Elliot Gaynon

Final Report:
Telco Customer Churn

Introduction/Problem Statement:

Telco was once a thriving telecommunications giant, adding millions of customers annually and earning a reputation for reliable coverage nationwide. As competitors introduced cheaper plans and better customer experiences, Telco failed to adapt, and its customer churn rose. In order to develop customer retention programs, Telco must first identify those customers most in danger of churning.

By using Telco's customer data, which includes demographic info, account information, and services for which the customer subscribed, I created a tool for Telco to identify those at-risk customers. By wrangling the data, performing an exploratory analysis, and utilizing supervised machine learning, I developed a model to predict which customers would most likely churn.

My XGBoost Classification model achieved a recall rate 0.76 and can be used by Telco to create retention programs to avoid losing more customers.

Dataset:

I received the dataset from Telco's website. The dataframe was 7,403 rows (each row representing a unique customer) and 21 columns. The columns can be summarized as:

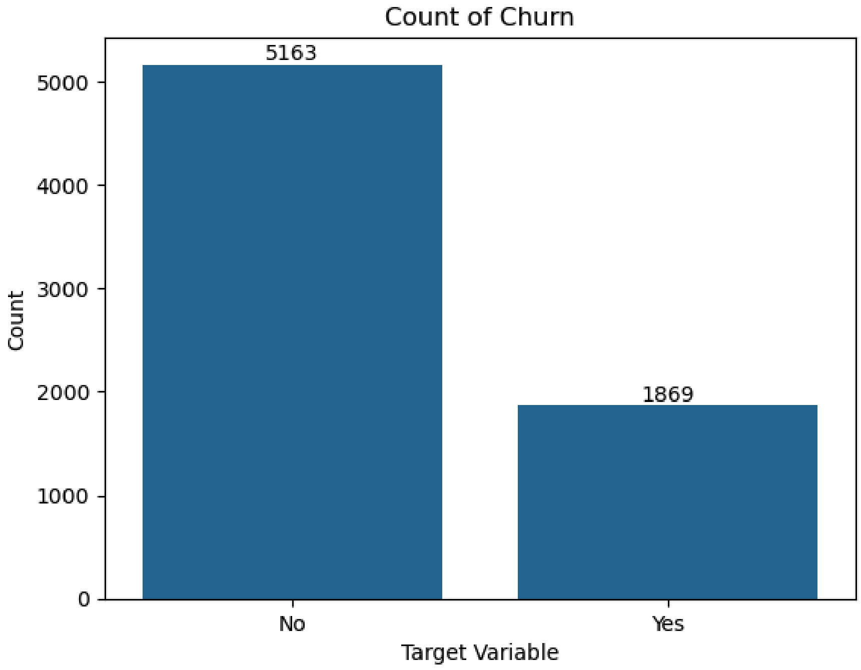
- Services that each customer has signed up for – phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies
- Customer account information – how long they've been a customer, contract, payment method, paperless billing, monthly charges, and total charges
- Demographic info about customers – gender, age range, and if they have partners and dependents
- *Customers who left within the last month – a.k.a* ***Churn***

The data arrived with data type mismatches. ‘Total Charges,’ normally presented in either float or integer form, was an object. ‘Senior Citizens’, was listed as a float, despite clearly being a category. I wrangled both columns to their correct data type. However, this resulted in 11 NaN values for total charges. Since it was only 11 values out of 7,043 (0.1 % of the data) I decided to drop those values.

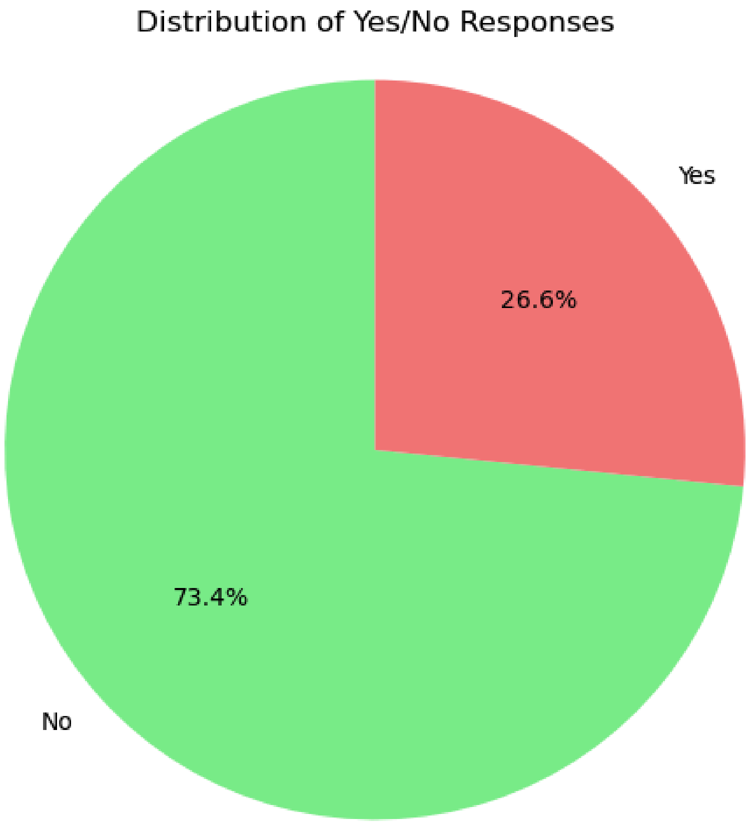
With my data set clean, I proceeded to explore the data of 7,032 rows and 21 columns.

