Analysis of Customer Churn in Telco Services

Executive Summary

The Telco Customer Churn Project aimed to analyze customer data and predict customer churn behavior to design targeted retention programs. The dataset provided insights into customer attributes and behaviors, services subscribed, account information, and demographic details, helping to identify factors contributing to customer churn.

Key Findings and Insights

1. ***Senior Citizens***: Senior citizens were found to have a significantly higher churn rate compared to other customer segments, suggesting the need for specialized services or support for this demographic.
2. ***Services and Contracts***: Customers who subscribed to a larger number of services and those on longer-term contracts were less likely to churn. This suggests the effectiveness of value-added services and contract stability in customer retention.
3. ***Fiber Optic Internet Service***: A significant number of customers who churned were subscribed to the fiber optic internet service, indicating potential dissatisfaction with this service.
4. ***New Customer***s: There was a higher churn rate amongst new customers (0-1 year tenure), underlining the need for more support.
5. ***Payment and Billing***: Customers using electronic check and paperless billing exhibited a higher churn rate, which may indicate potential issues with these methods.

Predictive Model Performance

- Several machine learning models were utilized to predict customer churn, including Logistic Regression, Support Vector Machines and Artificial Neural Networks.
- Among these, the Logistic Regression model showed the highest mean accuracy, followed closely by the Support Vector Machines model. These are good at predicting customers who will not churn, but poor at identifying who is likely to churn.
- Since the dataset has more non-churn customers than churn customers, the model may be biased towards the majority class, leading to a higher accuracy for non-churn predictions.

Introduction

In the increasingly competitive telecommunications industry, customer retention has become a vital strategy for business success. As acquiring a new customer can be significantly more costly than retaining an existing one, understanding the factors that lead to customer churn is of paramount importance.

Objective:

- Predict customer churn for a telecommunications company using the available dataset.
- Analyze customer data, including demographics, services, account information, and churn status, to develop targeted customer retention programs and help the company reduce customer attrition.
- Apply data exploration, preprocessing, feature engineering, and machine learning techniques to create a churn prediction model.

Variables

Variable	Object	DataType
CustomerID		Object
Gender	Male/female	Object
SeniorCitizen	1 = Senior, 0 = Not Senior	Int64
Partner	Yes = Has partner, No = No partner	Object
Dependents	Yes = Has dependent, No = No dependent	Object
Tenure	Number of months the customer has stayed	Int64
MulitpleLines	Yes = Has phone service, No = No Phone Service	object
InternetService	Internet service provider (DSL, Fiber optic, No)	Object
OnlineSecurity	online security (Yes, No, No internet service)	Object
OnlineBackup	online backup (Yes, No, No internet service)	Object
DeviceProtection	device protection (Yes, No, No internet service)	Object
TechSupport	tech support (Yes, No, No internet service)	Object
StreamingTV	streaming TV (Yes, No, No internet service)	Object
StreamingMovies	streaming movies (Yes, No, No internet service)	Object
Contract	The contract term (Month-to-month, One year, Two year)	Object
PaperlessBilling	paperless billing (Yes, No)	Object
PaymentMethod	Payment method (Electronic check, Mailed check, Bank Transfer (automatic))	Object
MonthlyCharges	The amount charged to the customer monthly	Float64
TotalCharges	The total amount charged to the customer	Object
Churn	The customer churned or not (Yes or No)	Object

Distribution of Categorical Variables (See Figure 1)

The dataset presents a balanced distribution between male and female customers, with 50.5% being male and 49.5% female. The majority of the dataset's subjects are not senior citizens, with seniors making up only 16.2% of the total.

Around half of the customers (51.7%) do not have partners, and most customers (90.3%) have phone service. Among those with phone service, 48.1% do not have multiple lines, while 42.2% do, and 9.7% have no phone service.

For internet service, 44% of the customers use fiber optic, 34.4% use DSL, and 21.7% do not have an internet service. Among those with internet service, fewer than half use online security (28.7%), online backup (34.5%), device protection (34.4%), and tech support (29.0%). The percentages of customers who do not use these services are slightly higher, except for those without an internet service.

Regarding entertainment services, the customer base is almost equally divided between those who use streaming TV (38.4%) and streaming movies (38.8%) and those who don't. The remaining 21.7% do not have internet service.

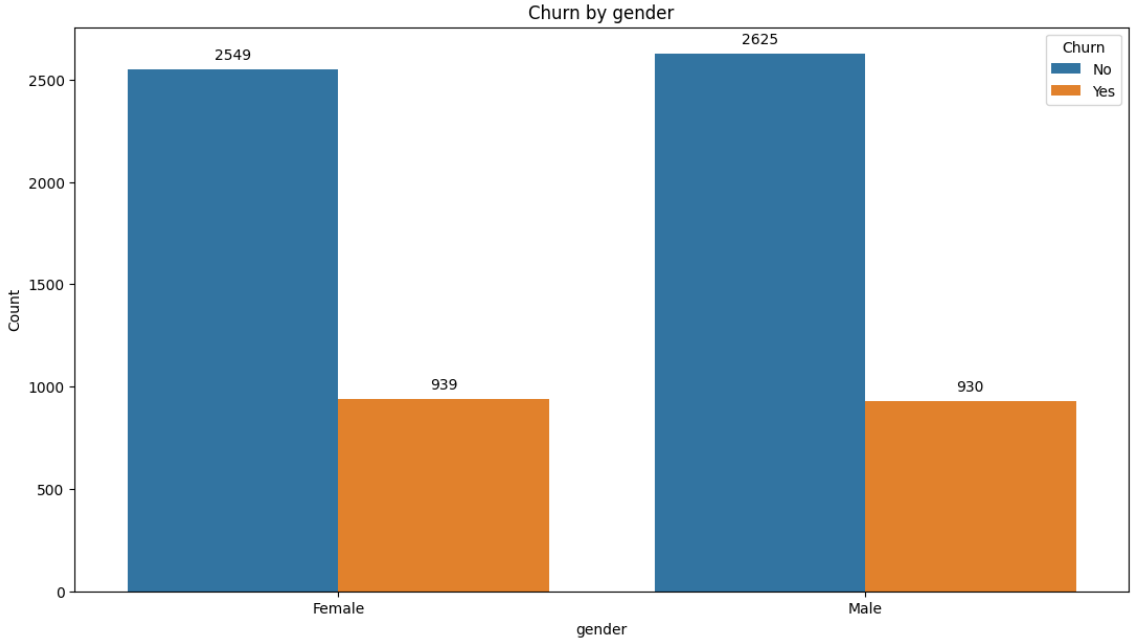
The distribution of contract type shows that most customers (55%) are on a month-to-month contract, followed by two-year contracts (24.1%) and one-year contracts (20.9%). A majority of customers (59.2%) use paperless billing.

Payment methods are relatively evenly distributed, with electronic check being the most common method (33.6%), followed by mailed check (22.9%), bank transfer (21.9%), and automatic credit card payments (21.6%).

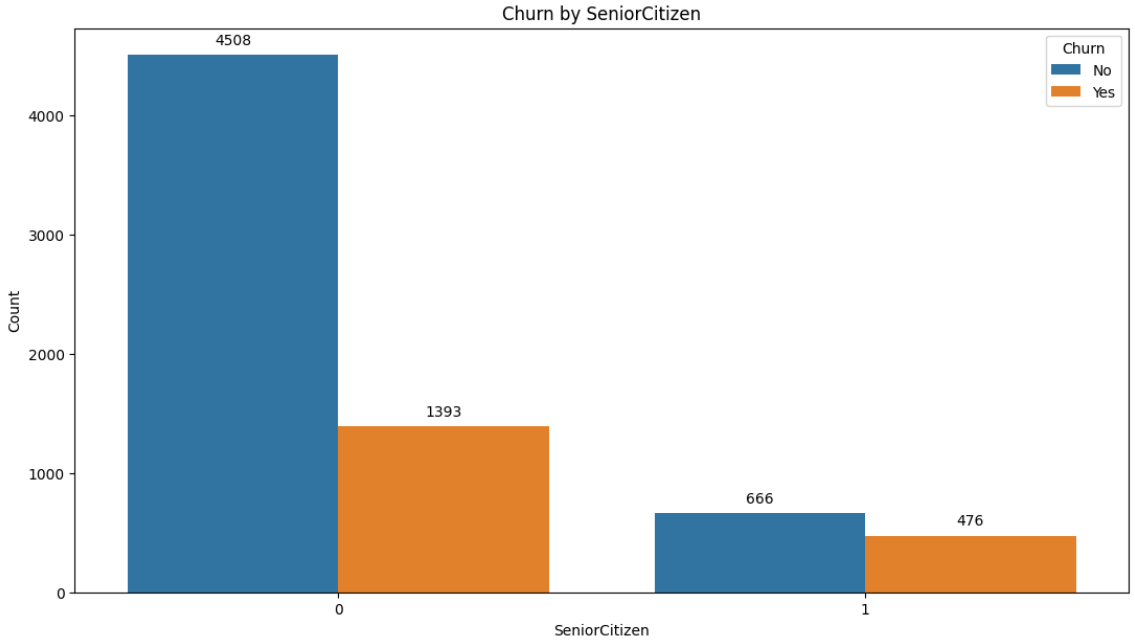
Finally, the churn variable shows that 26.5% of the customers in this dataset left the company within the last month, while the remaining 73.5% continued their services.

