PROJECT REPORT

CUSTOMER CHURN PREDICTION FOR TELECOM COMPANY

ALY 6040 Data Mining Applications

NORTHEASTERN UNIVERSITY

By
Ankit Bhalekar
Apurwa Sontakke
Vedant Wagh
&
Lazaree Worlikar

DATE OF SUBMISSION: 10-04-2023

Introduction:

The primary objective of this project is to investigate the factors that influence customer attrition within a telecommunications company and develop predictive models aimed at retaining customers. For this purpose, we will be using predictive modelling techniques such as Logistic Regression, Decision tree and Random Forest. Along with predicting customer who are more likely to churn, we will also identify important factors that lead to customer churn and give recommendations to the telecom company basis our findings.

Analysis:

Data Cleaning

In our project, we began by checking the data quality. Upon importing the dataset, we found it contained information for 7043 customers with 21 features. There were no duplicate rows. We inspected the dataset’s initial entries to understand its structure. Of the customers, 1869 (26.6%) had churned, while 5163 (73.4%) had not. We then ensured that all variables had appropriate formats. The ‘senior citizen’ variable, initially in integer format, was converted to a category with ‘Yes’ for 1 and ‘No’ for 0. The ‘Total charges’ data type was changed from object to numerical, resulting in correctly formatted columns.

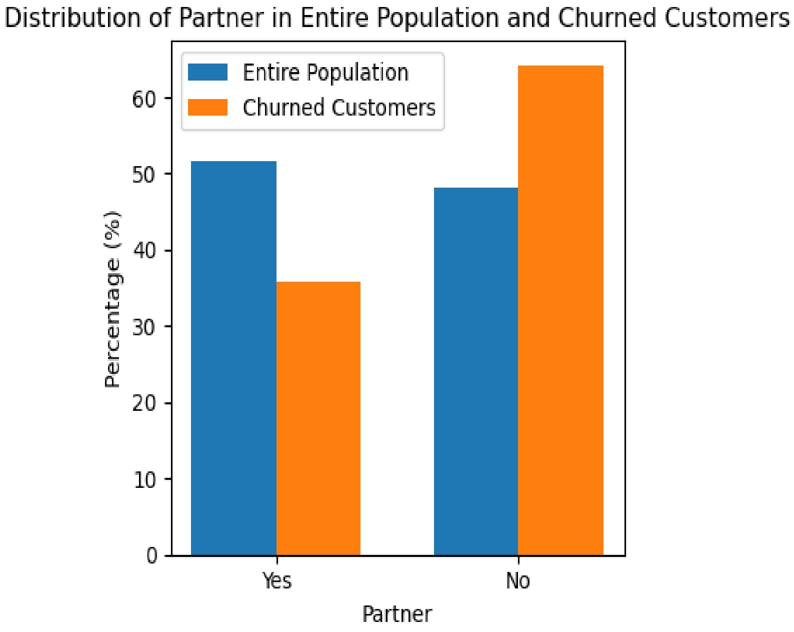
We, then examined missing values. Only 0.15% of instances had 11 missing values in the ‘Total Charges’ variable. To handle this, we decided to drop the corresponding rows as it was just 0.15%, leaving us with a complete dataset free of missing data.

Our dataset contained three numerical columns. To check for outliers, we used boxplots, the results of which are detailed in Appendix A. We observed that there were no outliers in these columns.

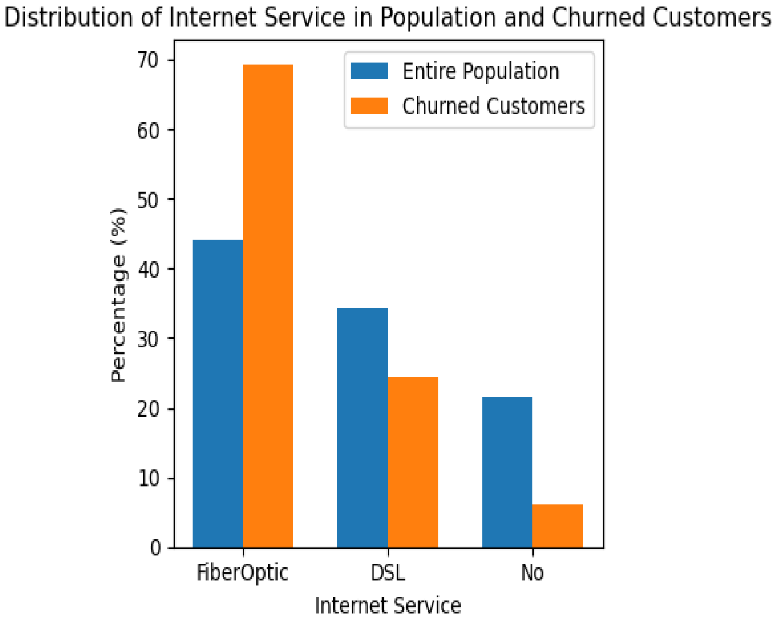
EDA

We first examined the gender variable, noting that it showed a roughly equal distribution of male (about 50.4%) and female (about 49.5%) in the entire population, with similar proportions in the churned dataset. A chi-square test confirmed that there was no association between gender and churn as seen in Appendix B.

