categorical variable successfully resolved the non-linearity violation observed in Figure 8.

2. **Multicollinearity:** LASSO regression helps manage multicollinearity by selecting one variable from groups of correlated predictors. For confirmation, the model was still checked for multicollinearity after LASSO-based variable selection. The Variance Inflation Factor (VIF) scores were all below 5, which confirmed that multicollinearity was not a concern in the final model.
3. **Outliers:** An analysis of the continuous variables (tenure, MonthlyCharges, TotalCharges) confirmed that there were no extreme outliers that could incorrectly influence the model.
4. **Binary Dependent Variable:** The churn outcome is binary (Yes/No), which fulfills this assumption
5. **Independence of Observations:** The data confirms that the observations are independent of each other.
6. **Sample Size:** The dataset contains 7,032 rows, which is suitable for binary logistic regression.

4.3 Model Training and Variable Selection

- The cleaned dataset was split into a training set (80%) and a testing set (20%).
- To prevent the model from being biased towards the majority class, the ROSE (Random Over-Sampling Examples) technique was applied to the training set to create a balanced dataset for model training.
- A LASSO regression was fit using 10-fold cross-validation on the training data to determine the optimal penalty parameter λ.
- The lambda.1se value (a higher value within one standard error of the minimum) was chosen for the final model. This value makes the model simpler and more efficient by using fewer predictors without losing much accuracy

The LASSO algorithm selected the following variables as the most important predictors of churn:
tenure, Contract, InternetService, PaperlessBilling, PaymentMethod, SeniorCitizen, and OnlineSecurity.

A standard logistic regression model was then refit using only these selected variables on the balanced training data.

4.4 Results and Interpretation

The final model's coefficients were changed into Odds Ratios (OR) to make them easier to understand. An OR greater than 1 means an increase in the odds of churning, while an OR less than 1 means a decrease.

Key Findings:

- **Tenure**: OR = 0.98. For each additional month of tenure, the odds of a customer churning decrease by about 2%. This makes loyalty a strong protective factor against churn.
- **Contract Type**: These were among the strongest predictors.
    - **Two Year Contract (OR = 0.41)**: Having a two-year contract reduces the odds of churning by 59% compared to a month-to-month contract.
    - **One Year Contract (OR = 0.70)**: Reduces the odds of churning by 30%.
- **Internet Service (**InternetServiceFiberOptic**)**: OR = 1.94. Customers with Fiber Optic internet have 94% higher odds of churning compared to those with DSL, this could potentially be due to cost or competition.
- **Payment Method (**PaymentMethodECheck**)**: OR = 1.33. Using an Electronic Check increases the odds of churning by 33% compared to other methods.

4.5 Model Evaluation

The model was evaluated on the held-out test set, which was not balanced to accurately represent the real-world class distribution.

Performance Metrics:

- **Accuracy**: 0.74
- **Sensitivity (Recall)**: 0.79 - The model correctly identified 79% of all customers who actually churned.

- **Specificity**: 0.72 - The model correctly identified 72% of all customers who did not churn.
- **Precision**: 0.51 - When the model predicts churn, it is correct 51% of the time.
- **F1-Score**: 0.62

Interpretation: The model demonstrates a good ability to identify customers who are likely to churn (high recall), which is the primary goal of a churn prediction model. The moderate precision score is expected because of the inherent class imbalance; even a well-tuned model will have false positives when the event of interest, which is churn in this case, is rarer. The model works well overall and can be used to help focus efforts on keeping customers from leaving.

4.6 Conclusion and Business Implications

The LASSO logistic regression model successfully identified key drivers of customer churn. The most significant factors are the type of contract and the customer's tenure.

Recommended Actions:

1. **Promote Long-Term Contracts:** Incentivize customers on month-to-month plans to switch to one-year or two-year contracts.
2. **Targeted Interventions for High-Risk Groups:** Proactively engage with customers who have fiber optic service, use electronic checks, and have paperless billing. Offer them personalized incentives or support to address their potential dissatisfaction.
3. **Loyalty Programs:** Develop programs that reward tenure to further reduce the churn risk among long-standing customers.