Exploratory Data Analysis Result

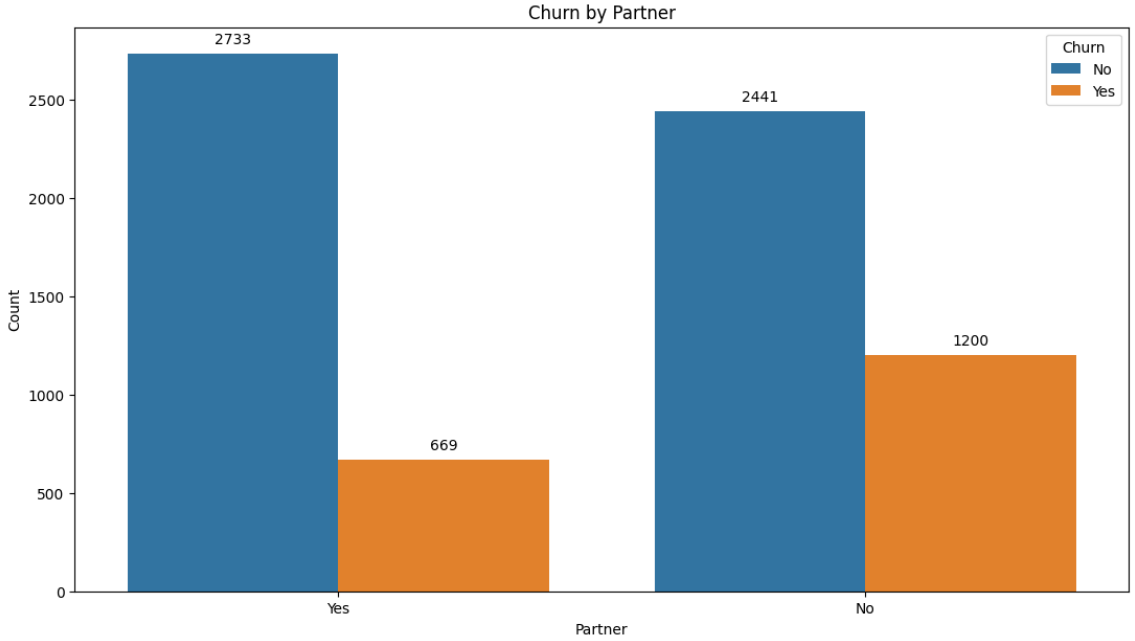
1. **Gender**: The churn rate appears to be quite balanced between genders, with both males and females having almost the same number of customers who churned and did not churn. Thus, gender may not be a significant factor in predicting customer churn.



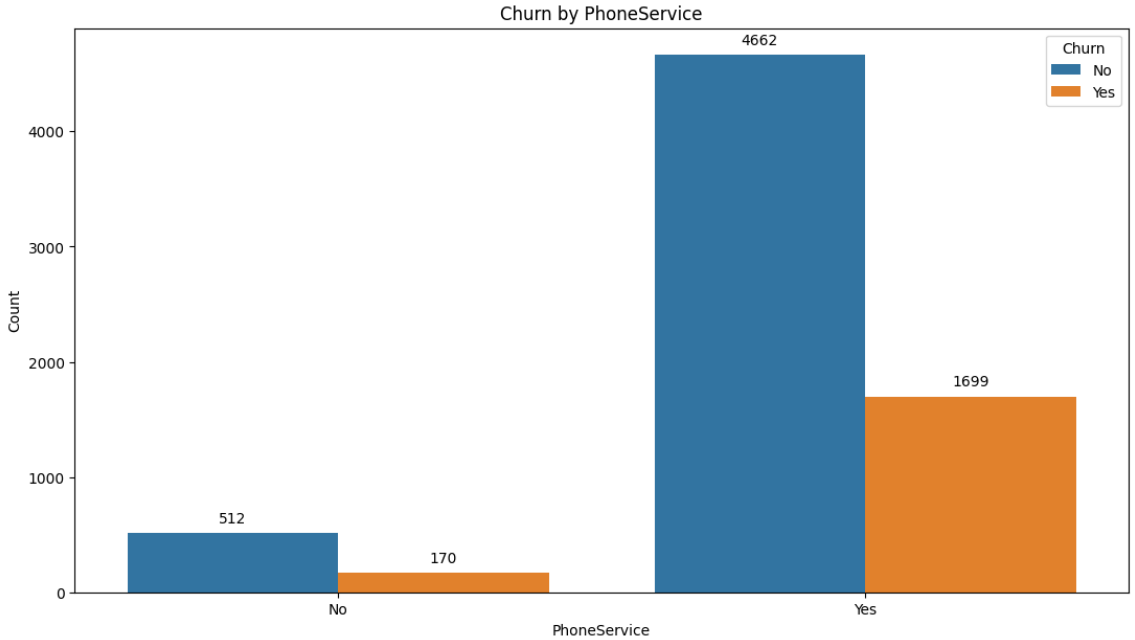
2. **Senior Citizen**: There is a higher churn rate among senior citizens compared to non-senior citizens. This might suggest that senior citizens find the services less satisfactory or have different needs that are not met.

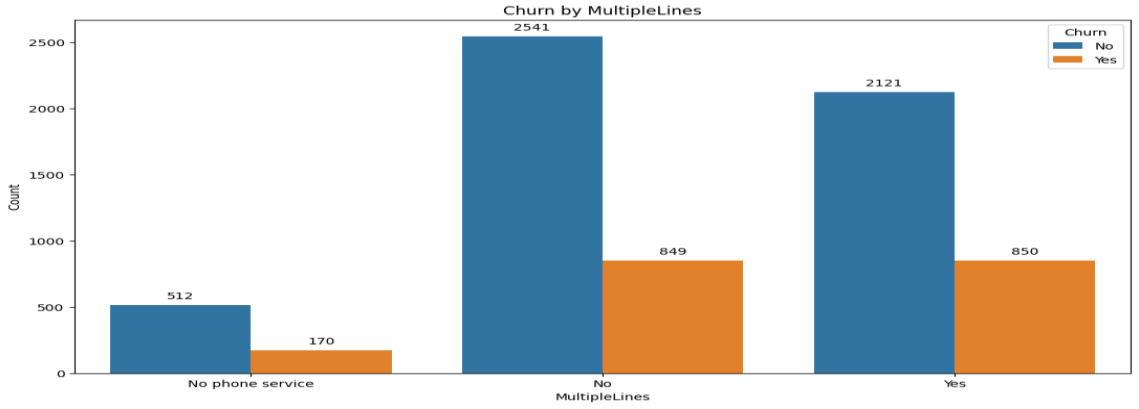


3. **Partner**: Customers without a partner tend to churn more than those with a partner. It might be because customers with partners may use more services or have more stable usage patterns.

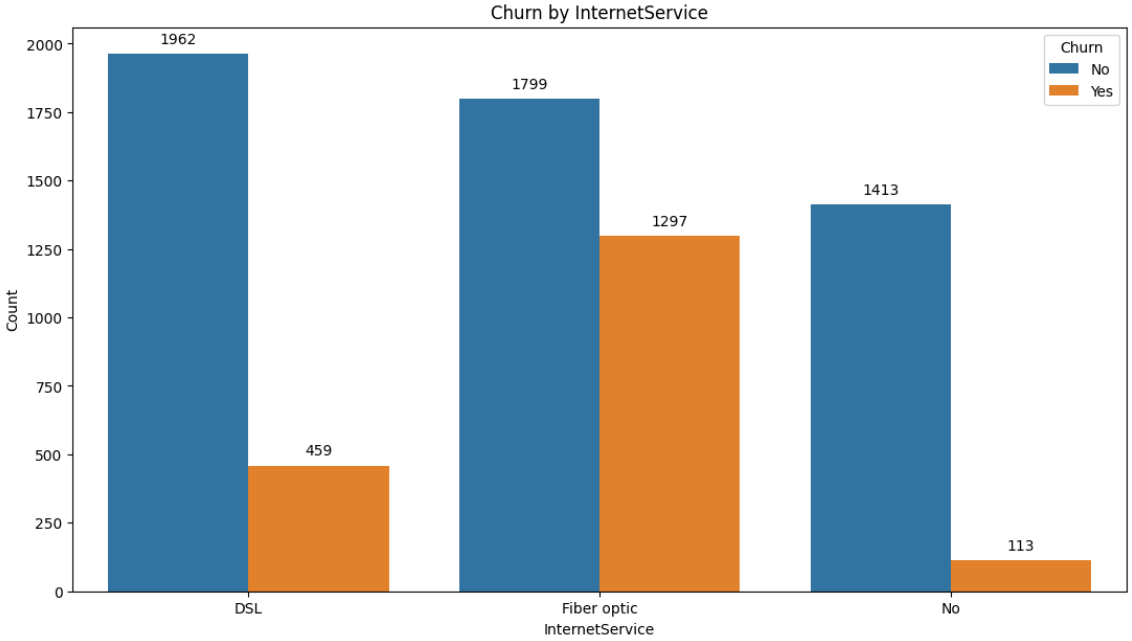


4. **Phone Service & Multiple Lines**: Most of the customers with phone service tend to stay. However, there's no significant difference in churn between customers with multiple lines and those without.