However, the distribution of the 'partner' variable appeared different between the overall population and the churn subset, suggesting significance. This could be due to additional benefits offered to couples who are both customers. A chi-square test supported our claim, establishing an association between 'partner' and churn as seen in Appendix B.



Additionally, we observed lower churn rates among customers with no internet service, and a chi-square test verified the relationship between internet service and churn as seen in Appendix B. Notably, the highest churn was among



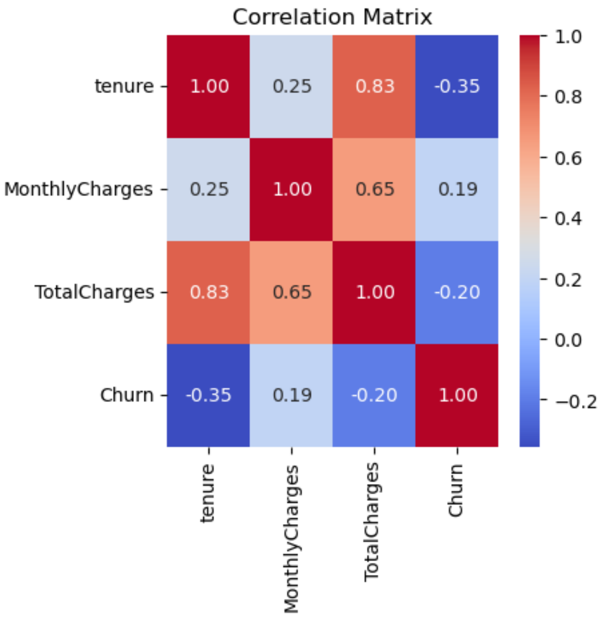
customers who opted for Fiber Optic, potentially indicating issues with service quality.

Remarkably, outstanding customer service plays a vital role in a company's success. Customer service connects businesses with their clients, enhancing relationships. From our analysis, 77.3% of those who churned did not have tech support. A chi-square test confirmed the significant association between tech support and churn as seen in Appendix B.

Data Modelling

After cleaning our data and performing the initial exploratory data analysis, we found that there is no association between “Gender” and “Churn”; there is association between having a partner and churning; there is association between having an internet service and churning, people having Fiber Optic were more likely to churn and there was association between Tech support service and churning. Now it was time to build a predictive model that will predict a customer will churn or not with the help of our data. Since this was a binary classification problem (Churn Yes/Churn No) we decided to start with Logistic Regression because logistic regression is well-suited for such problems as it predicts the probability of a customer belonging to one of the two classes. Before starting with modelling, we will do one hot to convert categorical variables into binary (0 or 1) numerical values and standard scaling for our numerical variables to help mitigate the impact of numerical variables with different units and scales on the model’s performance. Without standardization, variables with larger values may overshadow others in the model’s calculations, leading to bias. Before modelling we also checked the correlation between the variables to ensure that our independent variables are not highly correlated and these were the results:

We can see that variables "tenure" and "TotalCharges" have a correlation coefficient of 0.83 which is quite high. Also, "MonthlyCharges" and "TotalCharges" have high correlation of 0.65. Hence, we will drop "TotalCharges" variable and proceed with modelling.



Baseline Logistic Regression Model

We first started by creating a training (80%) and a testing (20%) dataset to help us validate our model. We set the random state to 90 to compare different models on the same dataset under the same conditions. Our initial logistic regression had the following confusion matrix:

		Actuals	
		Churn	Didn't churn
Predicted	Churn	935	114
	Didn't Churn	152	206

Our model has accuracy of 81.09%, precision of 64.34%, recall of 57.54% and F1-Score of 60.76%. In our case, recall is more important than accuracy and precision because it emphasizes the ability of the predictive model to identify and capture as many potential churners as possible. Maximizing recall will help the telecommunications company proactively address customer churn, reducing revenue loss and enhancing customer retention efforts. Our recall is very well, indicating that our model has high false negatives which is not good for our business, as we are missing on identifying a lot of potential customers who will churn.

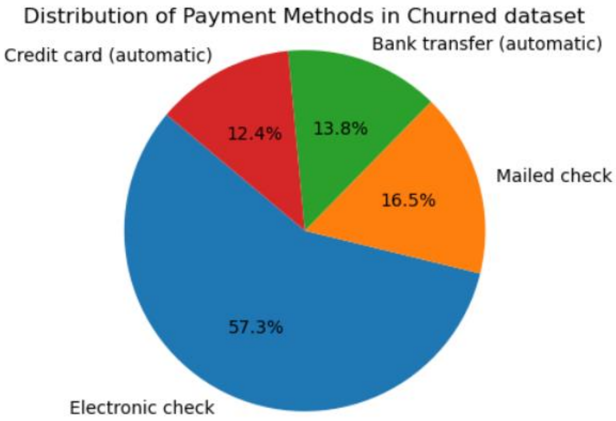
Optimized Logistic Regression

We then decided to fine tune the model by tweaking the threshold. Now instead of 0.5, we lowered the threshold to 0.3, as predicting churning was our priority. When we decreased the threshold, the model became more lenient in classifying observations as positive. One drawback of this is that it can also increase the number of false positives. We shall have a look on the results of this model and then make a call if we can proceed with this model or not. The confusion matrix was as follows:

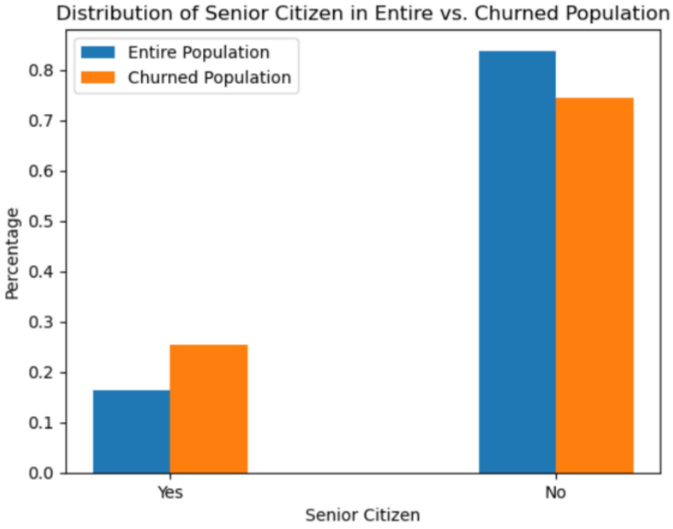
		Actuals	
		Churn	Didn’t churn
Predicted	Churn	811	238
	Didn’t Churn	70	288

The accuracy of this model was 78.10%, precision was 54.75%, recall was 80.44% and F1-Score was 65.15%. We can see that the recall of our model has significantly increased by approximately 23%. However, our accuracy has dropped by approximately 4% and false positives have significantly increased.

The significant variables with negative coefficients were Tenure (-0.8241), Contract (One year) (-0.294), Contract (Two years) (-0.568), PaymentMethod_Credit card (automatic)(-0.0264). This means that as a customer's tenure (the length of time they have been with the company) increases, they are less likely to churn. The longer a customer stays with the company, the lower the probability of them churning. This result is consistent with a common understanding in many industries. Customer retention tends to improve as customers stay with a company for a longer period. Customers with longer-term contracts are less likely to churn compared to those with shorter-term or month-to-month contracts. Businesses can consider promoting or encouraging customers to use credit card payments, especially in an automatic or recurring manner, as it appears to be associated with higher customer retention. We can see this in the pie chart too.



Variables like 'seniorcitizen_Yes' has a positive coeffient (0.0833) and is significant.This implies that they are more likely to churn. We can verify this in the graph too. Businesses may think about introducing special discounts for them as strategy to retain them. The choice of payment method, specifically electronic checks, is estimated to increase the likelihood of customers churning by approximately 14.19% when compared to customers using other payment methods. Our telecom should be aware that customers who opt for electronic check payments may have a higher churn rate. It can consider strategies to incentivize customers to choose alternative payment methods or to improve the customer experience for those using electronic checks. The detailed results of the variables are in AppendixC.



The variables like “Partner”, “Tech Support” and “Internet Service” which we found important distinguisher between churned customer and retained customer during exploratory data analysis (EDA) is not found significant by logistic regression. This could be due to Logistic regression starts with a simpler model and then adds complexity if necessary. It might not include certain variables if they do not significantly contribute to explaining the variance in the dependent variable (churn) based on statistical tests. EDA, on the other hand, identifies patterns without assuming specific functional relationships.

Decision Tree

Since logistic regression assumes linear relationship between the independent variables and the log-odds of the dependent variable, we will now try using decision tree to predict whether a customer will churn or not. Decision trees are a non-linear modeling technique that does not impose linear assumptions on the data. The results of our decision tree are as follows:

		Actuals	
		Churn	Didn’t churn
Predicted	Churn	827	206
	Didn’t Churn	183	191

Accuracy: 72.3%

Precision: 48.11%

Recall: 51.06%

As we can see our recall is just 51.06% which suggests that the model is able to identify only half of the churners in the dataset. We then checked the feature importance as per this model. The detailed results for this are in Appendix B. “MonthlyCharges”, “tenure”, “Contract”, “PaymentMethod”, “OnlineSecurity”, “InternetService” were among the top six important features.

Optimized Decision Tree

To improve the performance of the model we decided to use Grid Search CV for finding the best combination of hyperparameters for our model. After running the Grid Search CV, the Best Hyperparameters were: ‘class_weight’ as Balanced, this hyperparameter specifies how class weights are balanced. In this case, it is set to ‘balanced’, which means that the decision tree algorithm will automatically adjust the class weights based on the distribution of the target classes in the dataset. This is useful when the classes are imbalanced, as it gives more weight to the minority class during the training to prevent the model from being biased toward the majority class.