Exploratory Data Analysis:

The target variable is Churn. I wanted to take count of how many customers in the last month churned vs. how many did not churn.

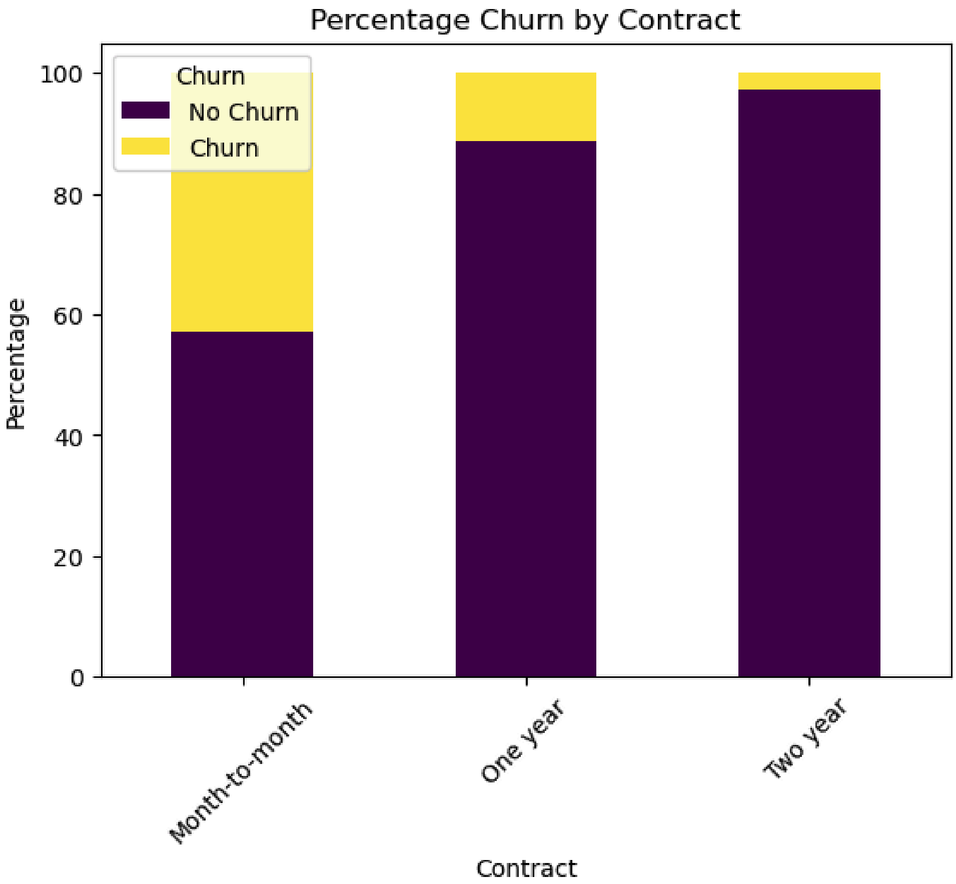


From the data, one can eyeball that Telco experienced about 1/4 of its customers churned last month. A pie chart reveals not only a better visualization, but also an actual ratio.



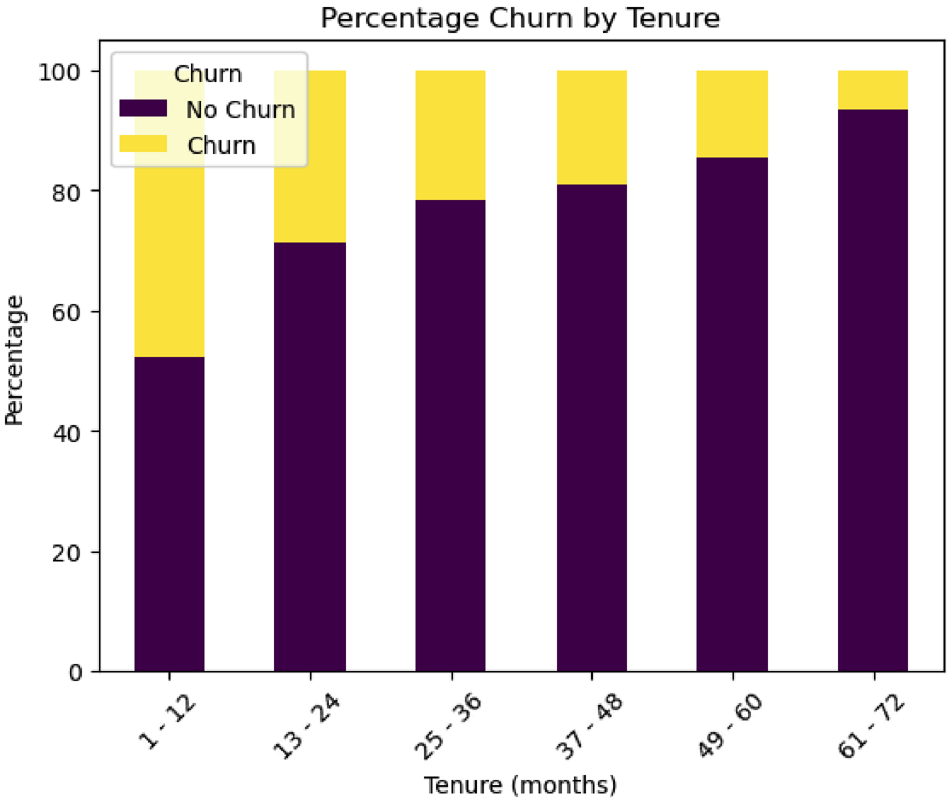
Telco experienced a churn of 26.6 percent. I explored which variables cause the ratio of yes/no to significantly change. In other words, I wanted to see which variables cause the ratio of churn to change from the approximately 1/4 ratio.