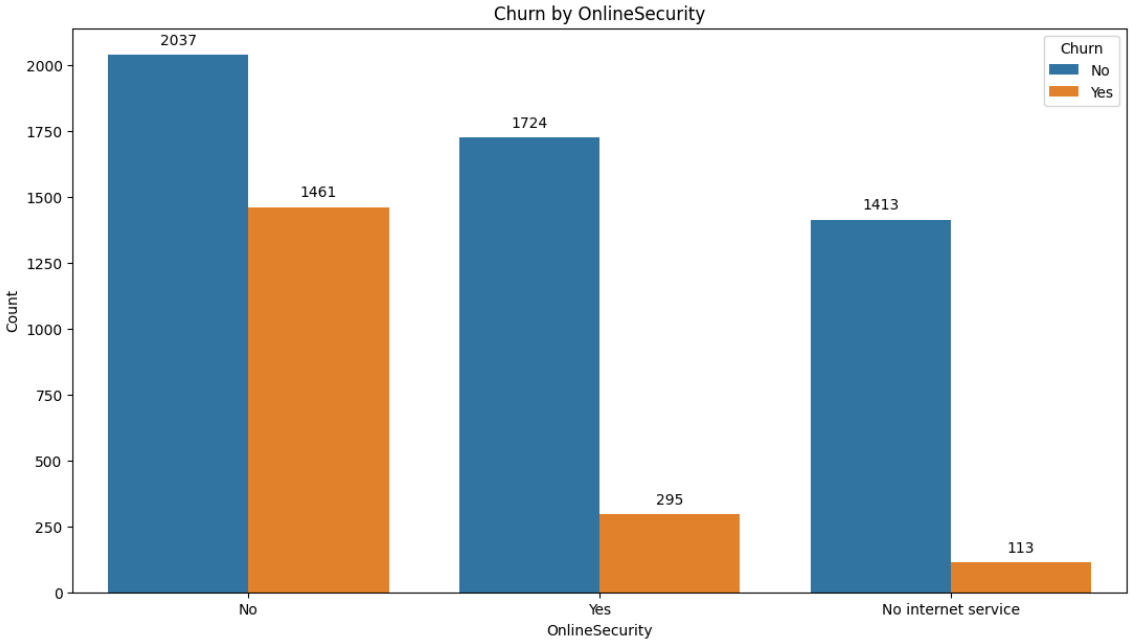


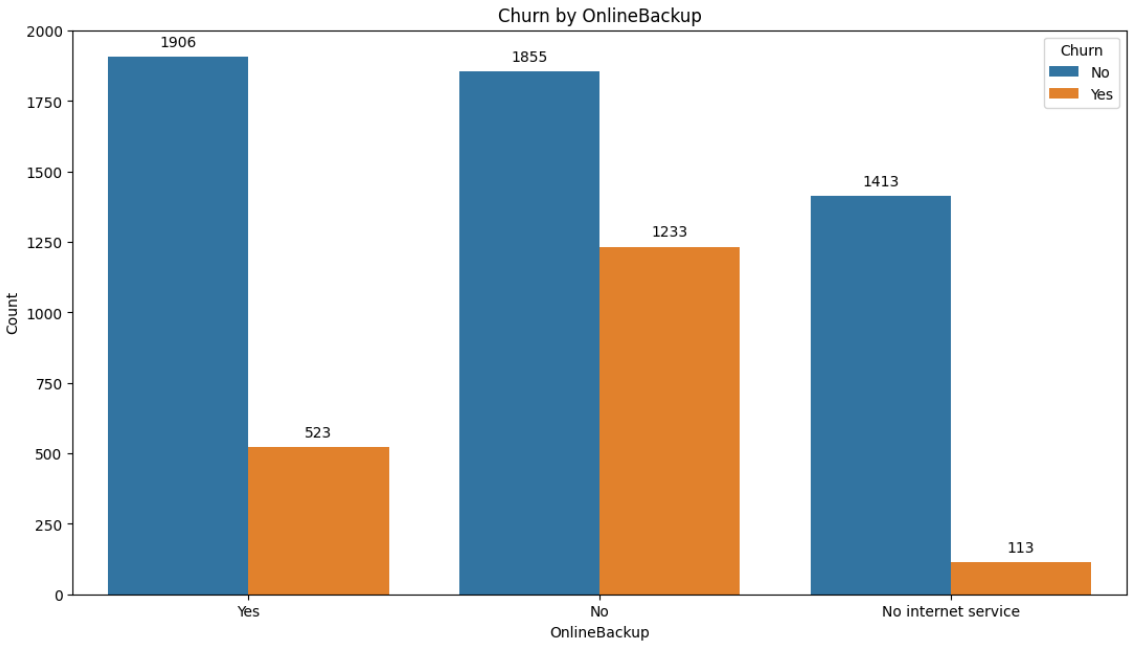


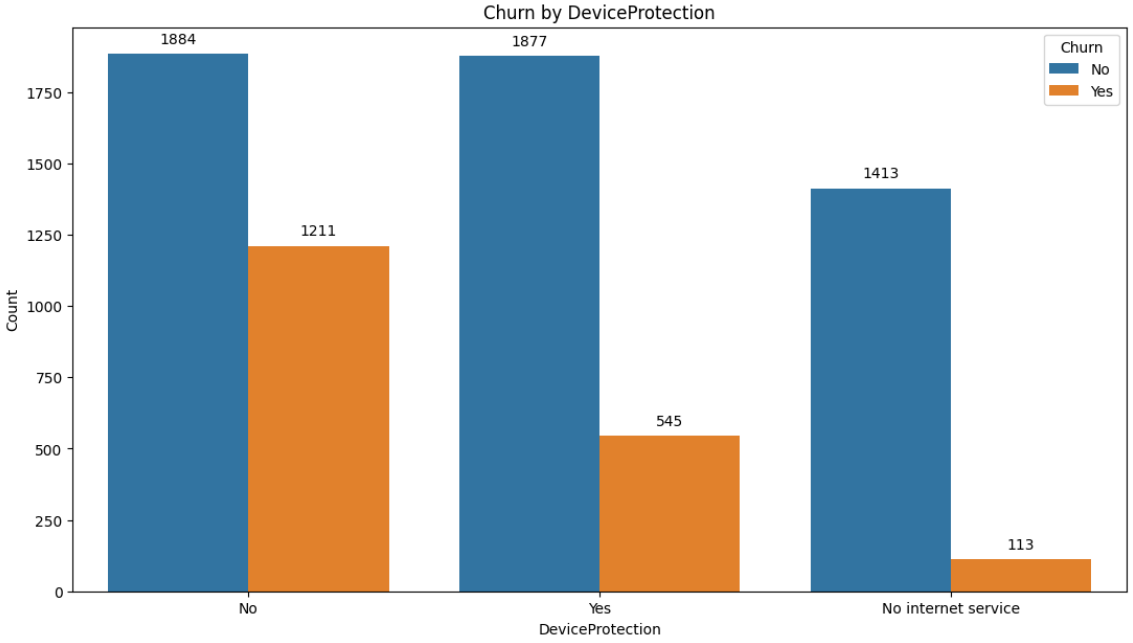
5. **Internet Service**: Customers using Fiber Optic services tend to churn more compared to those using DSL or no internet service. This might indicate a potential issue with the Fiber Optic service, such as cost or quality.

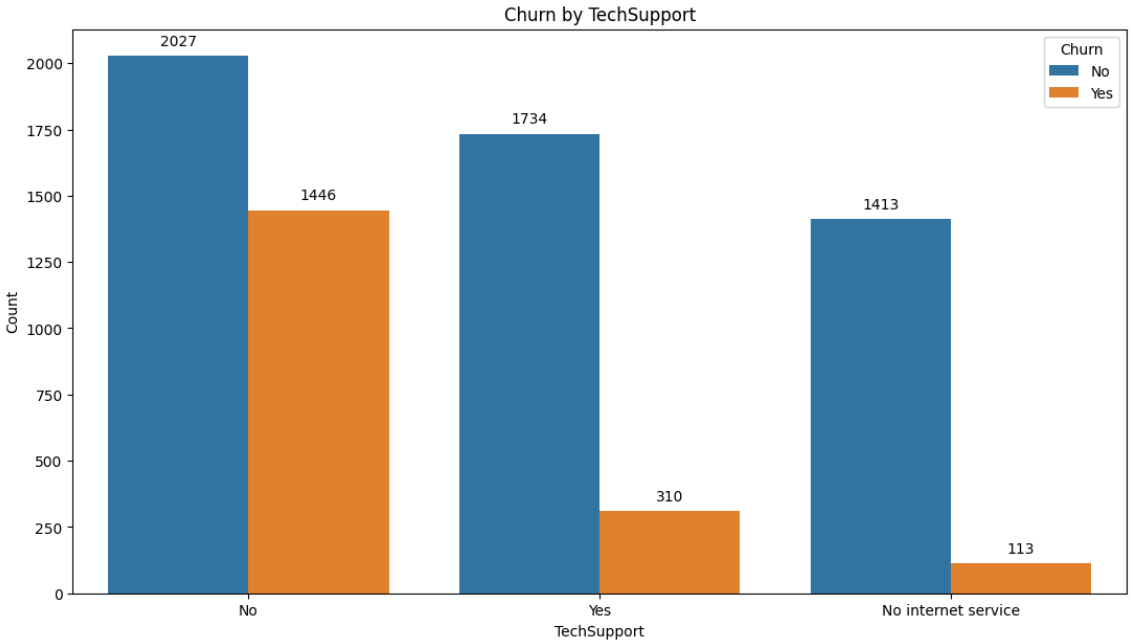


6. **Online Security, Online Backup, Device Protection, Tech Support**: Customers without these services tend to churn more. This might suggest that these features could improve customer retention.

