Customer Survival Analysis and Churn Prediction

App: https://churn-prediction-app.herokuapp.com/

About

In this project, I have utilized survival analysis models to see how the likelihood of the customer churn changes over time and to calculate customer LTV. I have also implemented the Random Forest model to predict if a customer is going to churn and deployed a model using the flask web app.

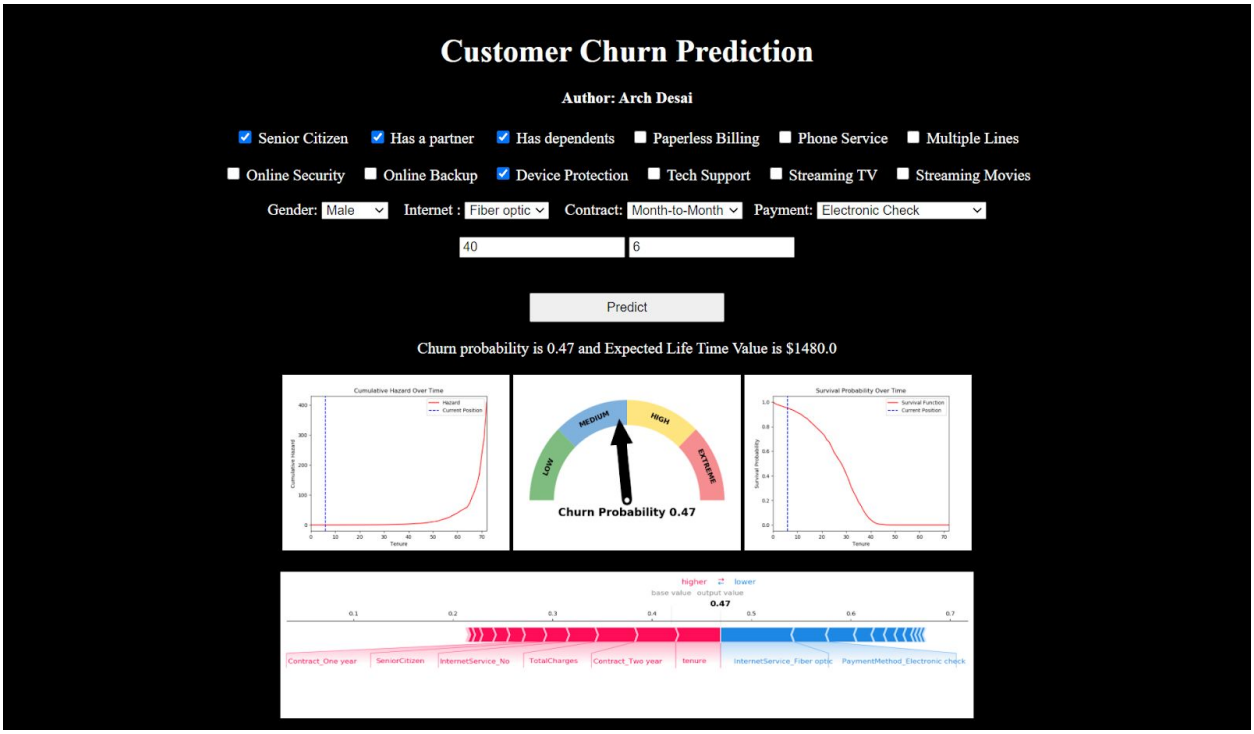
Customer attrition, also known as customer churn, customer turnover, or customer defection, is the loss of clients or customers.

Telephone service companies, Internet service providers, pay-TV companies, insurance firms, and alarm monitoring services, often use customer attrition analysis and customer attrition rates as one of their key business metrics because the cost of retaining an existing customer is far less than acquiring a new one. Companies from these sectors often have customer service branches that attempt to win back defecting clients, because recovered long-term customers can be worth much more to a company than newly recruited clients.

Predictive analytics use churn prediction models that predict customer churn by assessing their propensity of risk to churn. Since these models generate a small prioritized list of potential defectors, they are effective at focusing customer retention marketing programs on the subset of the customer base who are most vulnerable to churn.

In this project, I aim to perform customer survival analysis and build a model that can predict customer churn. I also aim to build an app that can be used to understand why a specific customer would stop the service and to know his/her expected lifetime value.

Final Customer Churn Prediction App



Project Organization

```
.
├── Images/                                : contains images
├── static/: plots to show gauge chart, hazard and survival curve, shap values
in Flask App
│   └── images/
│       ├── hazard.png
│       ├── surv.png
│       ├── shap.png
│       └── new_plot.png
├── templates/                             : contains html template for flask app
│   └── index.html
├── Customer Survival Analysis.ipynb: Survival Analysis Kaplan-Meier curve,
log-rank test, and Cox-proportional Hazard model
├── Exploratory Data Analysis.ipynb: Data Analysis to understand customer data
```

```
├── Churn Prediction Model.ipynb: Random Forest model to predict customer churn
├── app.py: Flask App
├── app-pic.png                          : Final App image
├── explainer.bz2: Shap Explainer
├── model. pkl: Random Forest model
├── survivemodel.pkl: Cox-proportional Hazard model
├── requirements.txt: requirements to run this model
├── Procfile: procfile for app deployment
├── LICENSE.md: MIT License
└── README.md: Report
```