In my bivariate analysis, several features produced interesting results.



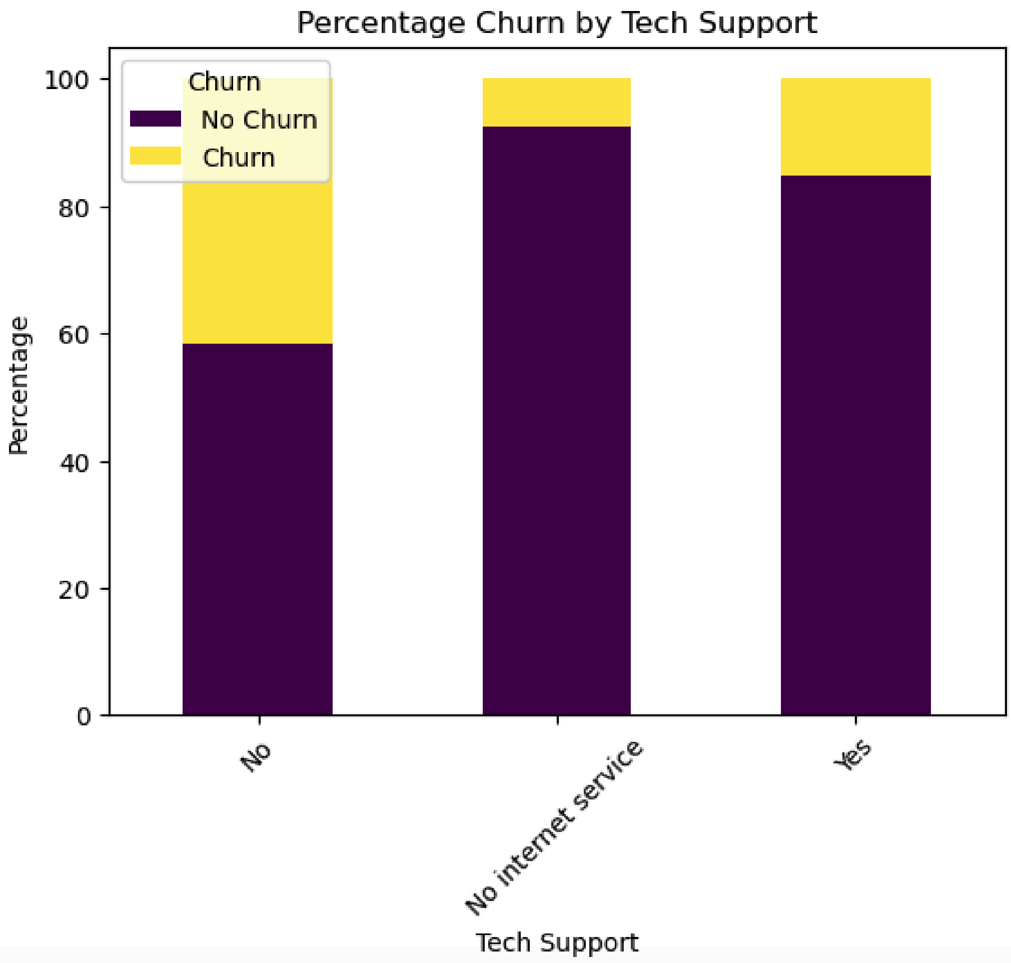
From the graph, one can see that as the Contract term gets longer (i.e., goes from “month-to-month” to “one year” to “two year”), the churn significantly decreases. This makes sense because if a customer is locked into a longer contract, they are more likely not to leave. There might also be a termination fee, further incentivizing the customer to remain with the company.

Another strong correlation we saw was with Tenure.