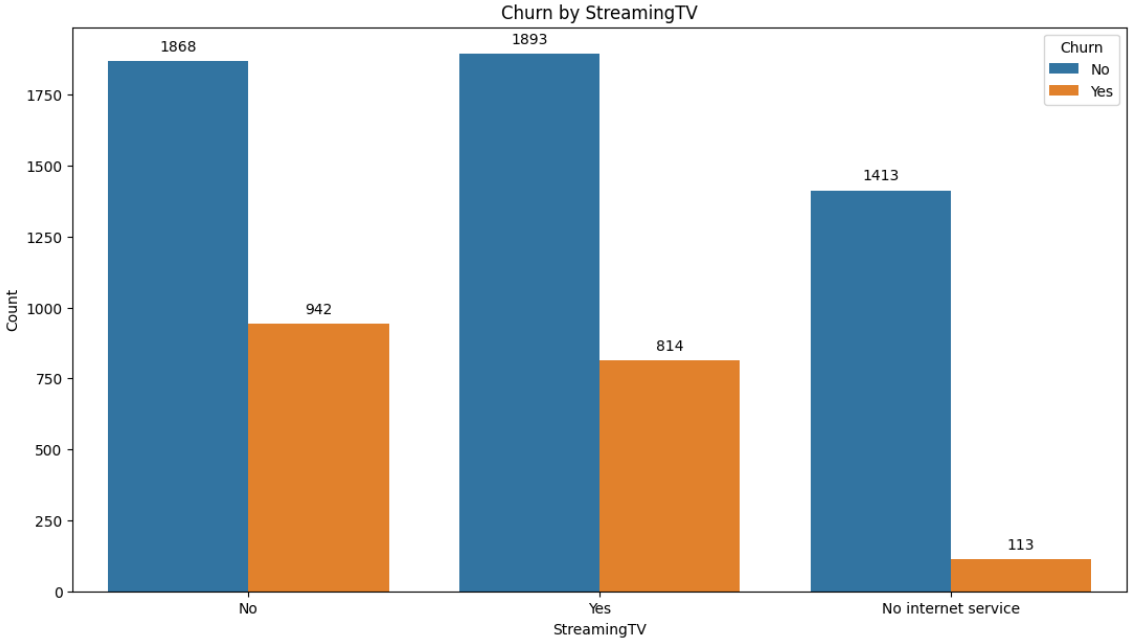


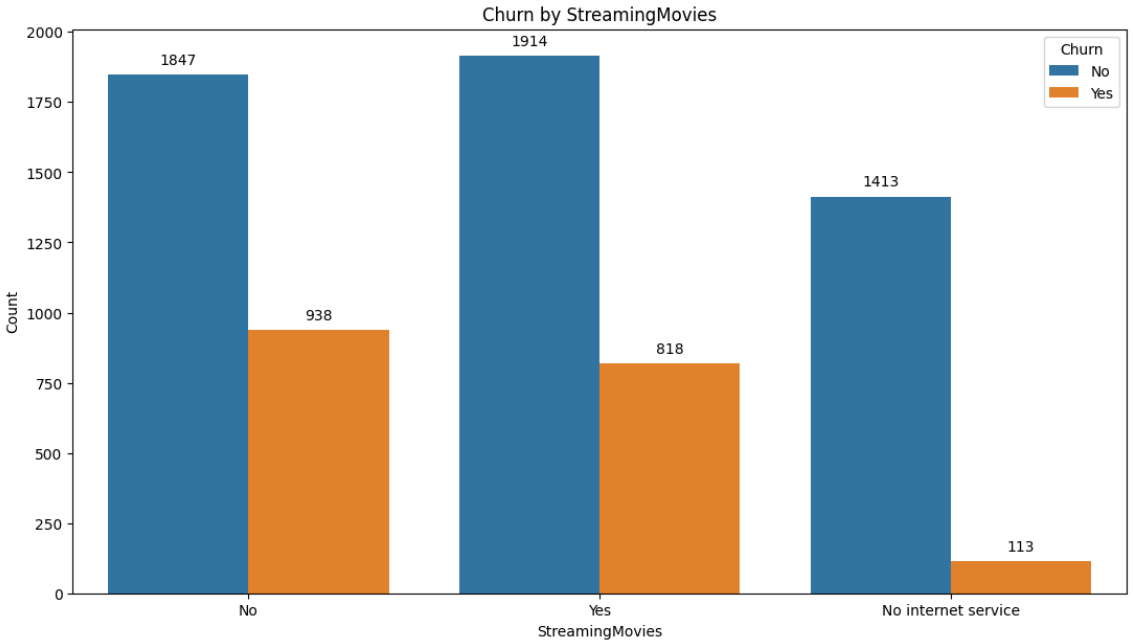

Churn by OnlineBackup
Churn
No
Yes
Count
OnlineBackup
Yes
No
No internet service
1906
523
1855
1233
1413
113


Churn by DeviceProtection
Churn
No
Yes
Count
DeviceProtection
No
Yes
No internet service
1884
1211
1877
545
1413
113

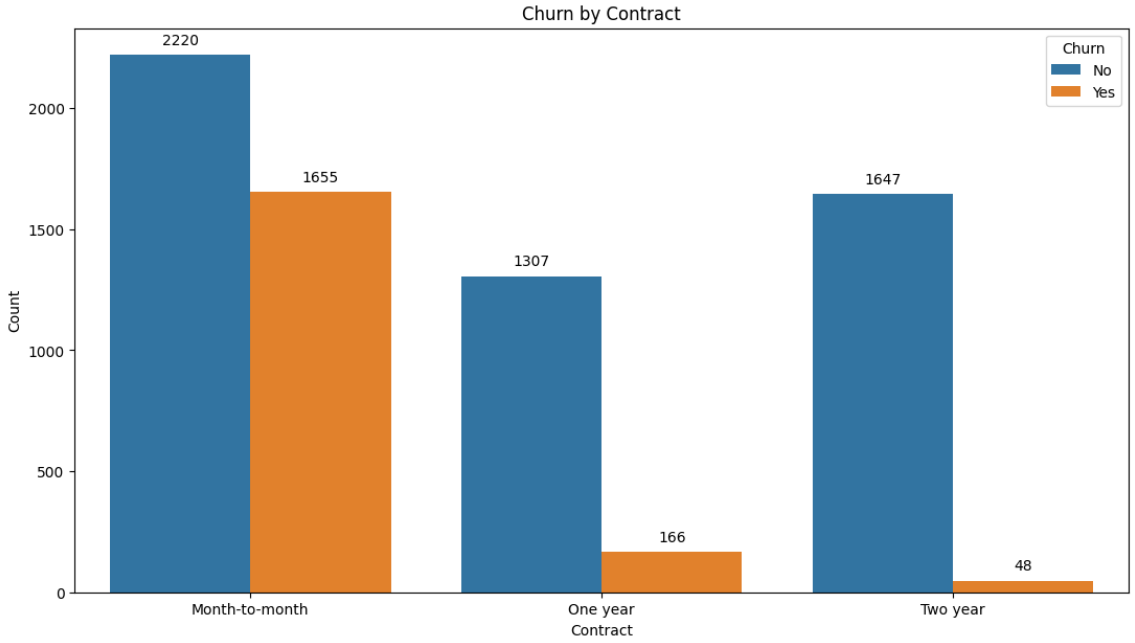

Churn by TechSupport
Churn
No
Yes
Count
TechSupport
No
Yes
No internet service
2027
1446
1734
310
1413
113

7. **Streaming TV, Streaming Movies**: There's no significant difference in churn between customers who use these services and those who don't. These features may not be major factors affecting customer churn.