Overall, this model provides a data-driven foundation for a customer retention strategy, which will allow the company to allocate resources efficiently and effectively to reduce churn.

5 Random Forest Model Approach

5.1 Model Rationale

While logistic regression provides a highly interpretable model, ensemble methods like Random Forest often have better predictive performance because they can capture complex, non-linear relationships and interactions within the data. The Random Forest approach is an ensemble learning method that works by building many decision trees during training and outputting the mode of the classes (for classification) of the individual trees. This addresses the risk of overfitting that is common in single decision trees and often results in high accuracy. The Random Forest algorithm was chosen for its ability to:

1. **Handle Non-Linearity:** It makes no assumptions about linear relationships between variables, which allows it to model complex patterns that logistic regression might miss.
2. **Manage Feature Interactions:** It automatically detects and leverages interactions between variables.
3. **Rank Feature Importance:** It provides a strong and reliable measure of which variables are most influential in predicting the outcome.
4. **Resist Overfitting:** By averaging the results of many decorrelated trees (built on bootstrapped samples and random feature subsets), it generalizes very well to new data.

5.2 Addressing Class Imbalance

As with the previous model, the original training data was highly imbalanced. To prevent the Random Forest from developing a bias towards the majority class ("No"), the ROSE (Random Over-Sampling Examples) technique was

also applied to the training set to create a balanced dataset for model training. The class distribution after ROSE was confirmed to be approximately even (No: 2834, Yes: 2793).

5.3 Variable Selection and Importance

An initial Random Forest model with 500 trees was built using all variables. The model's built-in feature importance metric, Mean Decrease in Gini, was used to identify the most predictive variables. The Gini index measures node impurity; a higher mean decrease indicates a variable that is more important for accurately splitting the data across all trees.

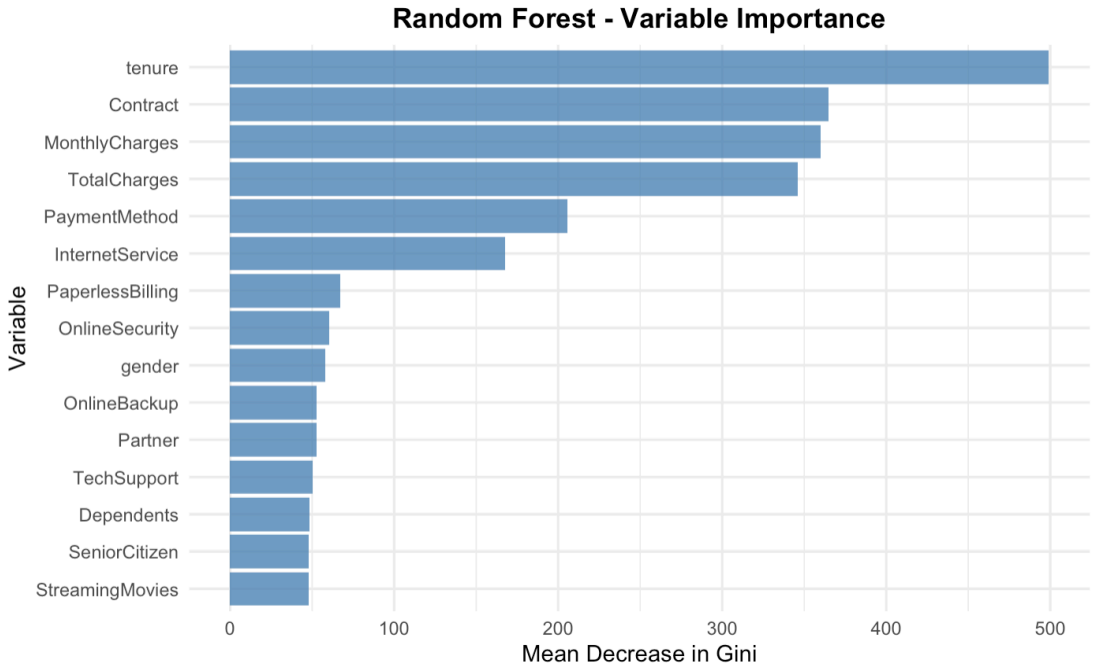

Figure 10 Random Forest Variable Importance Ranking: Predictive drivers of churn as determined by the Mean Decrease in Gini impurity. Tenure, Contract type, and Monthly Charges are the most influential factors in the model, which confirms the key insights from the exploratory data analysis.