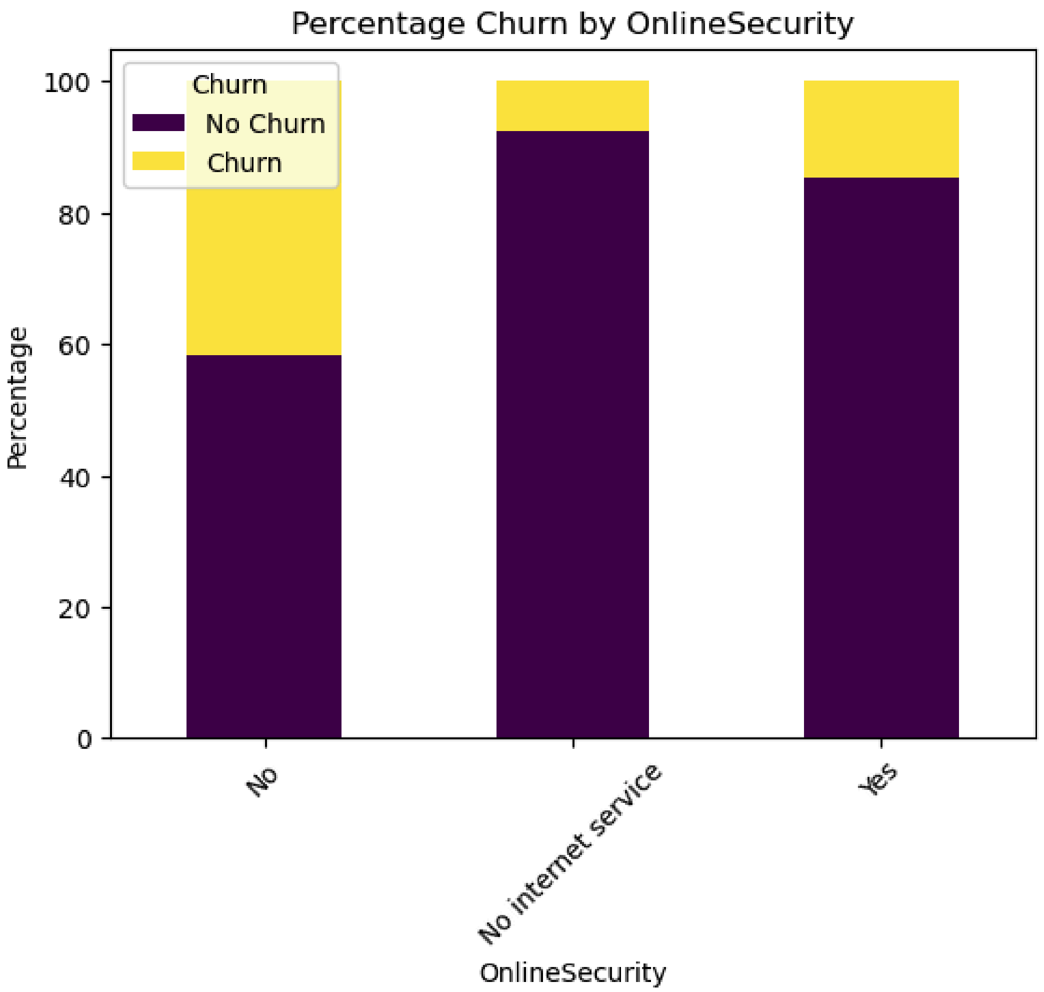


From the graph, one can see that as tenure (in months) increases, rate of churn significantly decreases. This makes sense because long-time customers are most likely satisfied with their service and feel no need to leave the company. Whereas newer customers may not feel the same loyalty and would more easily change service providers.

Tech Support also showed a unique relationship to Churn.