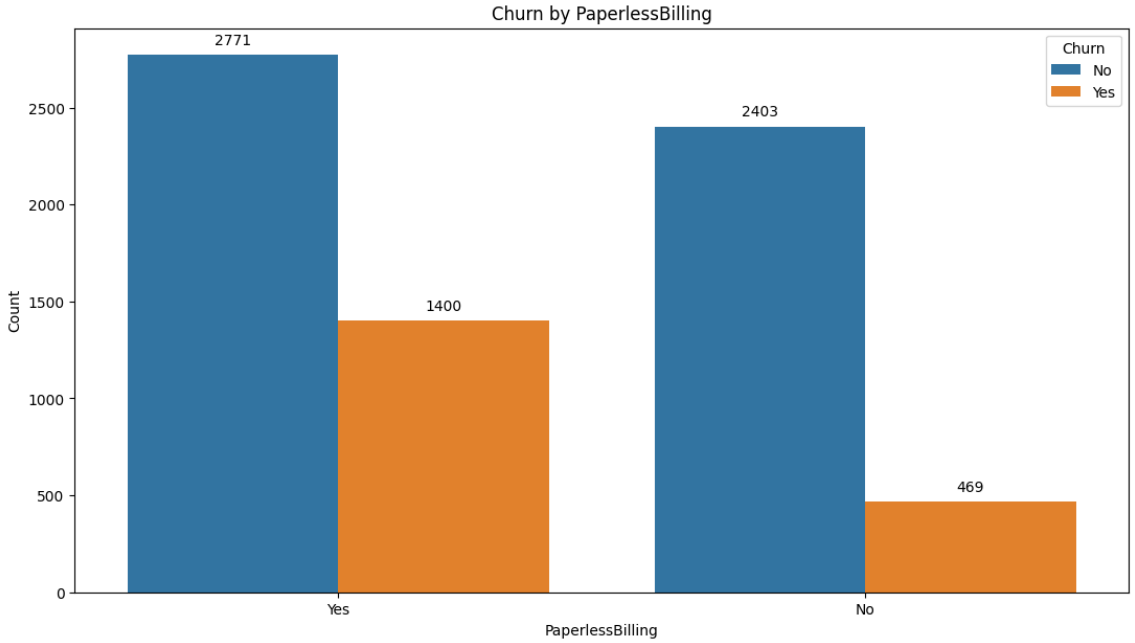




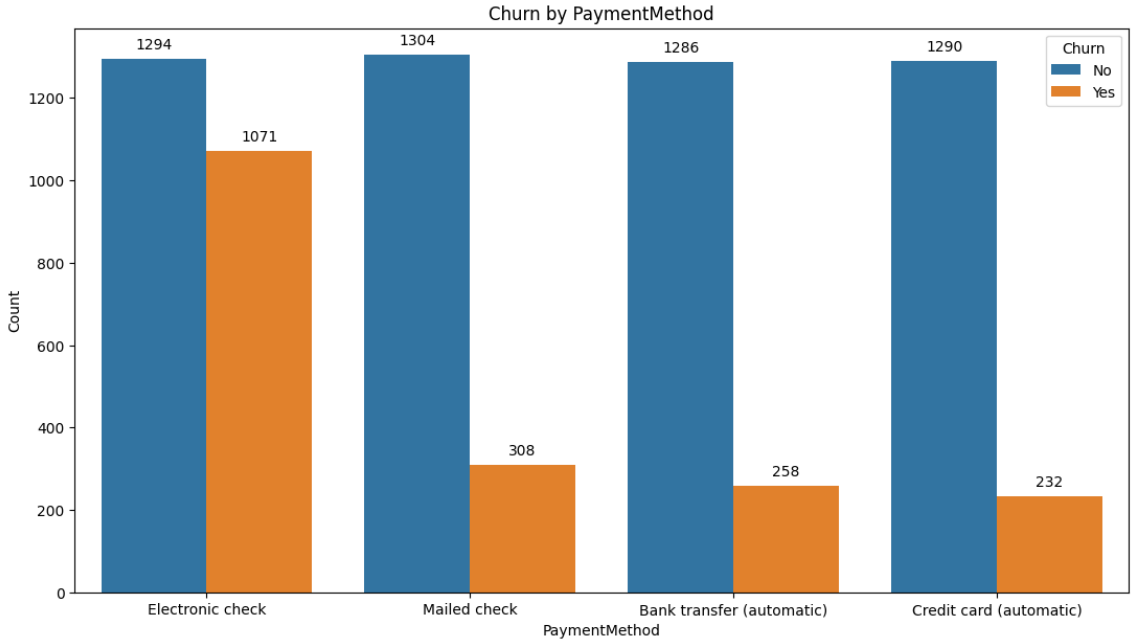
8. **Contract**: Customers with Month-to-month contracts tend to churn much more than those with one-year or two-year contracts. Longer contracts may provide a sense of stability and predictability that keeps customers from churning.



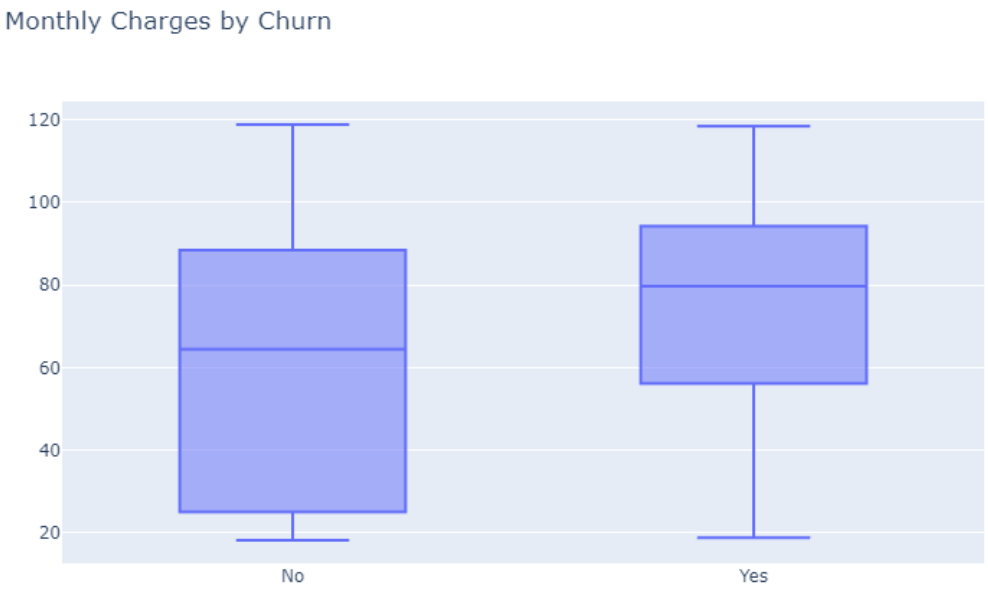
9. **Paperless Billing**: Customers using paperless billing tend to churn more. This could be due to factors like convenience or environmental consciousness not being strong enough retention factors, or issues with the paperless system itself.



10. **Payment Method**: Customers who pay via electronic check have a significantly higher churn rate compared to other payment methods. This might indicate some inconvenience or dissatisfaction associated with this payment method.



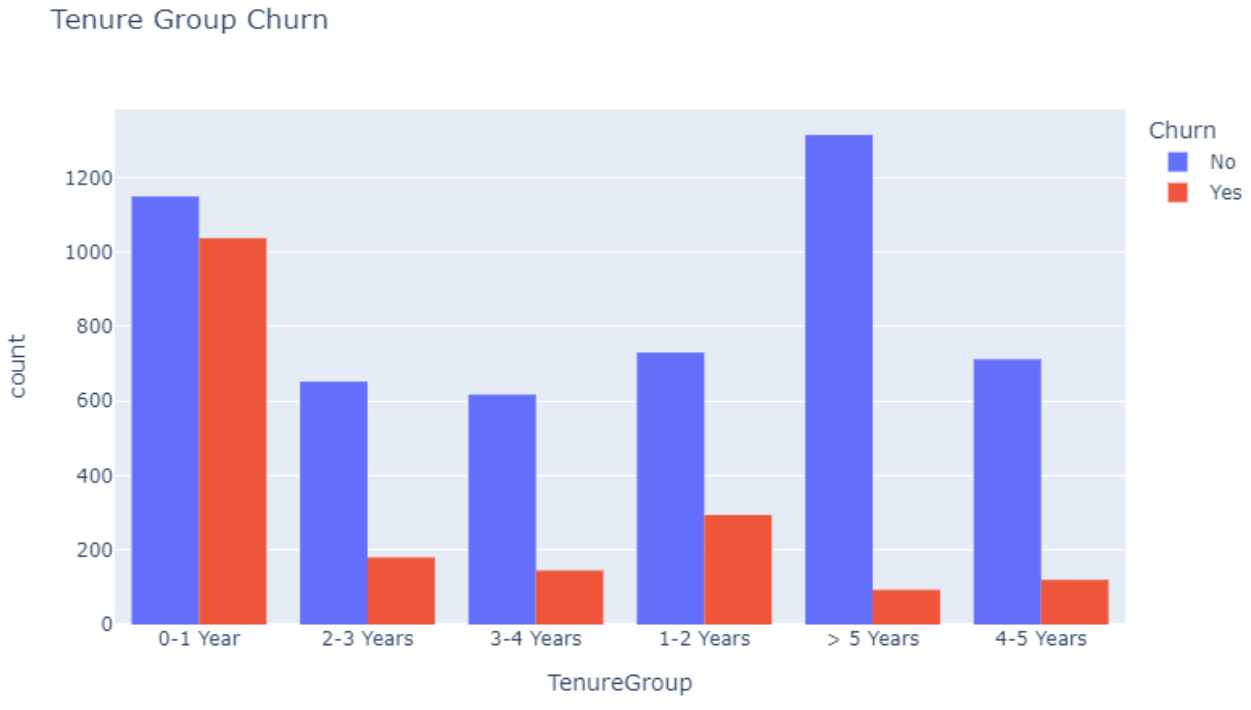
11. **Monthly Charges**:
    i. Customers who churned tend to have higher median monthly charges (79.65) compared to those who did not churn (64.425). This means higher monthly charges may be a contributing factor to customer churn.
    ii. The IQR for customers who churned is narrower than for customers who did not churn. This indicates that the monthly charges for customers who churned are more concentrated in a higher price range, while the charges for customers who did not churn are more spread out across different price points.
    iii. Both groups have a similar maximum monthly charge, but the minimum monthly charge for customers who did not churn is slightly lower than that for customers who churned. This may indicate that customers with lower monthly charges are less likely to churn.



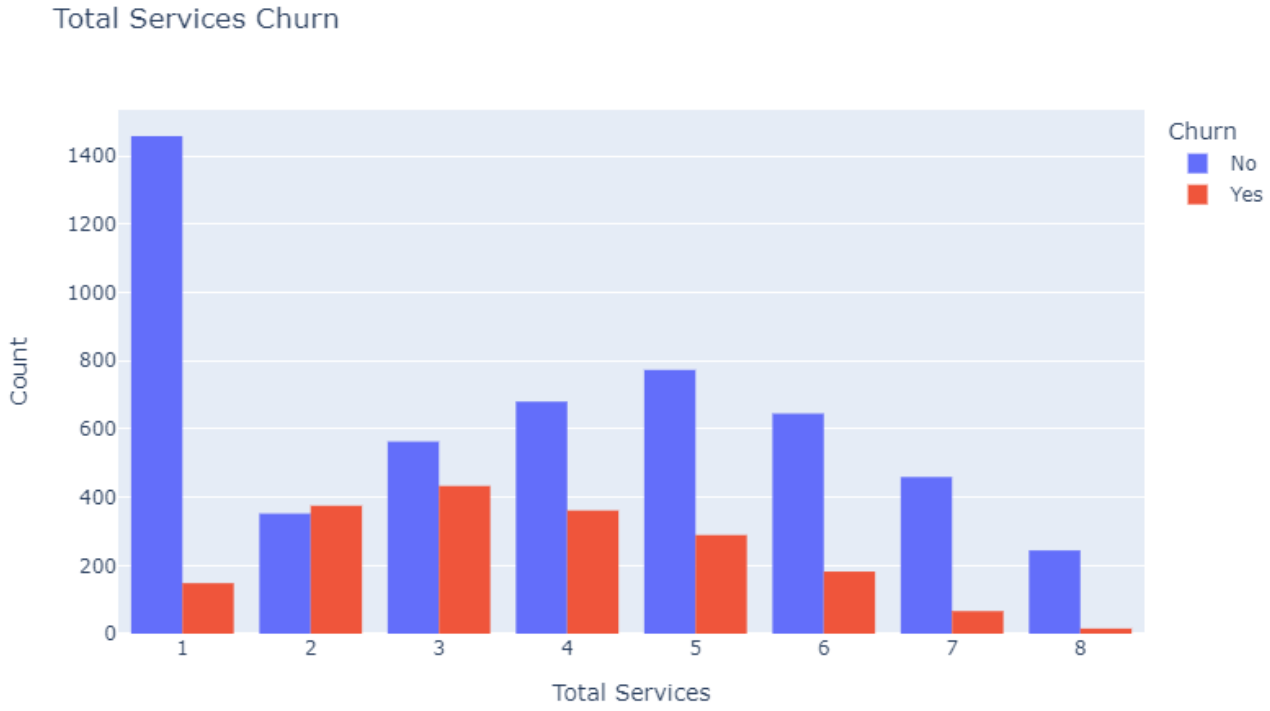
Feature Engineering and More Analysis Result

In this section, we created more features looking at more information.