The ‘criterion’ as ‘entropy’, which means that the entropy is used as the criterion to evaluate the splits. Entropy is a measure of impurity; it quantifies the disorder or randomness within a node. The Decision Tree algorithm selects splits that maximize the reduction in entropy, resulting in nodes with more homogeneous class distributions. ‘max_depth’ as 10, which means that the decision tree is limited to having a depth of 10 nodes (including the root node). This hyperparameter controls the complexity of the tree and helps prevent overfitting. A smaller value like 10 can make the tree more interpretable and less prone to overfitting. ‘min_samples_leaf’ as 30, meaning that a leaf node must have at least 30 samples. This hyperparameter also helps control the depth of the tree and can prevent it from growing too deep, which can lead to overfitting. ‘min_samples_split’ as 10 indicating that an internal node must have at least 10 samples to be considered for splitting. Similar to min_samples_leaf, this hyperparameter helps control the growth of the tree and can prevent it from splitting too early, which can also mitigate overfitting. The results of decision tree after tuning our parameters were

		Actuals	
		Churn	Didn’t churn
Predicted	Churn	729	304
	Didn’t Churn	90	284

Accuracy = 72%

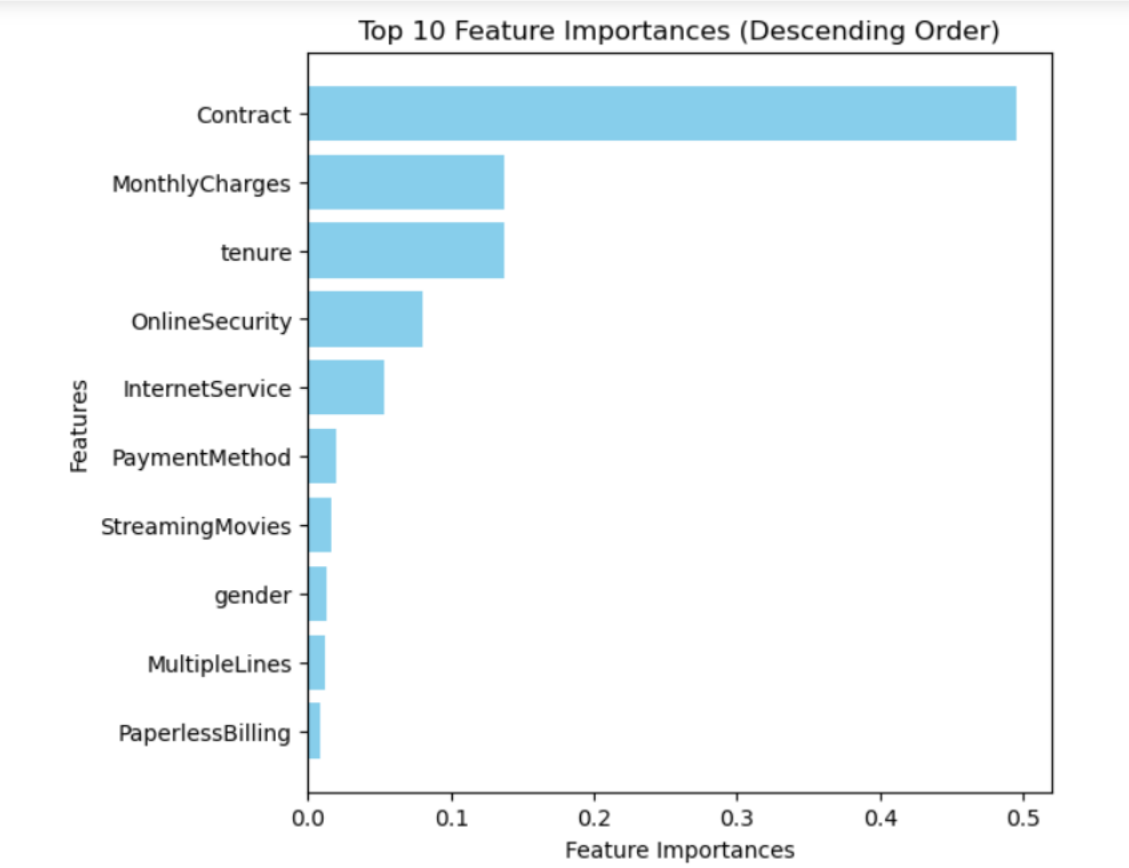
Precision = 48%

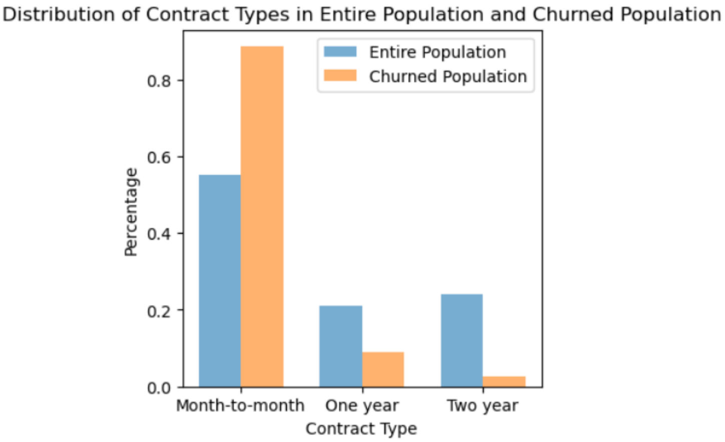
Recall = 76%

F1-Score = 59%

The accuracy of our model is 72% but our recall has significantly increased by approximately 25% to 76% which suggests that the model is able to identify 76% of the churners in the dataset.