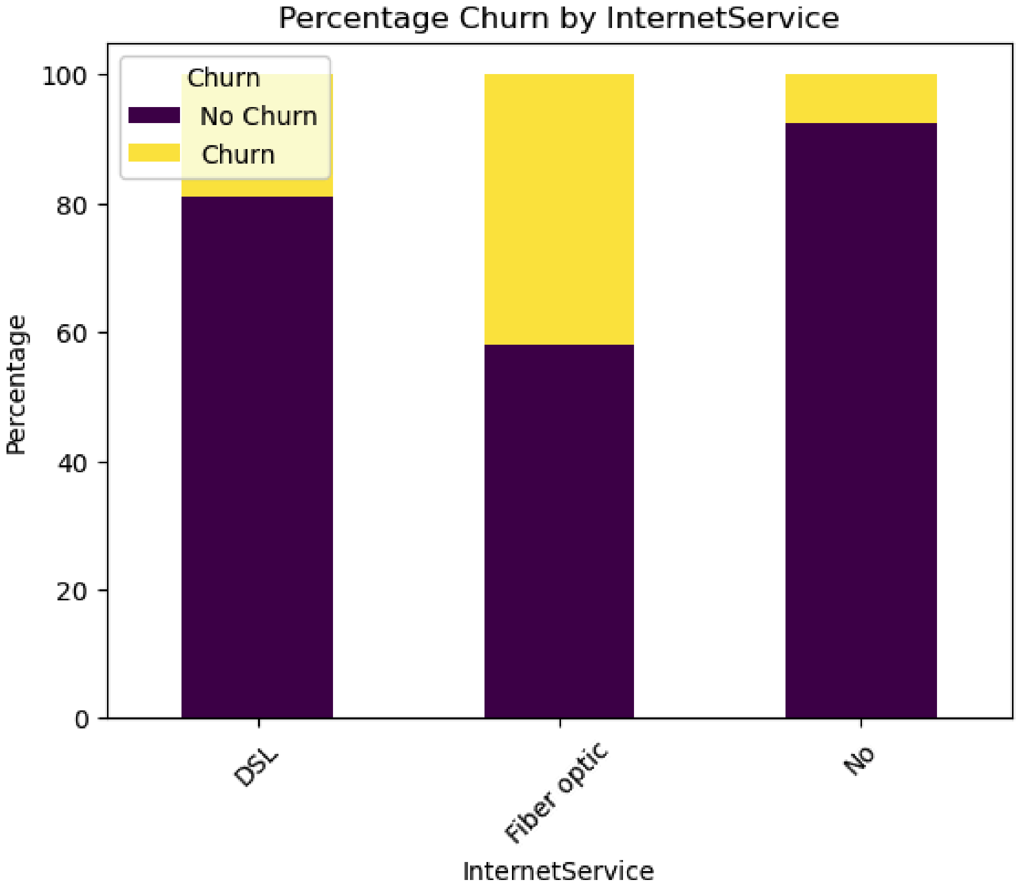


From the graph, one can see that if a customer received tech support from Telco, they would be less likely to leave the company. This makes sense because customers with tech support rely on Telco to help them with their day-to-day telecommunication needs. If they were to leave the company, they would also need to figure out where to get their new technical support services.



Online Security also showed an interesting correlation to churn. Those customers who subscribed to Telco's Online Security service were less likely to churn. Similar to Tech Support, Online Security is an additional service customers would have to seek out if they left the company. Thus, this feature is an incentive to stay with Telco.

Lastly the type of Internet Service the customers received uniquely affected likeliness to churn.



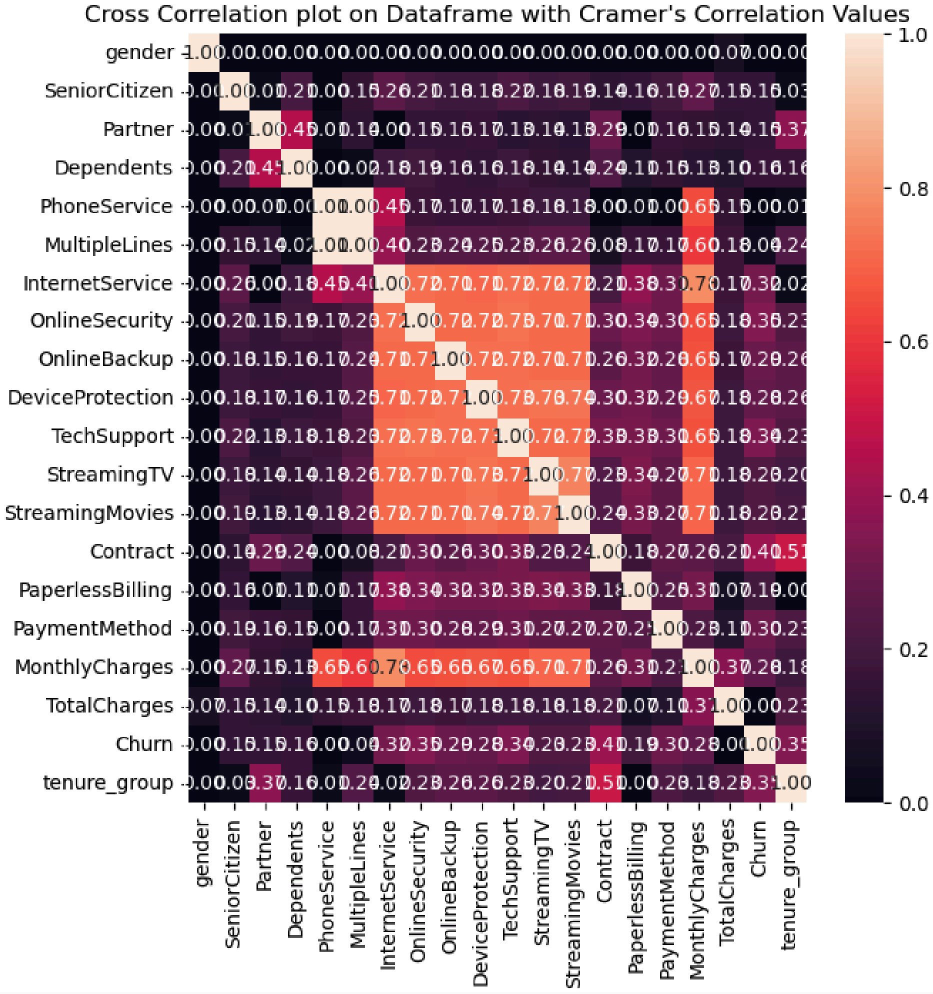
From this graph, one can see that customers with DSL were less likely to churn than those with fiber optic. This may be due to Telco being an older company and not fully having their fiber optic service up to snuff. In other words, there may be more enticing fiber optic services for those whose homes are adequate to handle such a service.

Cramer’s V Heatmap:

For a general overview of how the other categorical features related to our target, Churn(itself a categorical feature), I performed a Cramer’s V correlation.

Cramer’s V correlation is used to measure the association between two attributes. Its value varies from 0, which signifies no relationship between variables, to 1, a complete association between variables).