1. **Tenure Group**
   - Bucketing the 'tenure' variable into different groups (e.g., '0-12 months', '13-24 months', '25-48 months', '49-60 months', '> 60 months').
   - This can help reveal patterns related to customer loyalty and the likelihood of churn based on the length of their relationship with the company.
     i. ***0-1 Year Tenure***: There is a high churn rate among customers who have been with the company for less than a year. This could be due to new customers not being fully satisfied with the service or finding a better offer from a competitor.
     ii. ***1-2 Years Tenure***: The churn rate significantly decreases for customers who've been with the company for 1-2 years. This indicates that after the initial year, customers tend to stay with the company, suggesting a period of adjustment where they get used to the services and value offered.
     iii. ***2-5 Years Tenure***: The churn rate continues to decrease as the customer tenure increases to between 2 and 5 years. This indicates that the longer a customer stays with the company, the less likely they are to churn. They may have formed habits, become familiar with the service, and perceive switching to another company as inconvenient.
     iv. ***More than 5 Years Tenure***: For customers with over 5 years of tenure, the churn rate is the lowest, indicating high customer loyalty and satisfaction with the services provided.



2. **Total Services**
   - Identify if customers with more services are less likely to churn.
     i. ***Single Service***: Customers who have only subscribed to one service have a high churn rate. This could be due to a lack of engagement or satisfaction with the provided service. It could also be that these customers are more likely to compare the company's single service with competitors and switch if they find a better offer.
     ii. ***Multiple Services***: As the number of services increases from 2 to 4, the churn rate also increases. This could be due to the complexity and cost associated with managing multiple services.
     iii. ***High Service Subscriptions***: Interestingly, we observe a decline in churn as the number of services goes beyond 4. This could be due to the perceived value and convenience of bundled services. Customers with 5 to 7 services have lower churn rates compared to those with fewer services.
     iv. ***Maximum Services***: Customers who have subscribed to all 8 services have a significantly lower churn rate compared to other groups. This could be because these customers are highly engaged and see value in the complete package of services. However, the number of customers in this category is quite low, so this observation might not be as significant.



Summaries and Discussion

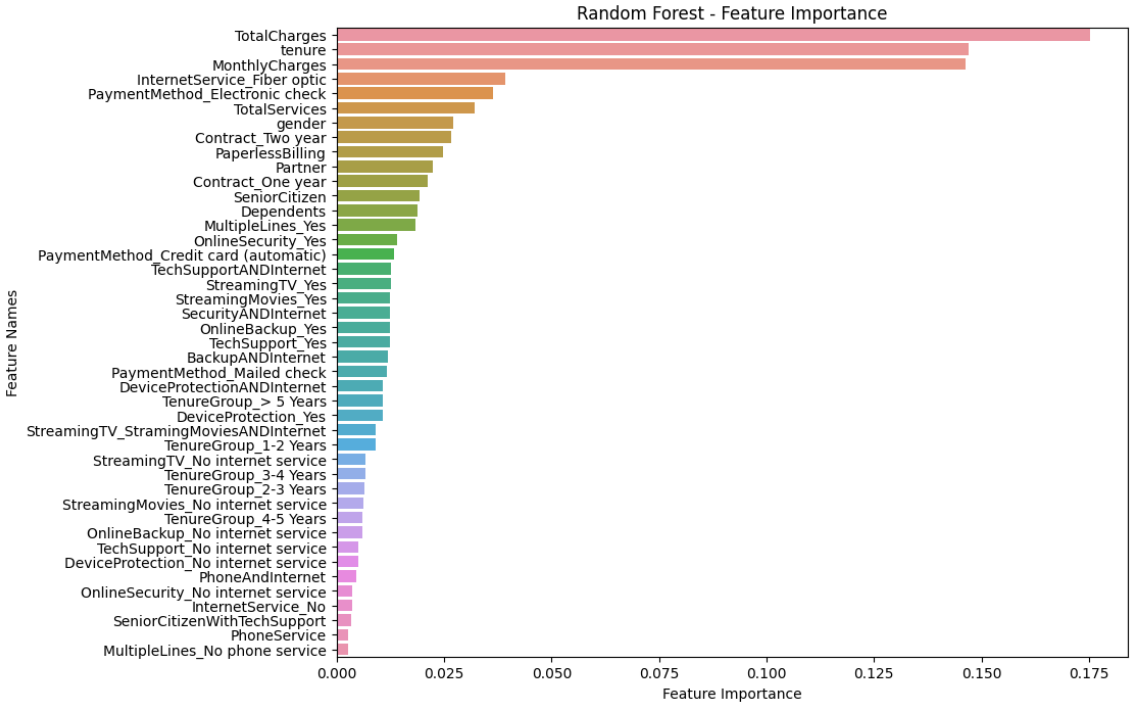
1. **Demographics**:
    I. ***Gender***: The churn rate is almost evenly split between males and females, indicating that gender may not significantly affect customer churn.
    II. ***Senior Citizens***: Senior citizens have a higher churn rate than non-seniors, suggesting that the company may need to address the unique needs and expectations of senior customers to retain them.
    III. ***Partner Status***: Customers without partners have a higher tendency to churn than those with partners. It might be that customers with partners find more value in the company's services, maybe due to shared usage or bundled offers.

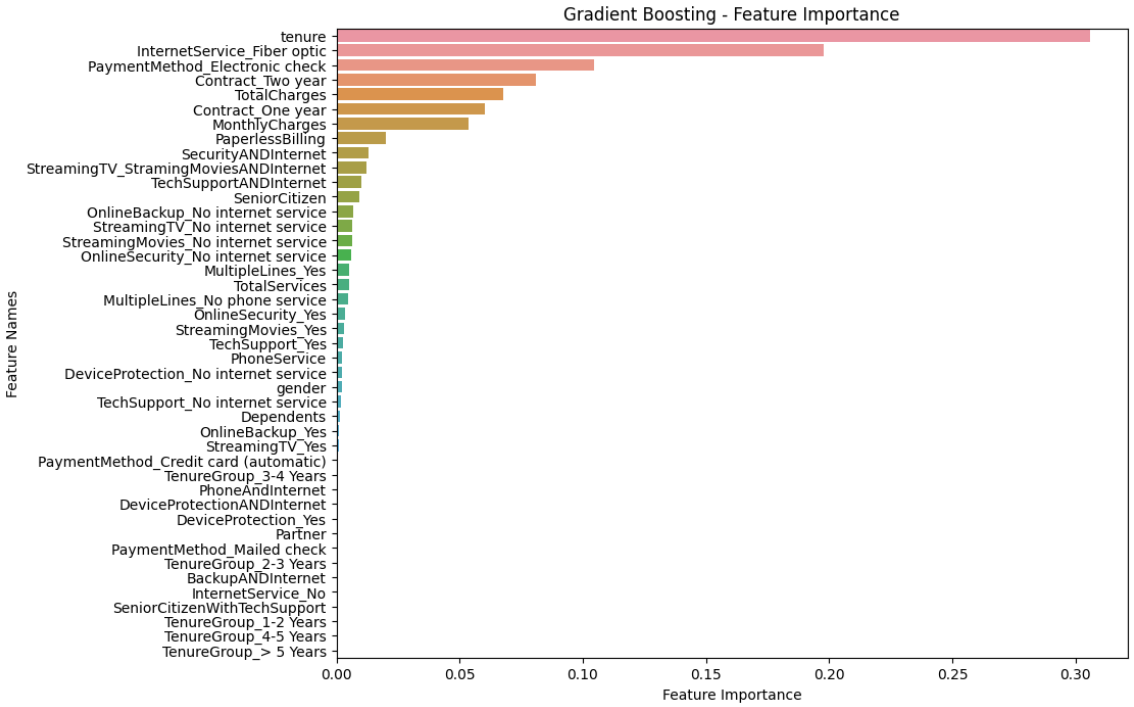
2. **Services**:
    I. Customers who have signed up for more services (indicated by a higher TotalServices value) are less likely to churn. This suggests that bundled services could be an effective strategy for customer retention.
    II. Internet Service type also influences churn. Customers with Fiber Optic services show a higher churn rate compared to those with DSL or no internet service. Perhaps customers are not finding the expected value in the Fiber Optic service, or there might be issues with the service quality.

3. **Contract and Billing**:
    I. Customers on a month-to-month contract are significantly more likely to churn compared to those on one-year or two-year contracts. Offering incentives for customers to switch to long-term contracts could reduce churn.
    II. Customers with paperless billing have a higher churn rate. This might be due to the digital literacy of the customers or the user experience of the online billing system.

Feature Section and Modeling

From the previous analysis, we know that Services, Contract, Billing, Partner Status, etc. are more important in predicting whether a customer will churn or not. This is shown in the feature importance of Random Forest and Gradient Boosting.