The top 10 most important variables were selected for the final model to optimize performance and reduce complexity:
tenure, Contract, MonthlyCharges, TotalCharges, PaymentMethod, InternetService, PaperlessBilling, OnlineSecurity, gender, OnlineBackup.

5.4 Hyperparameter Tuning

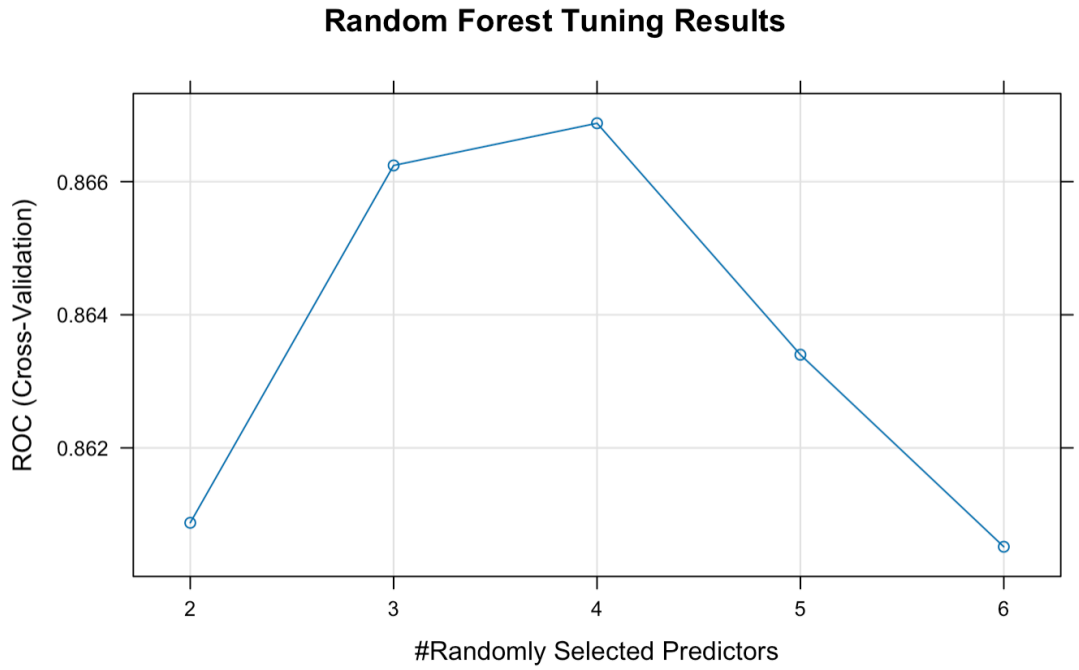

Figure 11 Selecting the Optimal mtry Hyperparameter: The performance of the Random Forest model across a range of mtry values. The peak AUC score at mtry = 4 shows that considering four variables at each node split produces the most generalizable and accurate model for this dataset.

To maximize the model's performance, the hyperparameter mtry was tuned using 5-fold cross-validation:

- mtry: The number of variables randomly sampled as candidates at each split. A grid of values (c(2, 3, 4, 5, 6)) was tested.
- The tuning process was managed using the caret package, with the Area Under the ROC Curve (AUC) as the evaluation metric. The cross-validation results identified mtry = 4 as the optimal value for the final model.

5.5 Results and Interpretation

The final Random Forest model was built with 500 trees and the tuned parameter (mtry = 4).

Out-of-Bag (OOB) Error Estimate: The model's OOB error rate was 24.84%. This is an unbiased estimate of the generalization error and is a key advantage of Random Forests, because it is computed internally during training without the need for a separate validation set.