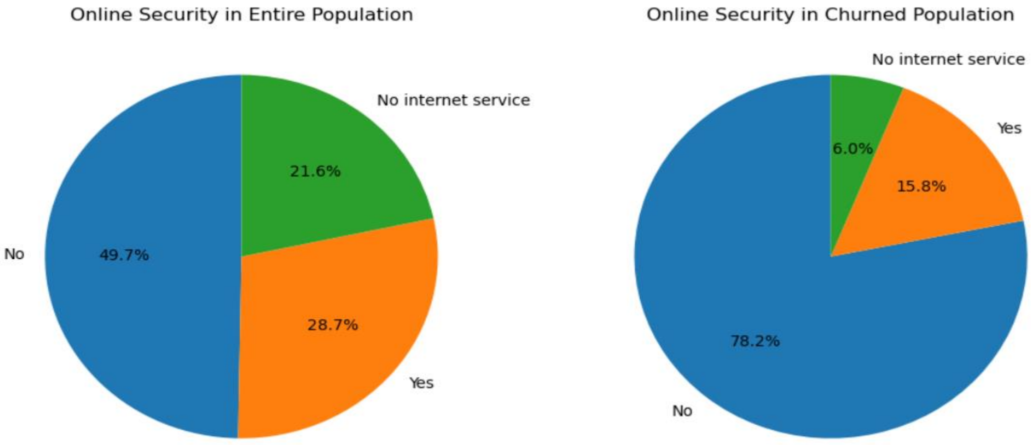
The top 10 important variables are shown below. The details are in Appendix D.





We can see that Contract Duration is the most critical factor affecting churn is the length of customer contracts. The “Contract” feature has the highest importance, indicating that it plays a crucial role in predicting customer churn. A high importance suggests that customer’s contract types have a substantial impact on their decision to churn. Longer contract durations, such as two-year contracts, are associated with lower churn rates as in the graph above.

“MonthlyCharges” is also a significant feature, with relatively high importance. This suggests that the amount customers are charged each month is an important factor in predicting churn. Higher monthly charges may lead to higher churn rates, indicating that pricing strategies may influence customer retention. “Tenure” which represents the length of time a customer has been with the company, has a similarly high importance. Longer tenure is typically associated with reduced churn rates, highlighting the importance of customer loyalty.



“OnlineSecurity” plays a notable role in predicting churn. Customers with online security services are less likely to churn as seen in above graph, indicating that providing security features can improve customer retention. The type of “InternetService” is moderately important. In particular, fiber optic service might be associated with higher churn rates which we had seen during EDA. This suggests that service quality or pricing related to internet services can influence churn. “Gender” has a lower importance, indicating that gender has a relatively minor impact on churn prediction.

Baseline Random Forest

We will now try Random Forest because decision trees can be prone to overfitting, creating models that are overly complex and don't generalize well. Random Forest mitigates this issue by aggregating the results of multiple trees, leading to more robust models. Also, random forests are effective at capturing non-linear relationships between features and the target variable unlike logistic regression which assumes linear relationship between the independent variables and the log-odds of the dependent variable. The results of our model were as follows:

		Actuals	
		Churn	Didn’t churn
Predicted	Churn	919	114
	Didn’t Churn	198	176

Accuracy = 77.82%

Precision = 60.68%

Recall = 47%

F1-Score = 53%

Optimized Random Forest

The recall in this case is not that good, so we tried adjusting the parameters using Grid Search CV and the best hyperparameters were {“class_weight”: “balanced”, “max_depth”: 20, “min_samples_leaf”: 30, “min_samples_split”: 10, “n_estimators”: 200}. We have included 200 decision trees in the Random Forest ensemble. A larger number of trees often leads to a more stable and generalizable model. It can reduce the risk of overfitting. The results of this optimized model were:

		Actuals	
		Churn	Didn’t churn
Predicted	Churn	744	289
	Didn’t Churn	78	296

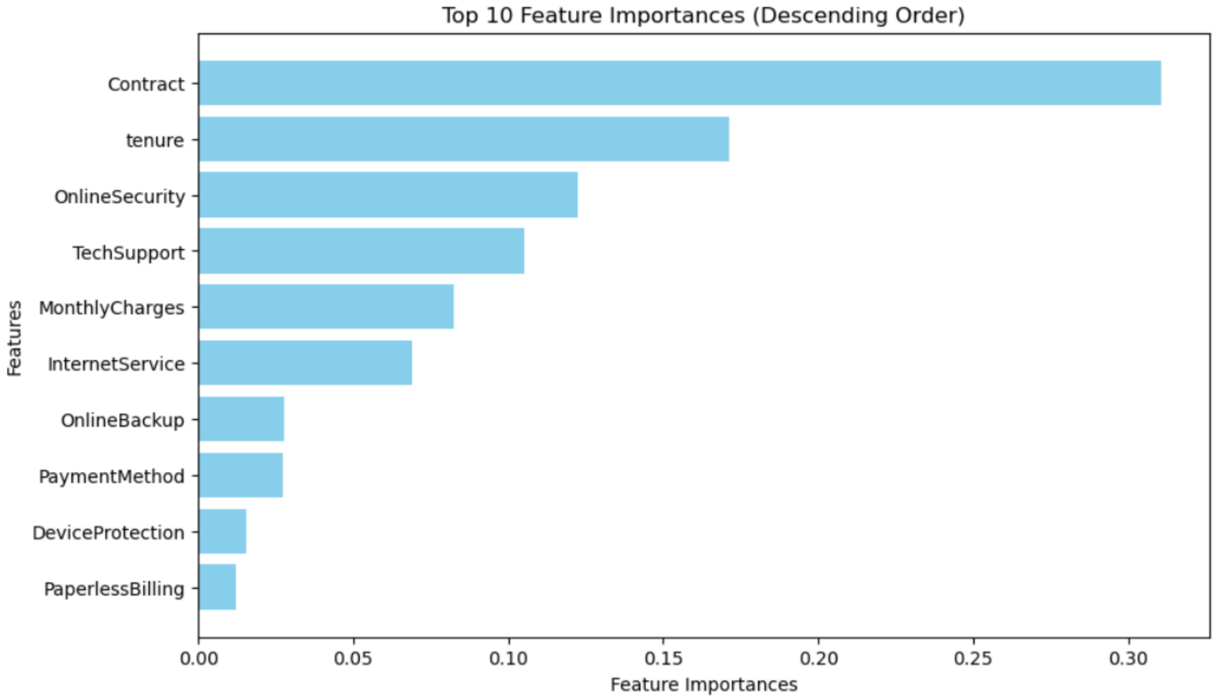
Accuracy = 74%

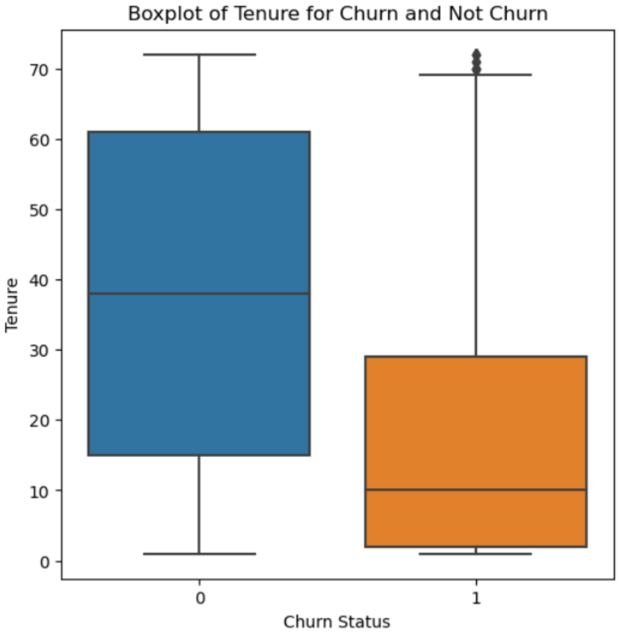
Precision = 51%

Recall = 79%

F1-Score = 62%

The feature importance with their details are in Appendix E. We shall now have a look on top 10 features:





As seen in decision tree, random forest also suggests "Contract", "tenure", "OnlineSecurity", "MonthlyCharges", "InternetService" are important variables in determining customer churn. It also suggestes "TechSupport" as one of the important variables while predicting customer churn. We can see from the boxplot too, the median tenure for people who churned is less than 10 months whereas for those who didn't is approximately 39 months. Customers access to technical support services significantly impacts their likelihood of churning as seen in our EDA.

Unlike seen in our Exploratory Data Analysis, our predictive models suggest that having a partner does not impact churning. Predictive models can identify interaction effects between variables. For example, the presence of a partner might have a different impact on churning when combined with other factors, such as contract type or monthly charges. These interactions can be challenging to detect through EDA alone. EDA had also suggested that churn is high among people who have opted for Fiber Optic internet service and even random forest and decision trees has suggested "Internet Services" as important variable in determining churn.

Comparing Model Performance

Model	Accuracy	Precision	Recall	F1-Score
Optimized Logistic Regression	78.10%	54.75%	80.44%	65.15%
Optimized Decision tree	72%	48%	76%	59%
Optimized Random Forest	74%	51%	79%	62%

The Optimized Logistic Regression model outperforms the Decision Tree and Random Forest models in terms of accuracy, precision, and recall. This suggests that the Logistic Regression model is more adept at predicting customer churn. Also, our aim was reducing False Negatives so that we don't miss out on potential churners hence the model with the highest recall (Logistic Regression) would be preferred.