Customer Survival Analysis

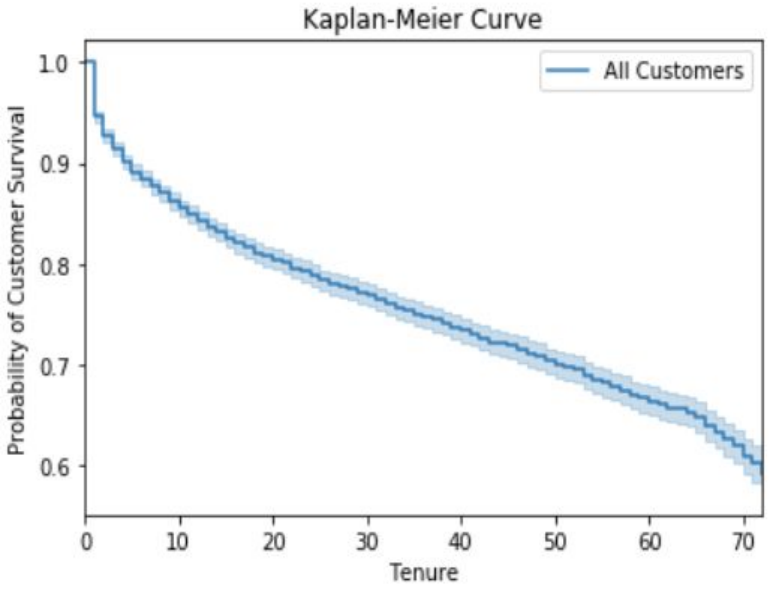
Survival Analysis: Survival analysis is generally defined as a set of methods for analyzing data where the outcome variable is the time until the occurrence of an event of interest. The event can be death, the occurrence of a disease, marriage, divorce, etc. The time to event or survival time can be measured in days, weeks, years, etc.

For example, if the event of interest is a heart attack, then the survival time can be the time in years until a person develops a heart attack.

Objective: The objective of this analysis is to utilize non-parametric and semi-parametric methods of survival analysis to answer the following questions.

- How the likelihood of customer churn changes over time?
- How we can model the relationship between customer churn, time, and other customer characteristics?
- What are the significant factors that drive customer churn?
- What is the survival and Hazard curve of a specific customer?
- What is the expected lifetime value of a customer?

Kaplan-Meier Survival Curve:

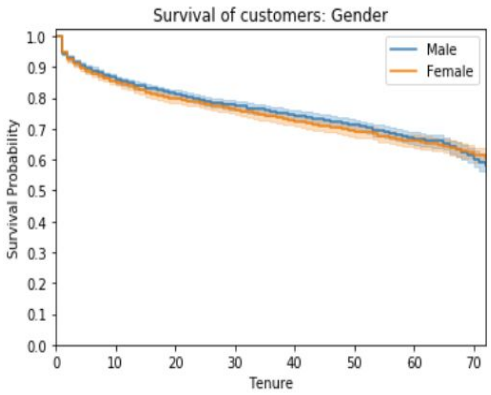


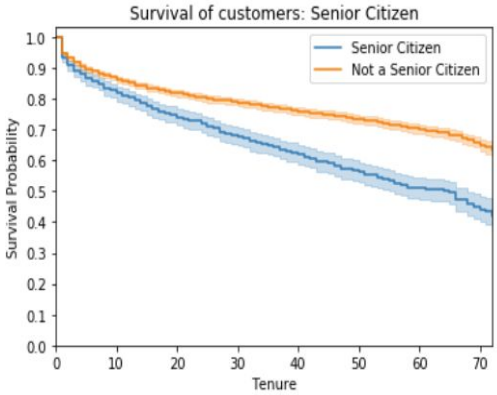
From the above graph, we can say that

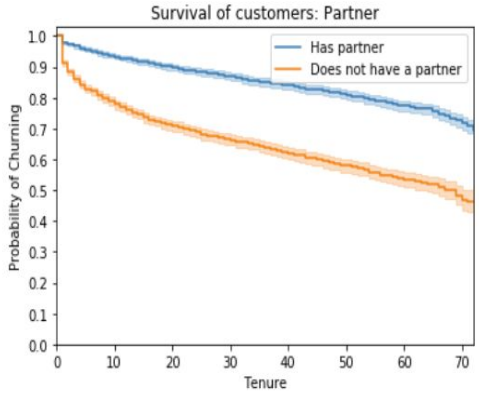
- AS expected, for telecom, churn is relatively low. The company was able to retain more than 60% of its customers even after 72 months.
- There is a constant decrease in survival probability between 3-60 months.
- After 60 months or 5 years, survival probability decreases with a higher rate.

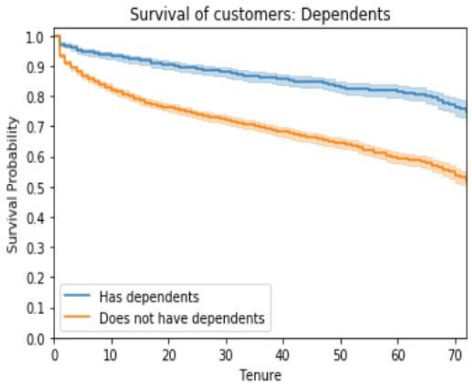
Log-Rank Test:

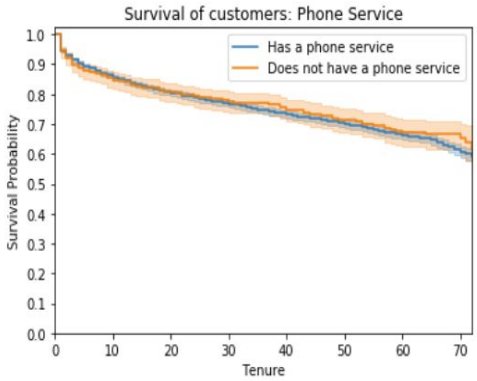
The log-rank test is carried out to analyze churning probabilities group-wise and to find if there is statistical significance between groups. The plots show the survival curve group wise.