Variable Importance: The analysis confirms the findings from the logistic regression model, while also adding more complex relationships and rankings:

1. **Tenure:** The most important predictor by a significant margin.
2. **Contract Type:** The second most important predictor, with long-term contracts being strongly protective.
3. **Monthly & Total Charges:** Financial metrics are critical, with their relationship to churn being complex and non-linear, which the Random Forest captures effectively.

5.6 Model Evaluation

The final model was evaluated on the untouched, imbalanced test set to simulate real-world performance.

Confusion Matrix Results:

	Actual: No	Actual: Yes
Predicted: No	646 (True Negatives)	54 (False Negatives)
Predicted: Yes	386 (False Positives)	319 (True Positives)

Performance Metrics:

- Accuracy: 0.687
- Sensitivity (Recall): 0.855
- Specificity: 0.626
- Precision: 0.452
- F1-Score: 0.592
- AUC (Area Under ROC Curve): 0.821

Interpretation: The Random Forest model does very well at its primary task, which is identifying customers who will churn. Its high Sensitivity (Recall) of 0.855 means it correctly identifies 85.5% of all customers who actually churn. This is better than the logistic regression model (0.786). However, the Precision is moderate (0.452), which means that there are many false positives. This is often an acceptable trade-off in churn prediction. The goal is to cast a wide net to make sure that very few at-risk customers are missed; the resulting list of flagged customers can then be prioritized by their predicted probability. The high AUC score of 0.821 confirms the model has excellent overall discriminatory power.

5.7 Conclusion and Business Implications

The Random Forest classifier is a highly effective tool for predicting customer churn, outperforming the logistic regression model in key areas like sensitivity and overall discriminative power (AUC).

The strategic advantages of this approach include:

- **High Recall:** It is exceptionally good at finding most of the customers who are likely to leave, which will make sure that retention campaigns have a comprehensive target list.
- **Actionable Insights:** The variable importance output provides a clear, data-backed priority list for the business, confirming that customer tenure and contract type are the strongest factors for reducing churn.

Recommended Actions:

1. **Deploy for Proactive Engagement:** Use this high-recall model to generate a list of customers with a high risk of churning. These customers should be the main focus of retention efforts.
2. **Prioritize within Predictions:** Since the model gives a risk score for each customer, prioritize those with the highest risk to make the best use of resources.
3. **Focus on Key Drivers:** Focus on what the model shows are important factors, such as encouraging long-term contracts and creating loyalty programs that reward customers who stay longer.

In summary, the Random Forest model provides a powerful model getting the information necessary to create a data-informed customer retention strategy.

6 Overall Conclusion and Recommendations

This project built and compared the two strong models LASSO logistic regression and Random Forest to predict customer churn for Telco. From cleaning the data to building and testing the models, the process provided strong classifiers and important insights into why customers leave.

6.1 Summary of Findings

The analysis from both the Logistic Regression model and the Random Forest identified a core set of factors that drive customer churn:

1. **Tenure and Contract Type:** The single strongest predictor of churn is low customer tenure. This risk is compounded by month-to-month contracts, which have higher churn rates compared to one or two-year agreements. This shows that customer loyalty is earned early and solidified through long-term commitments.
2. **Service Cost and Value Perception:** Customers with higher MonthlyCharges, particularly those with Fiber Optic internet, are at a significantly higher risk of churning. This suggests issues with perceived value, price sensitivity, or competitive pressure in the more premium service tier.
3. **Customer Experience:** The payment method is another telling indicator. Customers using Electronic Checks churn at a higher rate, this may be due to the manual effort required each month compared to the convenience of automatic payments.

6.2 Model Comparison

Each model offers its own distinct advantages:

- **LASSO Logistic Regression** is the best tool for explanation and targeted action. Its coefficients provide a clear, quantifiable measure of how each factor increases or decreases churn risk. This will be ideal for understanding the "why" behind customer behavior and for designing specific interventions.
- **Random Forest** is a great tool for prediction and for creating a list of customers to target. Its higher Recall (0.855) means it is very effective at identifying the vast majority of customers who are likely to leave. This makes sure that only a few at-risk customers are missed.