Conclusion:

Our analysis aimed to aid a telecommunications company in identifying the factors that affect customer attrition and constructing predictive models geared toward customer retention. In pursuit of this objective, we harnessed logistic regression, decision trees, and random forest techniques to establish the most precise predictive model, enabling the company to pinpoint churners and prioritize relevant features. After a comprehensive exploration, we determined that the optimized logistic regression, utilizing our standardized dataset, emerged as the most accurate model. Additionally, we recommend that the company should focus on monitoring and potentially revising pricing strategies to minimize monthly charges may help improve customer retention. Also, company can improve the quality of their Fiber Optic internet service and tech support to improve customer experience and thus prevent customer churn. Long-term contracts are crucial for retaining customers, as they are associated with lower churn rates. Companies can actively encourage customers to opt for long-term contracts, such as one-year or two-year plans. This can be done through marketing campaigns, promotional offers, or incentives like discounted rates for extended contracts.

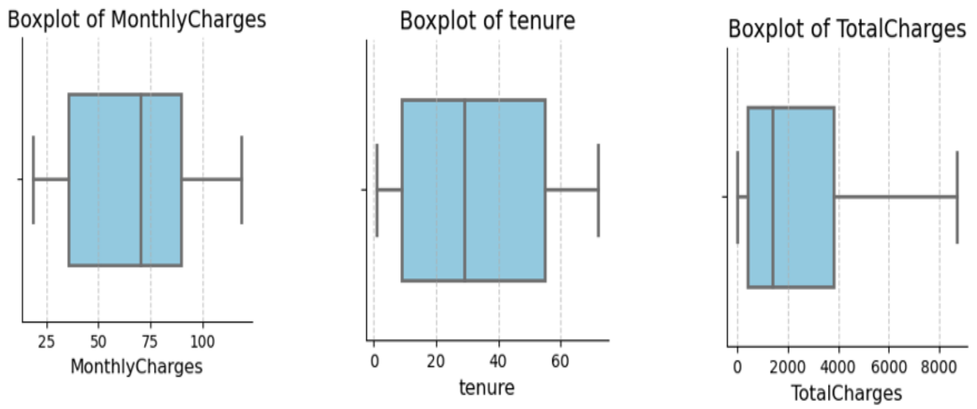
For future analysis we suggest incorporating customer feedback and competitor's pricing data. Feedback obtained from customer surveys and reviews on various platforms can be analyzed for sentiments. This data can reveal specific issues and suggestions directly from customers, helping the company proactively address concerns. Churn rates are often influenced by the pricing strategies of competitors. Accessing and analyzing competitor pricing data can provide insights into how our pricing compares. Understanding the price sensitivity of our customer base can help in setting competitive and retention-focused pricing.

References:

1. Koehrsen, W. (2018, January 17). Random Forest in python. Medium. https://towardsdatascience.com/random-forest-in-python-24d0893d51c0
2. Li, S. (2019, February 27). *Building a logistic regression in Python, step by step*. Medium. https://towardsdatascience.com/building-a-logistic-regression-in-python-step-by-step-becd4d56c9c8

Appendix A

Boxplot for the numerical variables



Appendix B

Test association between Gender and Churn

H0: There is no association between gender and churn.

H1: There is association between gender and churn.

Chi-Squared Statistic	0.4841
p-value	0.4866

Since p-value>0.05 we fail to reject null hypothesis and conclude that there is no association between gender and churn at 5% level of significance.

Test association between Partner and Churn

H0: There is no association between partner and churn.

H1: There is association between partner and churn.

Chi-Squared Statistic	158.7334
p-value	0.000

Since p-value<0.05 we reject null hypothesis and conclude that there is association between having a partner and churn at 5% level of significance.

Test association between having Internet Service and Churn

H0: There is no association between having Internet Service and churn.

H1: There is association between having Internet Service and churn.

Chi-Squared Statistic	732.3096
p-value	0.00

Since p-value<0.05 we fail to reject null hypothesis and conclude that there is association between having Internet Service and churn at 5% level of significance.

Test association between having Tech Support Service and Churn

H0: There is no association between having Tech Support Service and churn.

H1: There is association between having Tech Support Service and churn.

Chi-Squared Statistic	828.19
p-value	0.00

Since p-value<0.05 we fail to reject null hypothesis and conclude that there is association between having Tech Support Service and churn at 5% level of significance.