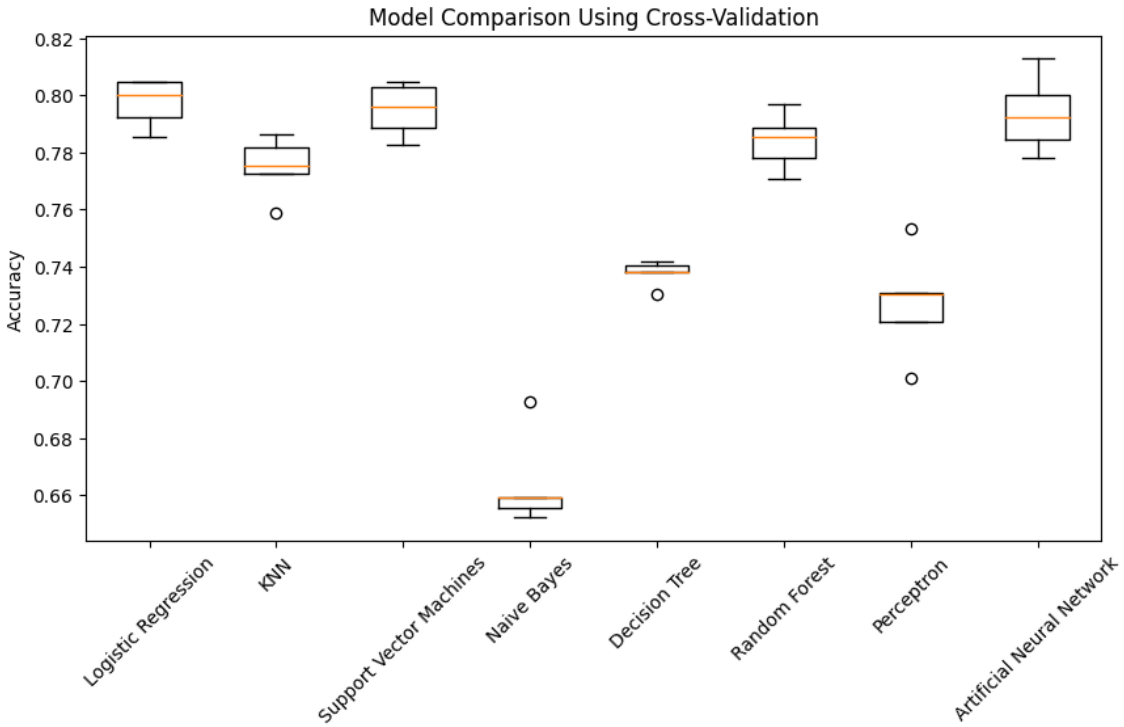
Here were the features selected to build our model:

1. InternetService_Fiber optic
2. PaymentMethod_Electronic check
3. PhoneAndInternet
4. MonthlyCharges
5. PaperlessBilling
6. Dependents

7. TechSupportANDInternet
8. TechSupport_Yes
9. OnlineSecurity_Yes
10. OnlineSecurity_Yes
11. SecurityANDInternet
12. Contract_One year
13. TotalCharges
14. TenureGroup_> 5 Years
15. StreamingMovies_No internet service
16. StreamingTV_No internet service
17. TechSupport_No internet service
18. DeviceProtection_No internet service
19. OnlineBackup_No internet service
20. InternetService_No
21. OnlineSecurity_No internet service
22. Contract_Two year
23. Tenure

Also, based on the comparison of models using cross-validation, we chose logistic regression, support vector machines, and artificial neural networks to build the models.

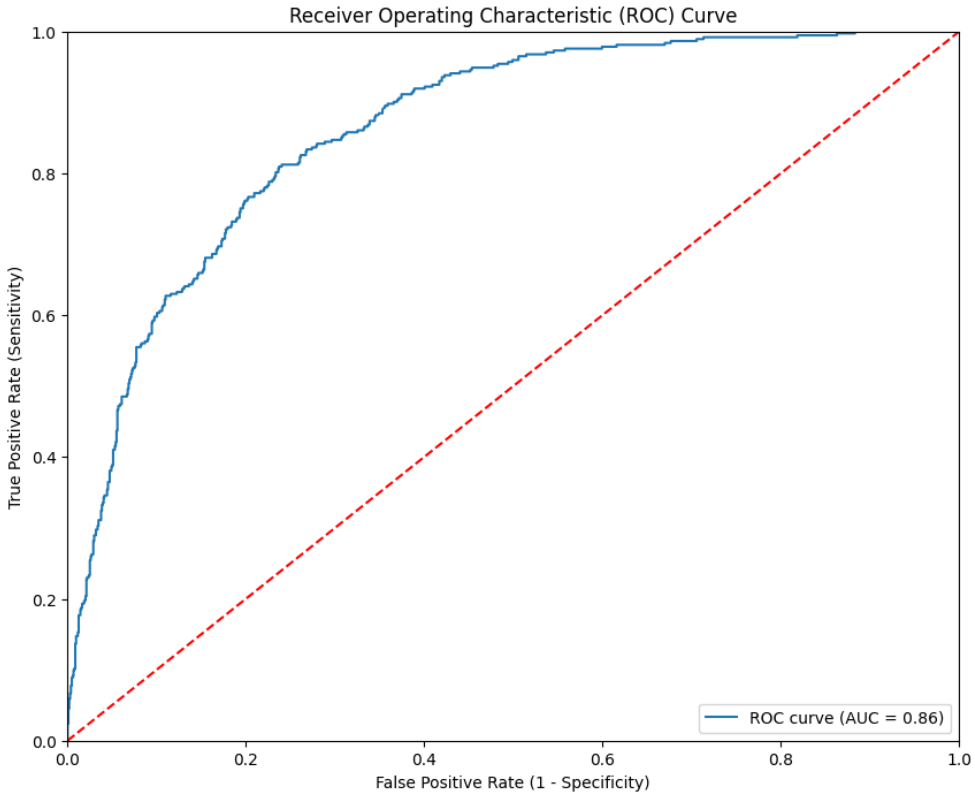
Logistic Regression: 0.797 (0.008)
KNN: 0.775 (0.009)
Support Vector Machines: 0.795 (0.008)
Naive Bayes: 0.664 (0.015)
Decision Tree: 0.738 (0.004)
Random Forest: 0.784 (0.009)
Perceptron: 0.727 (0.017)
Artificial Neural Network: 0.794 (0.012)



Performance of these models below:

A. Logistic Regression

```
Accuracy: 0.8211497515968772
Classification report:
              precision    recall  f1-score   support

           0       0.86      0.90      0.88      1036
           1       0.69      0.60      0.64       373

    accuracy                           0.82      1409
   macro avg       0.77      0.75      0.76      1409
weighted avg       0.81      0.82      0.82      1409

Confusion matrix:
[[935 101]
 [151 222]]
```

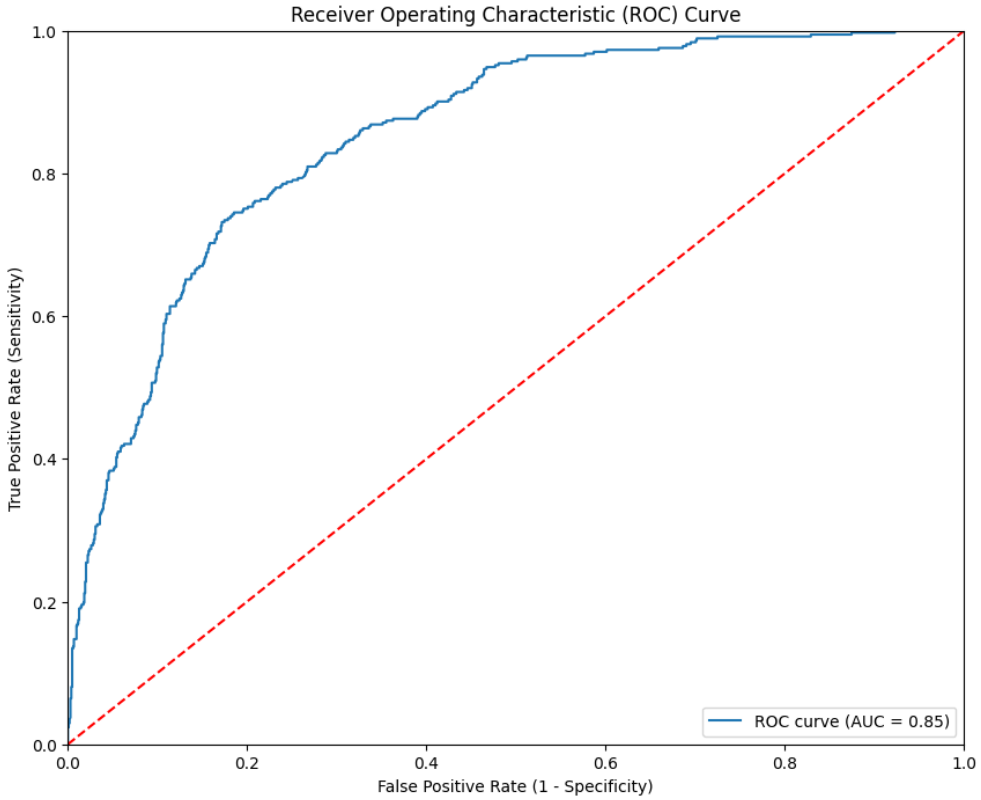


B. Support Vector Machines

```
Accuracy: 0.8069552874378992
Classification report:
              precision    recall  f1-score   support

           0       0.84      0.92      0.87      1036
           1       0.68      0.50      0.58       373

    accuracy                           0.81      1409
   macro avg       0.76      0.71      0.73      1409
weighted avg       0.80      0.81      0.80      1409

Confusion matrix:
[[949  87]
 [185 188]]
```

C. Artificial Neural Networks

```
Accuracy: 0.7977288857345636
Confusion Matrix:
 [[928 108]
 [177 196]]
Classification report:
              precision    recall  f1-score   support

           0       0.84      0.90      0.87      1036
           1       0.64      0.53      0.58       373

    accuracy                           0.80      1409
   macro avg       0.74      0.71      0.72      1409
weighted avg       0.79      0.80      0.79      1409
```



Discussion of modeling results

The performance of these models is evaluated by several metrics: accuracy, precision, recall, F1 score, AUC-ROC.

i. **Accuracy**: The ratio of correct predictions to the total number of predictions. While it is a straightforward metric, it can be misleading if the classes are imbalanced.

ii. **Precision**: Measures the percentage of correctly predicted positive observations out of the total predicted positives. High precision relates to a low false positive rate.

iii. **Recall (Sensitivity)**: Measures the percentage of correctly predicted positive observations out of the actual positives. High recall indicates a

model that is able to find all the positive samples.

iv. **F1-Score**: The weighted average of Precision and Recall, taking both metrics into account. It tries to find the balance between precision and recall.

v. **AUC-ROC**: Which is less sensitive to class imbalance, might provide a better indication of model performance in class imbalance cases.