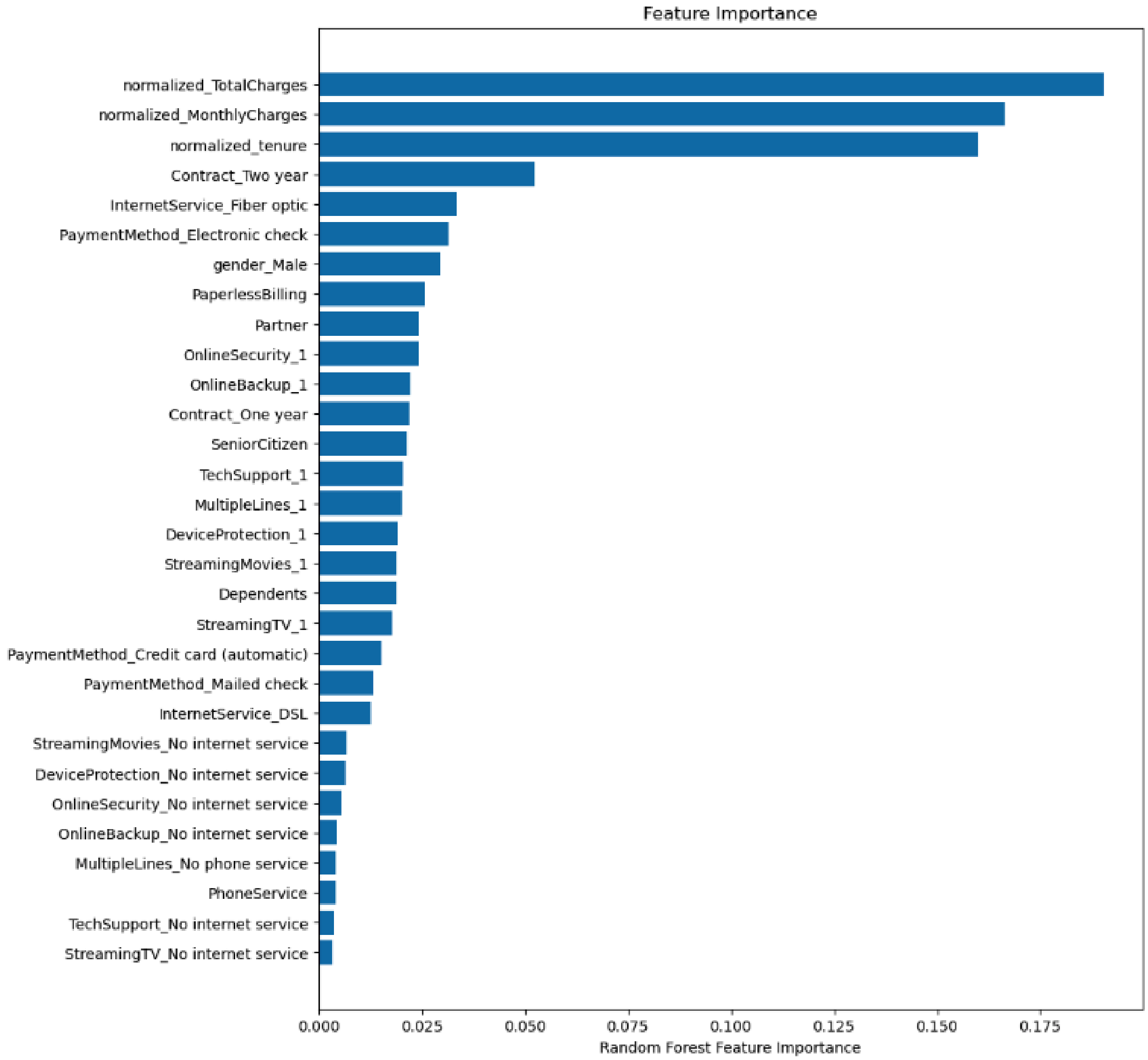
While Cramer's V itself is a measure of effect size (strength of association), its computation directly utilizes the chi-square test statistic. Thus, our associations should be statistically significant.



From the heatmap, one can see the greater correlations between 'Churn' and 'Contract', 'Online Security', 'Tenure', 'Tech Support', and 'Internet Service.' This is consistent with my charts above.

Feature Importance:

Feature importance was assessed using a Random Forest classifier. Before performing this analysis, all continuous variables were scaled using Min-Max normalization, and all categorical variables were one-hot encoded. This preprocessing ensured that all features were on the same scale, ranging from 0 to 1, making the data ready for modeling.



My most important features were 'normalized_MonthlyCharges', 'normalized_TotalCharges', 'normalized_tenure', 'Contract_Two year', 'InternetService_Fiber optic', 'gender_Male', 'PaymentMethod_Electronic check', 'OnlineSecurity_1', 'PaperlessBilling', 'Contract_One year', and 'Partner.'

This step was valuable because identifying the most important features helps us understand which variables are the strongest predictors of churn. At the same time, recognizing weaker features can help reduce noise and overfitting, improving the model's ability to generalize to new data.

I evaluated all features, the top 10 features, and the top 5 features using a Random Forest classifier. I used F1 score as my evaluation metric because it balances recall (catching actual churners) and precision (avoiding unnecessary outreach to non-churners).

From testing different feature sets, the highest F1 score was achieved using all features. This indicates that all variables contribute meaningfully to the model and that no features introduce significant noise. Therefore, I will include all features in the final model.

Modeling:

Since this is a **classification problem**, I evaluated the following models: **Logistic Regression, Random Forest, K-Nearest Neighbors,** and **XGBoost**.