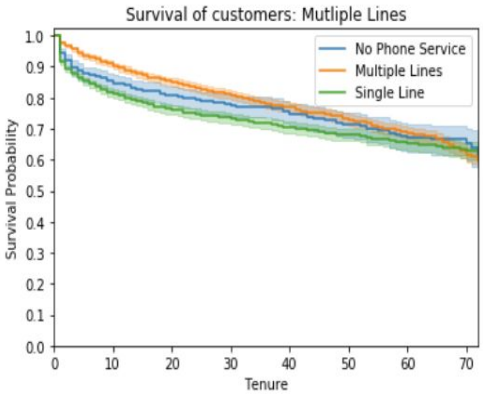

Survival of customers: Gender
Male
Female
Survival Probability
Tenure

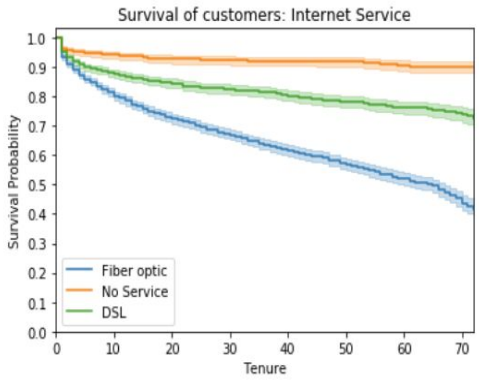

Survival of customers: Senior Citizen
Senior Citizen
Not a Senior Citizen
Survival Probability
Tenure

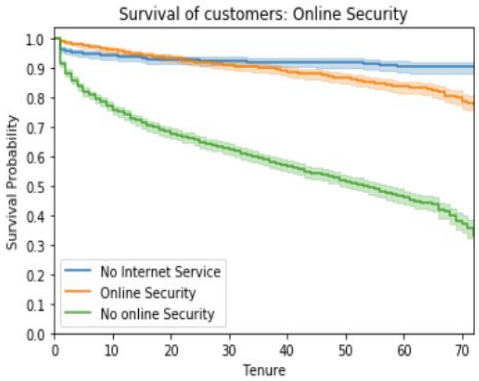

Survival of customers: Partner
Has partner
Does not have a partner
Probability of Churning
Tenure

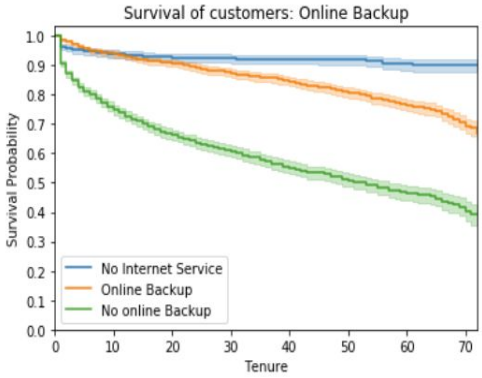

Survival of customers: Dependents
Has dependents
Does not have dependents
Survival Probability
Tenure

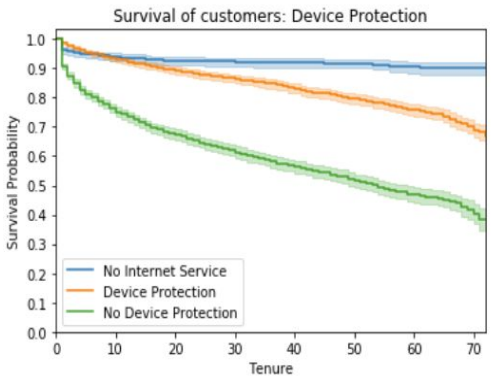

Survival of customers: Phone Service
Has a phone service
Does not have a phone service
Survival Probability
Tenure

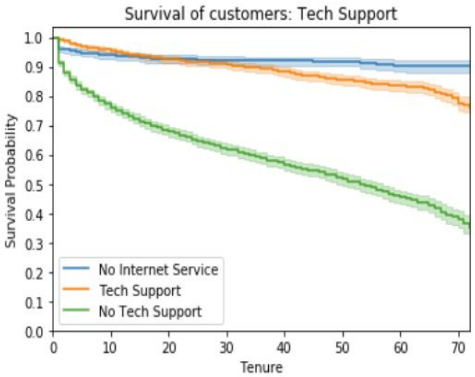

Survival of customers: Mutliple Lines
No Phone Service
Multiple Lines
Single Line
Survival Probability
Tenure

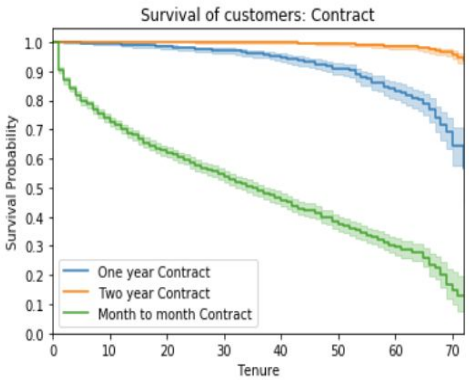

Survival of customers: Internet Service
Survival Probability
Tenure
Fiber optic
No Service
DSL

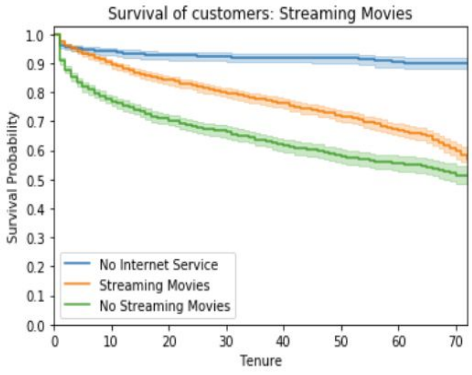

Survival of customers: Online Security
Survival Probability
Tenure
No Internet Service
Online Security
No online Security