Appendix C: Logistic Regression Output

	Variable	coef	stderr	z	P>|z|	[0.025	0.975]	Decision
const		-1.5895	0.051	-31.116	0	-1.69	-1.489	Significant
x1	'tenure'	-0.8241	0.065	-12.772	0	-0.951	-0.698	Significant
x2	'MonthlyCharges'	-0.6326	1.07	-0.591	0.554	-2.729	1.464	Not Significant
x3	'gender_Male'	-0.0145	0.036	-0.403	0.687	-0.085	0.056	Not Significant
x4	'SeniorCitizen_Yes'	0.0833	0.035	2.388	0.017	0.015	0.152	Significant
x5	'Partner_Yes'	0.0163	0.043	0.378	0.706	-0.068	0.101	Not Significant
x6	'Dependents_Yes'	-0.0594	0.045	-1.306	0.192	-0.149	0.03	Not Significant
x7	'PhoneService_Yes'	-0.0174	3.58E+06	-4.87E-09	1	-7.01E+06	7.01E+06	Not Significant
x8	'MultipleLines_No phone service'	0.0174	3.58E+06	4.87E-09	1	-7.01E+06	7.01E+06	Not Significant
x9	'MultipleLines_Yes'	0.2106	0.098	2.142	0.032	0.018	0.403	Significant
x10	'InternetService_Fiber optic'	0.7048	0.444	1.586	0.113	-0.166	1.576	Not Significant
x11	'InternetService_No'	-0.0728	3.65E+06	-2.00E-08	1	-7.15E+06	7.15E+06	Not Significant
x12	'OnlineSecurity_No internet service'	-0.0728	3.65E+06	-2.00E-08	1	-7.15E+06	7.15E+06	Not Significant
x13	'OnlineSecurity_Yes'	-0.0875	0.091	-0.966	0.334	-0.265	0.09	Not Significant
x14	'OnlineBackup_No internet service'	-0.0728	3.65E+06	-2.00E-08	1	-7.15E+06	7.15E+06	Not Significant
x15	'OnlineBackup_Yes'	-0.0363	0.094	-0.388	0.698	-0.22	0.147	Not Significant
x16	'DeviceProtection_No internet service'	-0.0728	3.65E+06	-2.00E-08	1	-7.15E+06	7.15E+06	Not Significant
x17	'DeviceProtection_Yes'	0.0483	0.094	0.513	0.608	-0.136	0.233	Not Significant
x18	'TechSupport_No internet service'	-0.0728	3.65E+06	-2.00E-08	1	-7.15E+06	7.15E+06	Not Significant
x19	'TechSupport_Yes'	-0.1321	0.092	-1.441	0.15	-0.312	0.048	Not Significant
x20	'StreamingTV_No internet service'	-0.0728	3.65E+06	-2.00E-08	1	-7.15E+06	7.15E+06	Not Significant
x21	'StreamingTV_Yes'	0.2543	0.178	1.428	0.153	-0.095	0.603	Not Significant
x22	'StreamingMovies_No internet service'	-0.0728	3.65E+06	-2.00E-08	1	-7.15E+06	7.15E+06	Not Significant
x23	'StreamingMovies_Yes'	0.2726	0.178	1.53	0.126	-0.077	0.622	Not Significant
x24	'Contract_One year'	-0.294	0.048	-6.067	0	-0.389	-0.199	Significant
x25	'Contract_Two year'	-0.5683	0.082	-6.969	0	-0.728	-0.408	Significant
x26	'PaperlessBilling_Yes'	0.1545	0.04	3.817	0	0.075	0.234	Significant
x27	'PaymentMethod_Credit card (automatic)'	-0.0264	0.052	-0.511	0.609	-0.128	0.075	Significant
x28	'PaymentMethod_Electronic check'	0.1419	0.049	2.884	0.004	0.045	0.238	Significant
x29	'PaymentMethod_Mailed check'	0.0157	0.053	0.299	0.765	-0.088	0.119	Significant

Appendix D:

Decision tree output

Feature Importances	
MonthlyCharges	0.2766
tenure	0.18785
Contract	0.17408
PaymentMethod	0.05943
OnlineSecurity	0.0485
InternetService	0.03383
Partner	0.03126
Dependents	0.02678
PaperlessBilling	0.02537
gender	0.02203
SeniorCitizen	0.02021
MultipleLines	0.01832
OnlineBackup	0.01796
DeviceProtection	0.01555
StreamingTV	0.01318
TechSupport	0.01305
StreamingMovies	0.01284
PhoneService	0.00317

Optimized decision tree output

Feature Importances (Descending Order)	
Contract	0.50
MonthlyCharges	0.14
tenure	0.14
OnlineSecurity	0.08
InternetService	0.05
PaymentMethod	0.02
StreamingMovies	0.02
gender	0.01
MultipleLines	0.01
PaperlessBilling	0.01
PhoneService	0.01
TechSupport	0.01
OnlineBackup	0.00
DeviceProtection	0.00
Partner	0.00
Dependents	0.00
SeniorCitizen	0.00
StreamingTV	0.00

Appendix E:

Random Forest results

Feature Importances	
MonthlyCharges	0.23798
tenure	0.22618
Contract	0.09528
PaymentMethod	0.06076
OnlineSecurity	0.04608
TechSupport	0.04335
gender	0.03412
InternetService	0.02961
PaperlessBilling	0.02825
OnlineBackup	0.02758
Partner	0.02697
MultipleLines	0.02613
DeviceProtection	0.02452
SeniorCitizen	0.02441
Dependents	0.02303
StreamingMovies	0.02034
StreamingTV	0.01953
PhoneService	0.00588

Optimized Random forest results

Feature Importances	
Contract	0.31086
tenure	0.1711
OnlineSecurity	0.12261
TechSupport	0.10498
MonthlyCharges	0.08253
InternetService	0.06885
OnlineBackup	0.02748
PaymentMethod	0.02741
DeviceProtection	0.01555
PaperlessBilling	0.01191
StreamingMovies	0.01045
StreamingTV	0.00878
MultipleLines	0.00817
Dependents	0.00726
Partner	0.00697
gender	0.00623
SeniorCitizen	0.00547
PhoneService	0.00341