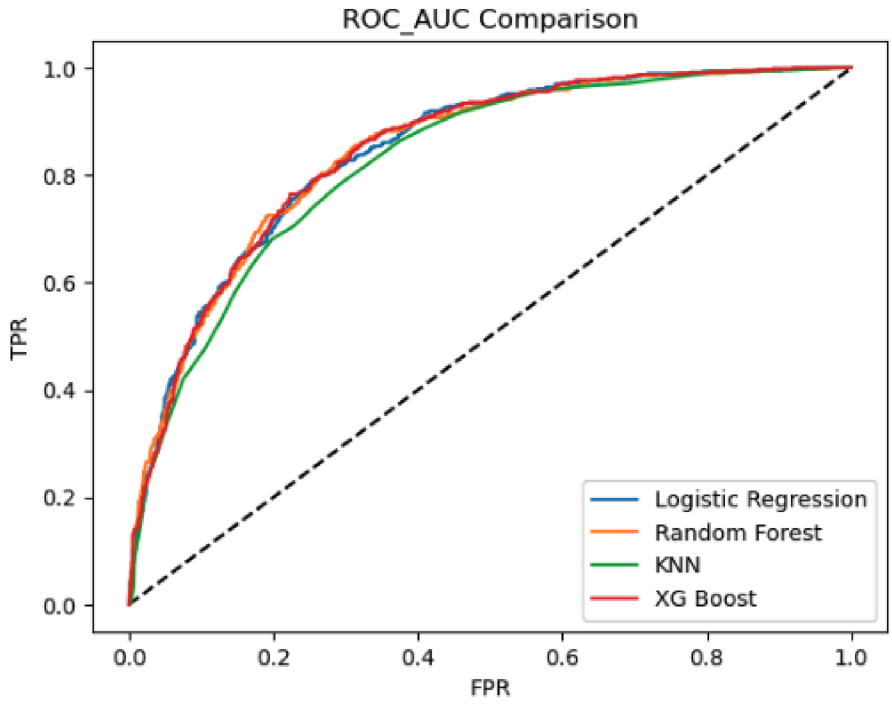
One challenge during modeling was the **class imbalance** in the dataset. As noted in the EDA section, the ratio of non-churners to churners is approximately **3:1**, which could bias models toward predicting non-churn.

For each model, I performed **hyperparameter tuning using GridSearchCV** to identify the optimal parameters.

For evaluation, I used the **ROC-AUC score**. This metric measures how well the model ranks churners higher than non-churners across all possible thresholds. The model with the highest ROC-AUC score was selected as the **best-performing model**.

Telco Churn Modeling Comparison

Name	Best Parameters	ROC_AUC Score
Logistic Regression	'C': 10, 'max_iter': 5000, 'penalty': 'l2', 'solver': 'saga'	84.37%
Random Forest Classifier	'bootstrap': True, 'max_depth': 10, 'max_features': 'log2', 'min_samples_leaf': 4, 'min_samples_split': 2, 'n_estimators': 200	84.41%
K-Nearest Neighbors	'metric': 'manhattan', 'n_neighbors': 29, 'weights': 'uniform'	82.55%
XG Boost	'colsample_bytree': 0.8, 'gamma': 5, 'learning_rate': 0.1, 'max_depth': 4, 'min_child_weight': 5, 'n_estimators': 1000, 'subsample': 0.8	84.5%



Overall, the CV Scores and ROC_AUC scores ranged in the low to mid 80s. This means that all four models had good to excellent discriminatory power.

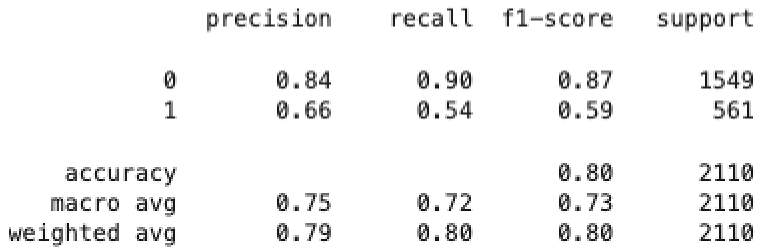
However, the model that scored the highest was XG Boost with an ROC_AUC of 84.5%. Thus, XG Boost is my best model.

Further assessment via confusion matrix and classification report below.

Confusion Matrix:

	Actual 1	Actual 0
Predicted 1	1391	158
Predicted 0	259	302

Classification Report:



```
              precision    recall  f1-score   support

           0       0.84      0.90      0.87      1549
           1       0.66      0.54      0.59       561

    accuracy                           0.80      2110
   macro avg       0.75      0.72      0.73      2110
weighted avg       0.79      0.80      0.80      2110
```