- An AUC of 0.5 represents a model that is as good as random. This is represented by the diagonal line in the ROC plot.
- An AUC of 1.0 represents a perfect model that has no false positives or false negatives.
- An AUC of 0.0 represents a model that is perfectly wrong, i.e., it always makes the wrong prediction.

Logistic Regression: This model performs well with an accuracy of approximately 82%. The recall for class 1 (churn) is 0.59 which means the model is able to correctly identify 59% of the churn cases. This could be improved, but the F1-score of 0.64 indicates a decent balance between precision and recall. The AUC-ROC of 0.86 indicates that your model has a high chance of correctly distinguishing between positive and negative classes.

Support Vector Machines (SVM): The SVM model has an accuracy of 80.7%, slightly lower than the Logistic Regression model. The recall for class 1 is lower (0.49) compared to the Logistic Regression model, meaning it identifies fewer true positive churn cases. The F1-score is also lower (0.57) indicating a worse balance between precision and recall compared to the Logistic Regression model.

Artificial Neural Networks (ANN): The ANN model has a similar accuracy to the ANN model (80%). The recall for class 1 is a bit higher than SVM's but still lower than Logistic Regression's (0.53). The F1-score is similar to SVM's (0.58), indicating a similar balance between precision and recall. The AUC-ROC of 0.85 indicates that your model has a high chance of correctly distinguishing between positive and negative classes.

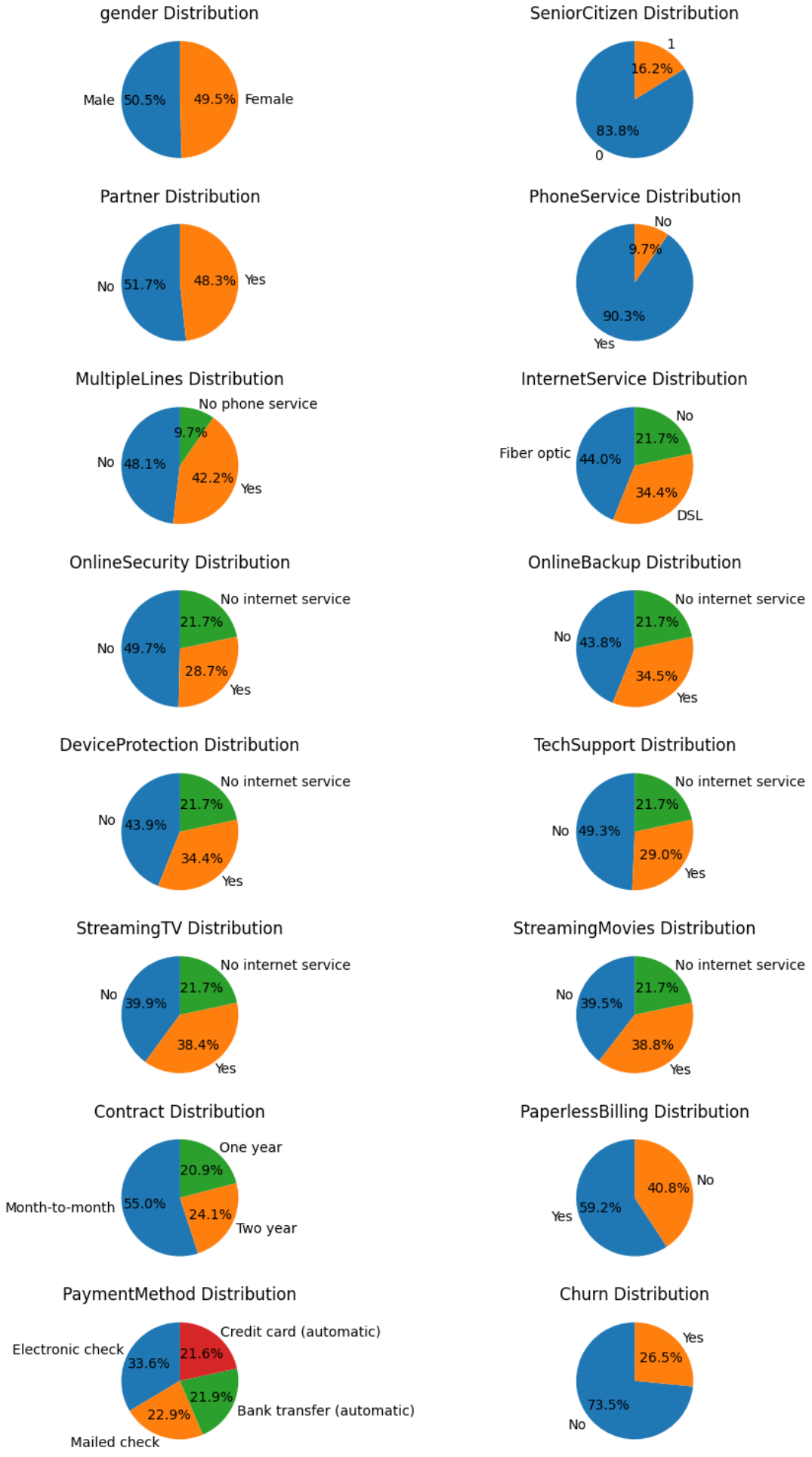
Overall, all models have decent accuracy, but the Logistic Regression model performs the best out of the three in terms of both accuracy and ability to correctly identify churn cases (recall for class 1). These three models are good at predicting customers who will not churn, but poor at identifying who is likely to churn.

Since the dataset has more non-churn customers than churn customers, the model may be biased towards the majority class, leading to a higher accuracy for non-churn predictions.

Recommendations

1. **Improve Services for Senior Citizens**: As senior citizens are found to have a higher churn rate, special attention should be given to cater to their needs. This could include providing more user-friendly interfaces, offering senior citizen discounts, or providing dedicated customer service.
2. **Bundle Services**: Encourage customers to subscribe to more services, as customers who subscribed to more services were less likely to churn. Bundling services together at a discounted rate can be an attractive offer for customers and can increase their likelihood of staying.
3. **Revisit Fiber Optic Pricing or Services**: Customers with fiber optic services churned more. It would be beneficial to look into their concerns or complaints. It might be possible to improve the quality of service or adjust the pricing to make it more competitive.
4. **Encourage Longer-term Contracts**: Customers with month-to-month contracts were more likely to churn than those with one or two-year contracts. Incentives could be provided to encourage customers to commit to longer-term contracts.
5. **Enhance Support for New Customers**: New customers had a higher churn rate. This could be reduced by providing enhanced customer support in the initial months, ensuring that they have a good understanding of the services and are satisfied.
6. **Examine Payment and Billing Methods**: Customers using electronic check and paperless billing churned more. It would be worth examining if these methods of payment and billing are causing any issues or dissatisfaction.
7. **Leverage Predictive Models for Proactive Retention**: These models can be used to identify at-risk customers before they churn. Proactive measures can then be taken to retain these customers. Although they seem to be good at predicting customers who will not churn and not so good at identifying who is likely to churn, it still has some reference value.

Figure 1:


gender Distribution
Male
50.5%
49.5%
Female
SeniorCitizen Distribution
1
16.2%
83.8%
0
Partner Distribution
No
51.7%
48.3%
Yes
PhoneService Distribution
No
9.7%
90.3%
Yes
MultipleLines Distribution
No phone service
9.7%
No
48.1%
42.2%
Yes
InternetService Distribution
No
21.7%
Fiber optic
44.0%
34.4%
DSL
OnlineSecurity Distribution
No internet service
21.7%
No
49.7%
28.7%
Yes
OnlineBackup Distribution
No internet service
21.7%
No
43.8%
34.5%
Yes
DeviceProtection Distribution
No internet service
21.7%
No
43.9%
34.4%
Yes
TechSupport Distribution
No internet service
21.7%
No
49.3%
29.0%
Yes
StreamingTV Distribution
No internet service
21.7%
No
39.9%
38.4%
Yes
StreamingMovies Distribution
No internet service
21.7%
No
39.5%
38.8%
Yes
Contract Distribution
One year
20.9%
Month-to-month
55.0%
24.1%
Two year
PaperlessBilling Distribution
40.8%
No
Yes
59.2%
PaymentMethod Distribution
Credit card (automatic)
Electronic check
21.6%
33.6%
21.9%
Bank transfer (automatic)
22.9%
Mailed check
Churn Distribution
Yes
26.5%
73.5%
No