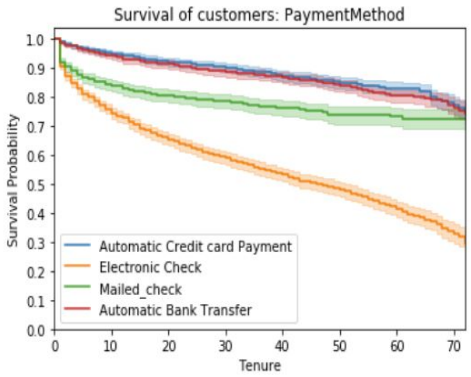

Survival of customers: Online Backup
Survival Probability
Tenure
No Internet Service
Online Backup
No online Backup

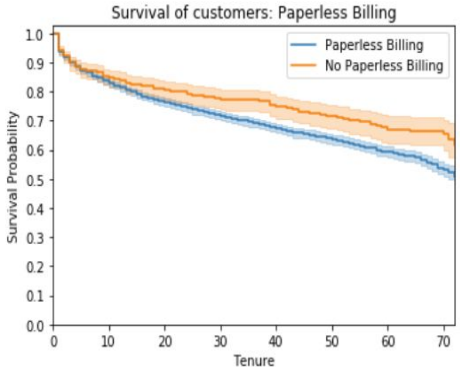

Survival of customers: Device Protection
Survival Probability
Tenure
No Internet Service
Device Protection
No Device Protection


Survival of customers: Tech Support
Survival Probability
Tenure
No Internet Service
Tech Support
No Tech Support


Survival of customers: Contract
Survival Probability
Tenure
One year Contract
Two year Contract
Month to month Contract



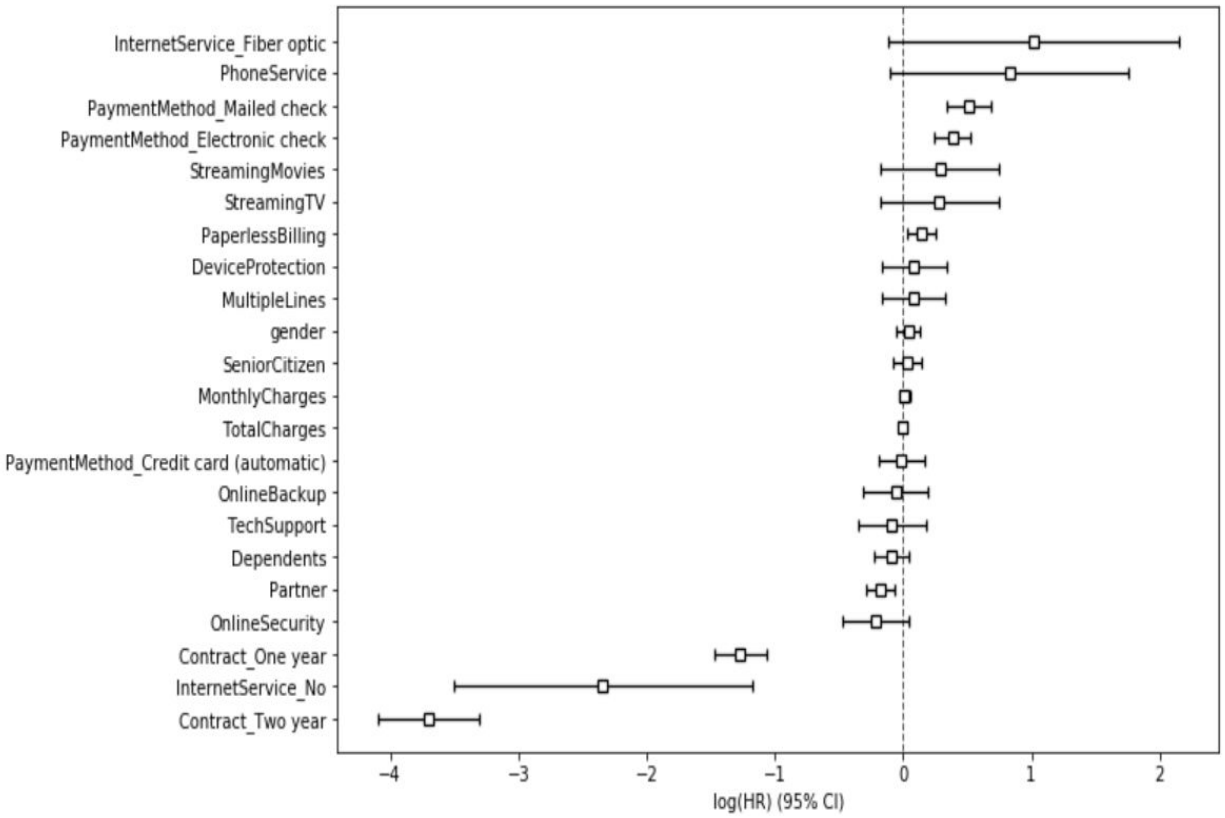




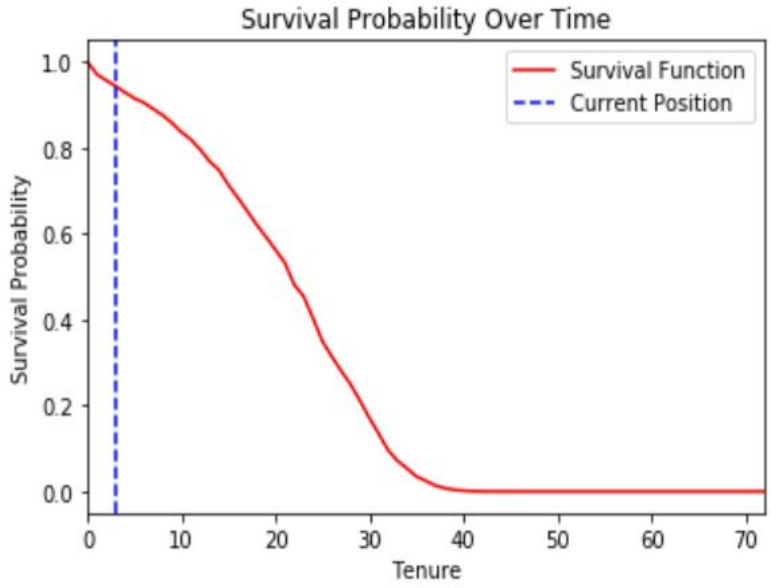
From the above graphs we can conclude the following:

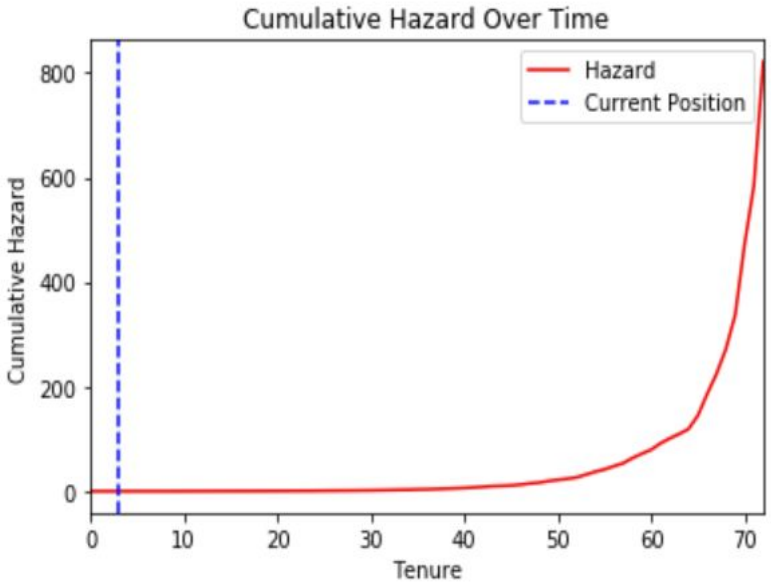
- Customer's Gender and the phone service type are not indicative features and their p-value of the log-rank test is above threshold value 0.05.
- If the customer is young and has a family, he or she is less likely to churn. The reason might be a busy life, more money, or another factor.
- If the customer is not enrolled in services like online backup, online security, device protection, tech support, streaming Tv, and streaming movies even though having active internet service, the survival probability is less.
- The company should target customers who opt for internet service as their survival probability constantly decreases. Also, Fiber Optic type of Internet Service is costly and fast compared to DSL and this might be the reason for higher customer churn.
- More offers should be given to customers who opt for month-to-month contracts and the company should target customers to subscribe for long-term service.
- If the customer's paying method is automatic, he or she is less likely to churn. The reason is in the case of electronic checks and mailed checks, a customer has to make an effort to pay and it takes time.

Survival Regression: I use the cox-proportional hazard model to perform survival regression analysis on customer data. This model is used to relate several risk factors or exposures simultaneously to survival time. In a Cox proportional hazards regression model, the measure of effect is the hazard rate, which is the risk or probability of suffering the event of interest given that the participant has survived up to a specific time. The model fits the data well and the coefficients are shown below.



Using this model we can calculate the survival curve and hazard curve of any customer as shown below. These plots are useful to know the remaining life of a customer.





Customer Lifetime Value:

To calculate customer lifetime value, I would multiply the Monthly charges the customer is paying to Telcom and the expected lifetime of the customer.

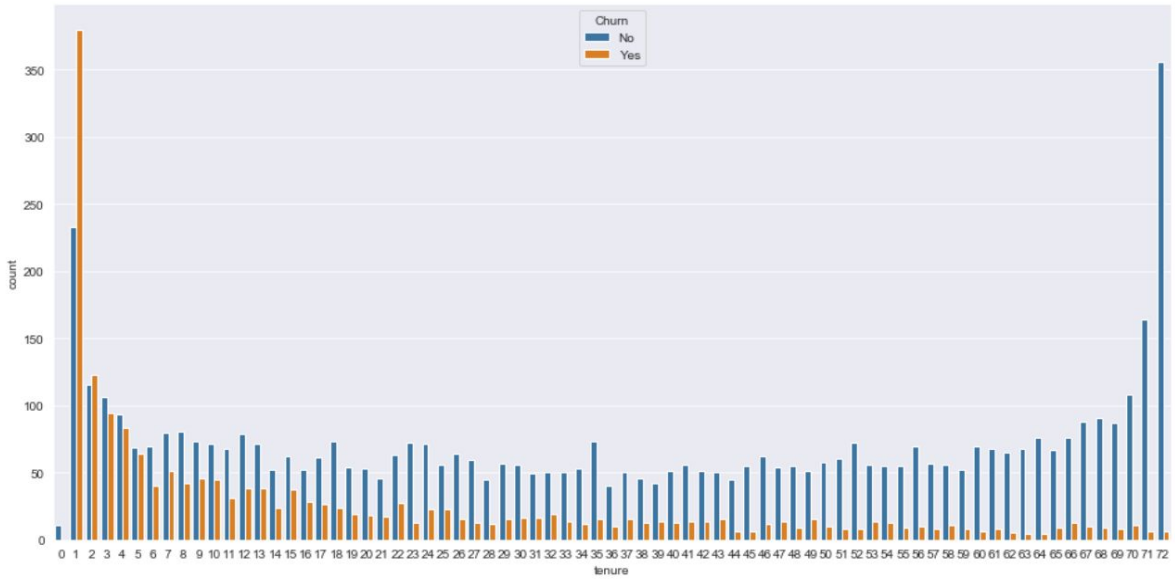
I utilize the survival function of a customer to calculate its expected lifetime. I would like to be a little bit conservative and consider the customer is churned when the survival probability of him is 10%.