Recommendation: The best thing to do would be to use the information from both of these models. Use the Random Forest to generate a complete list of high-risk customers, and then use the insights from both the Logistic Regression and Random Forest to prioritize and tailor the retention strategies for different segments on that list.

6.3 Recommendations

1. **Implement a Tiered Onboarding and Loyalty Program:**
    - **New Customers:** Introduce strong incentives for new customers to sign 1 or 2-year contracts immediately. For example, Telco can offer a discounted rate or a waived activation fee as a signing bonus.
    - **Existing Month-to-Month Customers:** Launch a targeted campaign offering existing month-to-month customers a compelling reason to switch to an annual contract, such as a guaranteed rate lock.
    - **Long-Tenure Customers:** Create a formal loyalty program that rewards tenure with services such as exclusive discounts, priority support, or annual service credits to really reinforce their decision to stay.

2. **Revise the Value Proposition for High-Cost Services:**
    - For Fiber Optic and other high-monthly-cost customers, look into conducting satisfaction surveys to understand if their expectations are being met.
    - **Bundle Essential Services:** To improve the perceived value, it would be good to look into bundling critical retention services like OnlineSecurity with high-cost plans like Fiber Optic, rather than offering them as costly add-ons.
3. **Proactive Retention Campaigns:**
    - Use the Random Forest model to routinely score all customers and flag those with a high probability of churning.
    - For these high-risk customers, use targeted interventions such as personalized offers or a call from a retention specialist to better understand their situation.
4. **Reduce Friction in the Customer Payment Experience:**
    - **Incentivize Automatic Payments:** Encourage customers to switch from manual payments (checks) to automatic payments by for example offering a small monthly discount of $5 off. This will reduce the monthly "touchpoint" where a customer might decide to cancel.

In conclusion, this project demonstrates that by using machine learning, Telco can transition from a reactive to a proactive business model. By understanding the reasons behind customer churn and predicting who is most at risk, the company can allocate their resources efficiently, preserve revenue, and build a more stable, loyal customer base. The information that was gained from the analysis can serve as a strong and data-driven foundation for a successful customer retention strategy.

7 Limitations

While this analysis gives useful insights and strong prediction models, it is important to recognize its limits to better understand the results and plan improvements.

7.1 Limitations

- **Temporal Snapshots:** The data is only a single snapshot in time and so it does not track changes over a customer's lifetime. For example, there is no information for when services are added or removed or when changes to payment methods are made. This makes it hard to model how churn happens over time.
- **Absence of Key Business Metrics:** The data does not include important information like:
    - **Customer Service Interactions:** The number of calls to support, complaint tickets filed, or customer satisfaction scores are often leading indicators of churn.
    - **Competitor Offers:** A major missing external factor is having a customer churned due to a better offer from a competitor.
    - **Price Changes:** There is no information on whether a customer experienced a recent price hike, which is a common trigger for churn.
- **Inherent Class Imbalance:** Even though techniques like ROSE were used to balance the training data, the underlying reality is that churn is a relatively rare event. This imbalance can limit the maximum achievable precision. This means that even a small number of false positives can result in a large list of customers to contact, which impacts resource allocation.
- **Causality vs. Correlation:** The models identify strong correlations, but they cannot definitively prove causation. For example, the analysis shows that paperless billing is correlated with churn, but it is more likely a proxy for being a month-to-month customer, who are more likely to churn, rather than the direct cause itself.

7.2 Future Work

To build upon this analysis and address its limitations, the following steps are recommended:

- **Incorporate Temporal Data:** Add data from Telco's IT department to create a dataset that includes customer history over time, like service changes and support interactions, which would improve the model's accuracy.
- **Secure Additional Data Sources:** Integrate data from customer satisfaction surveys and marketing campaigns to understand the customer's sentiment and responsiveness to offers.
- **Explore Alternative Models:** Experiment with more complex models designed for sequences and time-series data, such as Gradient Boosting Machines (XGBoost, LightGBM) or recurrent neural networks (RNNs), if temporal data becomes available.