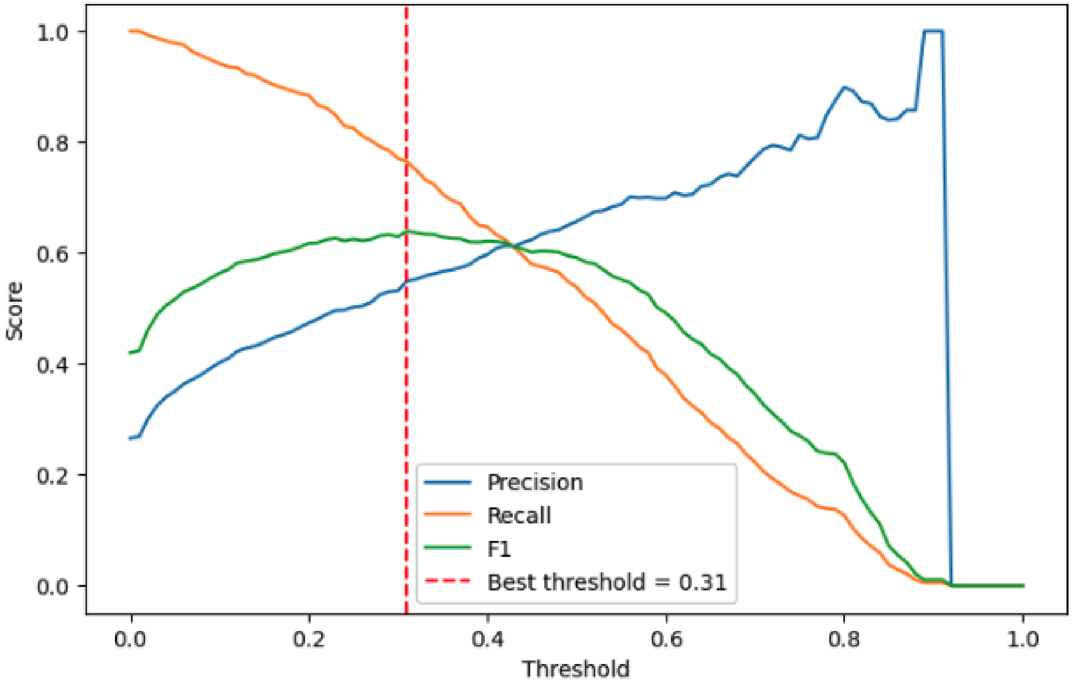
Overall, the model achieved 80% accuracy. However, because the dataset is imbalanced, accuracy alone can be misleading, so additional metrics must be considered.

For class 0 (No Churn), the model performed very well: precision, recall, and F1 score were all strong, indicating that the model is highly effective at identifying customers who are not likely to churn.

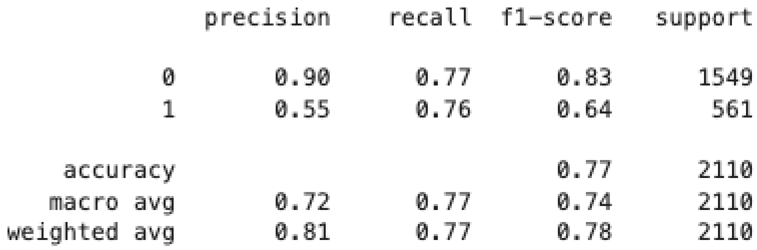
In contrast, the model struggled with class 1 (Churn), correctly identifying only 54% of actual churners. This is expected, as churners are the minority class, making recall the primary challenge. The model missed a significant number of churn cases.

To address this, I optimized the decision threshold to improve performance for class 1, prioritizing recall so that more at-risk customers could be correctly identified and targeted for retention efforts.

Precision-Recall Chart



From the chart, I concluded that the best probability threshold is 0.31. I added this threshold to my baseline XG Boost model and assessed it via a new classification report.



```
              precision    recall  f1-score   support

           0       0.90      0.77      0.83      1549
           1       0.55      0.76      0.64       561

    accuracy                           0.77      2110
   macro avg       0.72      0.77      0.74      2110
weighted avg       0.81      0.77      0.78      2110
```

After optimizing the decision threshold, recall for class 1 (Churn) improved dramatically to 76%, allowing the model to correctly identify many more true positives.

The F1 score also increased from 59% to 64%, reflecting a better balance between precision and recall. However, precision decreased, meaning the model now generates more false positives. This is a typical trade-off when prioritizing recall.

In churn prediction, recall for churners is more important than precision. Missing a customer who is likely to churn means losing the opportunity to intervene and retain them. On the other hand, a lower precision—sending a retention offer to a customer who wouldn’t have churned—has a relatively small economic cost.

Conclusion:

A hyperparameter-tuned XGBoost churn model was developed using Telco’s customer data, with the probability threshold adjusted to address class imbalance. The model identifies at-risk customers with approximately 80% accuracy, enabling proactive and targeted retention efforts.

With this model, Telco can focus resources on customers most likely to churn and understand the key drivers behind their risk. High-risk customers can receive personalized interventions such as tailored service plans, loyalty rewards, or temporary discounts. By prioritizing recall, the model ensures that customers who are genuinely at risk are not overlooked, maximizing the effectiveness of retention initiatives.

Three concrete recommendations for Telco:

1. Target high-risk customers with personalized retention campaigns – use the model’s insights to deliver service packages, loyalty programs, or discounts to customers most likely to churn.

2. Monitor and refine the model regularly – update the model with new data and adjust thresholds as needed to maintain high recall and improve precision, ensuring retention efforts remain effective.

3. Optimize resource allocation – focus retention budget and outreach efforts on the top segment of at-risk customers, minimizing spend on low-risk customers while maximizing ROI.

Future enhancements, such as alternative data scaling or resampling techniques, could further improve both precision and recall. These improvements would reduce unnecessary outreach to low-risk customers while continuing to capture those most likely to churn, ultimately increasing customer lifetime value, lowering acquisition costs, and maximizing ROI on retention efforts.