Customer Churn Prediction

I aim to implement a machine learning model to accurately predict if the customer will churn or not.

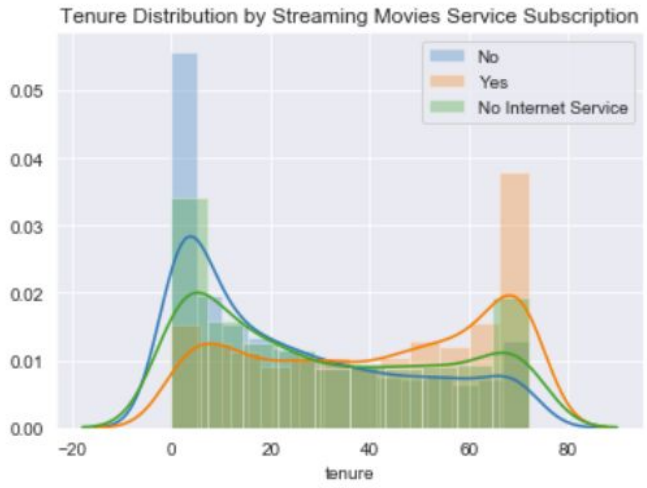
Analysis

Churn and Tenure Relationship:



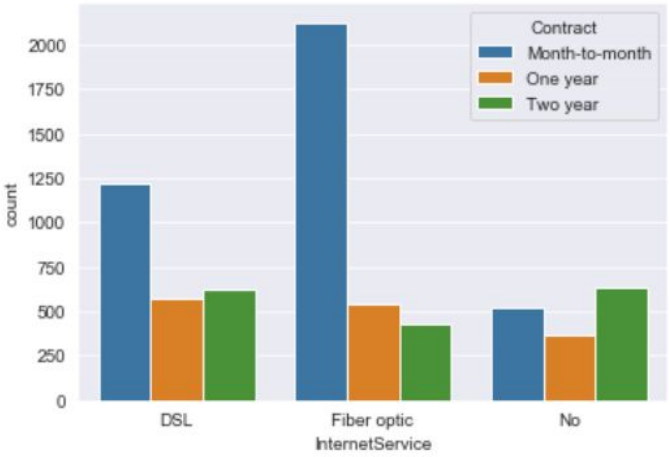
- As we can see the higher the tenure, the lesser the churn rate. This tells us that the customer becomes loyal with the tenure.

Tenure Distribution by Various Services:



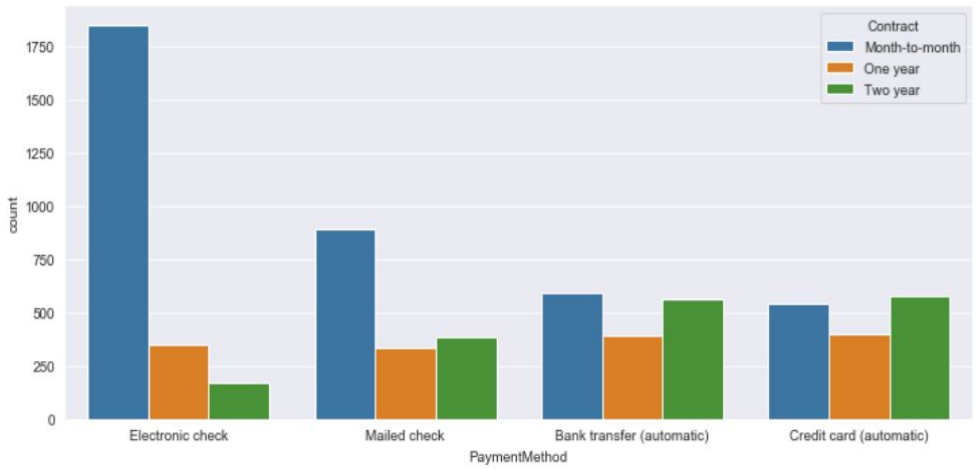
- When the customers are new they do not opt for various services and their churning rate is very high. This can be seen in the above plot for Streaming Movies and this holds true for all various services.

Internet Service By Contract Type:



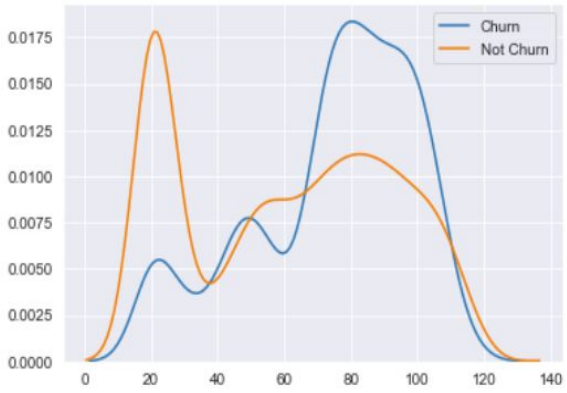
- Many of the people who opt for month-to-month Contracts choose Fiber optic as an Internet service and this is the reason for the higher churn rate for fiber optic Internet service type.

Payment method By Contract Type:



- People having a month-to-month contract prefer paying by Electronic Check mostly or mailed check. The reason might be a short subscription cancellation process compared to automatic payment.

Monthly Charges:

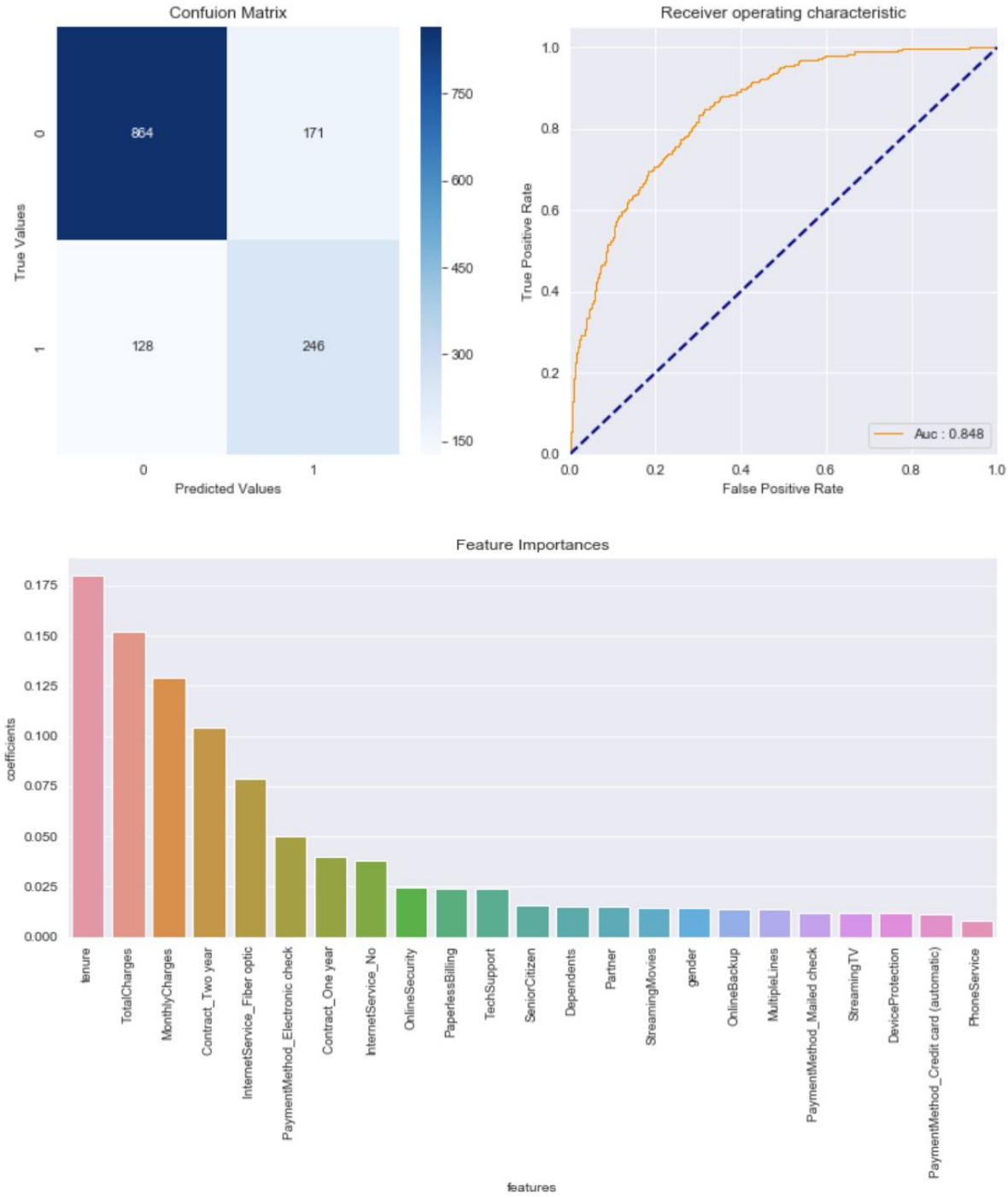


- As we can see the customers paying high monthly fees to churn more.

Modeling

For the modeling, I will use tress based Ensemble method as we do not have linearity in this classification problem. Also, we have a class imbalance of 1:3 and to combat it I will assign a class weightage of 1:3 which means false negatives are 3 times costlier than false positives. I built a model on 80% of data and validated the model on the remaining 20% of data keeping in mind that I do not have data leakage. The random forest model has many hyperparameters and I tuned them using Grid Search Cross-Validation while making sure that I do not overfit.

The final model resulted in a 0.62 F1 scores and 0.85 ROC-AUC. The resulting plots can be seen below.



From the feature importance plot, we can see which features govern the customer churn.

Explainability

We can explain and understand the Random forest model using explainable AI modules such as Permutation Importance, Partial Dependence plots, and Shap values.

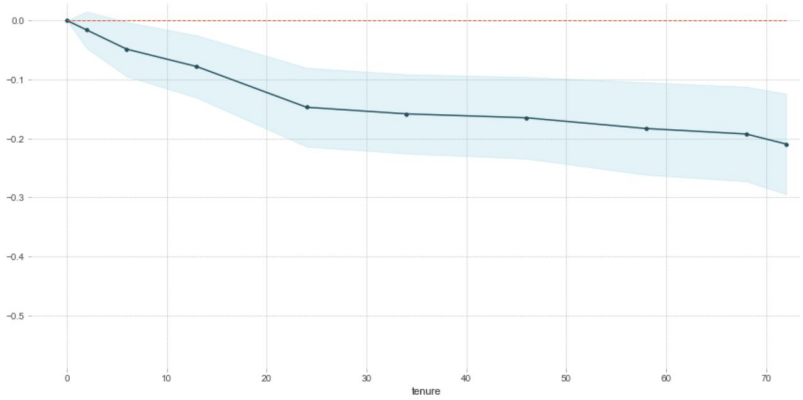
1. Permutation Importance shows feature importance by randomly shuffling feature values and measuring how much it degrades our performance.

Weight	Feature
0.0185 ± 0.0058	InternetService_Fiber optic
0.0064 ± 0.0088	Contract_Two year
0.0045 ± 0.0058	OnlineSecurity
0.0041 ± 0.0134	Contract_One year
0.0038 ± 0.0086	PaymentMethod_Electronic check
0.0037 ± 0.0071	InternetService_No
0.0028 ± 0.0094	tenure
0.0026 ± 0.0011	OnlineBackup
0.0020 ± 0.0078	MonthlyCharges
0.0010 ± 0.0014	DeviceProtection
0.0009 ± 0.0083	PaperlessBilling
0.0007 ± 0.0030	TechSupport
0.0004 ± 0.0032	StreamingMovies
0.0003 ± 0.0017	gender
0.0001 ± 0.0019	PhoneService
-0.0000 ± 0.0009	MultipleLines
-0.0001 ± 0.0006	StreamingTV
-0.0004 ± 0.0044	SeniorCitizen
-0.0009 ± 0.0033	Dependents
-0.0020 ± 0.0026	PaymentMethod_Credit card (automatic)
-0.0040 ± 0.0064	TotalCharges
-0.0040 ± 0.0039	Partner
-0.0075 ± 0.0033	PaymentMethod_Mailed check

2. Partial dependence plot is used to see how churning probability changes across the range of a particular feature. For example, in the below graph of the tenure group, the churn probability decreases at a higher rate if a person is in tenure group 2 compared to 1.

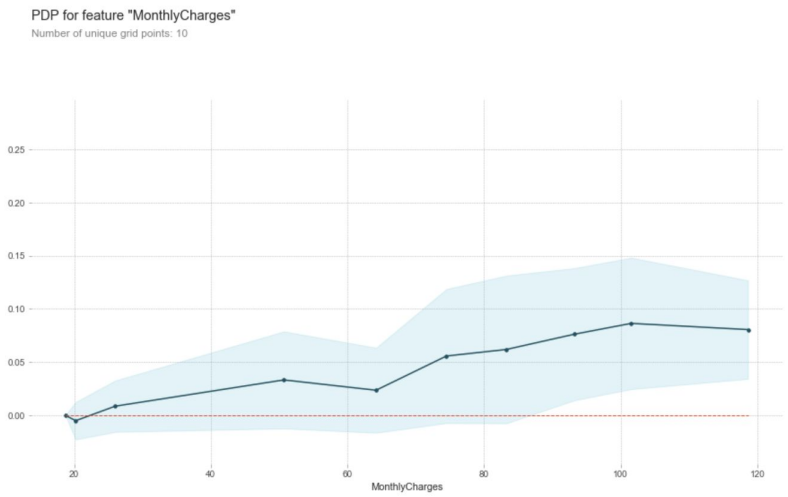
PDP for feature "tenure"

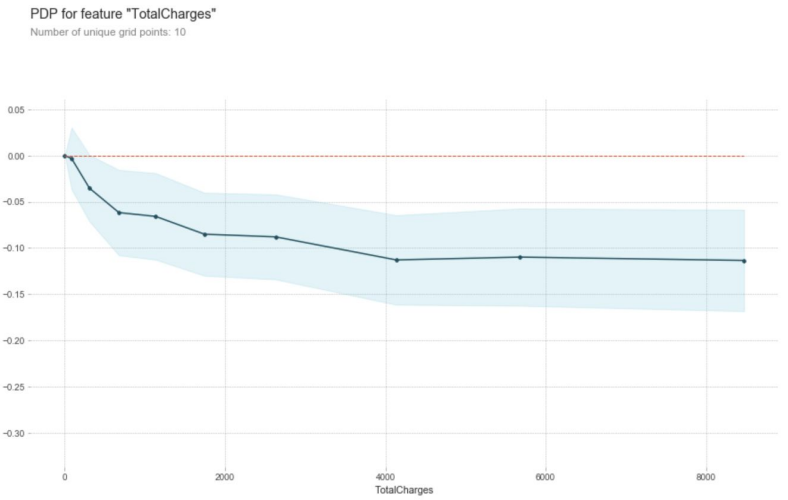
Number of unique grid points: 10



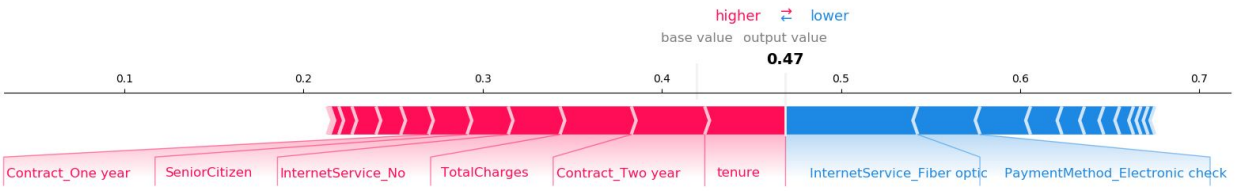
PDP for feature "Contract_Two year"

Number of unique grid points: 2





3. Shap values (SHapley Additive exPlanations) is a game-theoretic approach to explain the output of any machine learning model. In the below plot, we can see why a particular customer's churning probability is less than the baseline value and which features are causing them.



Flask App

I saved the final tuned Random Forest model and deployed it using the Flask web app. Flask is a micro web framework written in Python. It is designed to make getting started quick and easy, with the ability to scale up to complex applications. I saved the shap value explainer tuned using random forest model to show shap plots in-app. I have also utilized the cox-proportional hazard model to show the survival curve and hazard curve and to calculate the expected customer lifetime value.

The final app shows churning probability, a gauge chart of how severe a customer is, and shap values based on customer's data. The final app layout can be seen above.