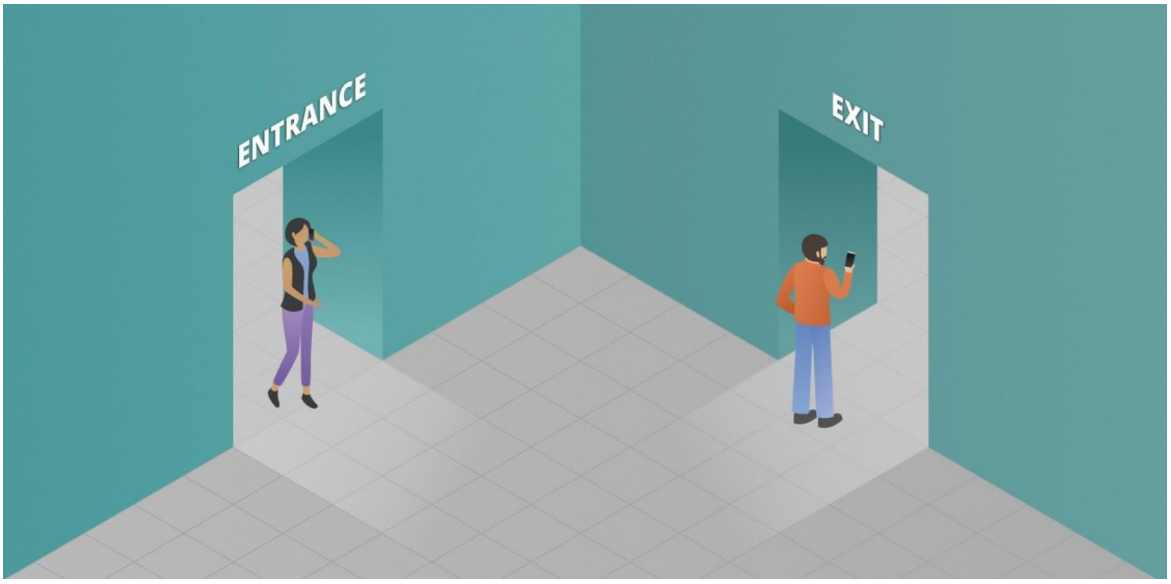
Understanding Customer Churn Using Survival Analysis



STATISTICAL MODELS AND COMPUTING

INSTRUCTOR: JACK MARDEKIAN

FINAL PROJECT

Manisha Gayatri Damera - *md1723*

Neha Thonta - *nt446*

Vishal Reddy Mekala - *vm574*

Vivek Reddy Chittari - *vc508*

Contents

Abstract

'Survival Analysis' or 'Time-to-event Analysis' are widely used techniques in medicine and biological research. It can also be used in the marketing industry to understand customer churning pattern. This study aims to implement survival analysis methods for understanding customer churn in the telecommunication industry. Multiple studies have revealed that the cost of acquiring new consumers outweighs that of keeping current ones. Therefore, it is essential for companies to retain their customers for the longest to maximise customer lifetime-value. In this study, we're going to perform descriptive analytics to predict the most influential variables in customer churning. The inferences made can be utilized to aid strategic decision-making and business-planning.

1 Introduction

Survival Analysis, as the name implies, uses statistics to estimate how long it will be before a specific event, such as the failure of a mechanical component or a consumer canceling their subscription, takes place. Although conventional regression techniques successfully predict customer churn, they fail to give information about how long they would stay with the company. We don't want to discard this information as it's valuable.

The main objective of this project is exploring different survival analysis methods and performing telecommunication customer churn analysis. The report summarizes the methodologies used and its results. We combined the traditional survival analysis methods with the advanced survival analysis using machine learning techniques. In this study, we investigated the non-parametric, parametric, and semi-parametric methods to delineate the survival analysis algorithms and procedures. Different features in the data which lead to churn are interpreted. The work presents experimental approaches to enumerate the significance of survival analysis.

2 Terminology in Survival Analysis

The survival time is the period between the start and end of an event. The survival time in this experiment is defined as the difference (in months) between the subscriber churn and the subscriber entry time point. The survival time of users in this paper is discrete, i.e., the survival time is a finite period, so the range of survival time is set to T= 0,1,2..., Tmax, where Tmax is the maximum.

We have survival time(T) and censoring time (C). And the event of interest is Y = min(T, C). If the event occurs before censoring (i.e. T < C) then we observe the true survival time T; however, if censoring occurs before the event (T > C) then we observe the censoring time.

2.1 Survival Function

The survival function S(t) is the probability that the duration will be more than time 't', or, the probability of an individual surviving past time 't'.

$$S(t) = 1 - F(t) = \Pr[T \geq t]$$

2.2 Hazard Function

Hazard function H(t) expresses the conditional probability that the event will occur within [t,t+dt], given that it has not occurred before. In other words, it is the risk for death at that moment.

3. Data Description and Exploratory Data Analysis

3.1 Data Description

The data used in this project is publicly available online on multiple platforms. We obtained it from Kaggle and studied using Python. The dataset contains information about customers subscribed to a Telecom Service Provider.

Each row depicts a single customer uniquely identified by a customerID. The variables of interest to us is ’tenure’ and ’Churn’. ’Tenure’ is a continuous variable and essentially specifies how long (in months) the customer has stayed with the company.

’Churn’ is a categorical variable with two values where: Churn = ’Yes’ - the customer has canceled subscription with the company Churn = ’No’ - the customer has been censored (or) lost track of the customer.
Number of customers - 7043
Number of features - 21

3.2 Exploratory Data Analysis

The relationship between tenure (the time for the event) and churn (the event of our interest) is interpreted in the below plots.

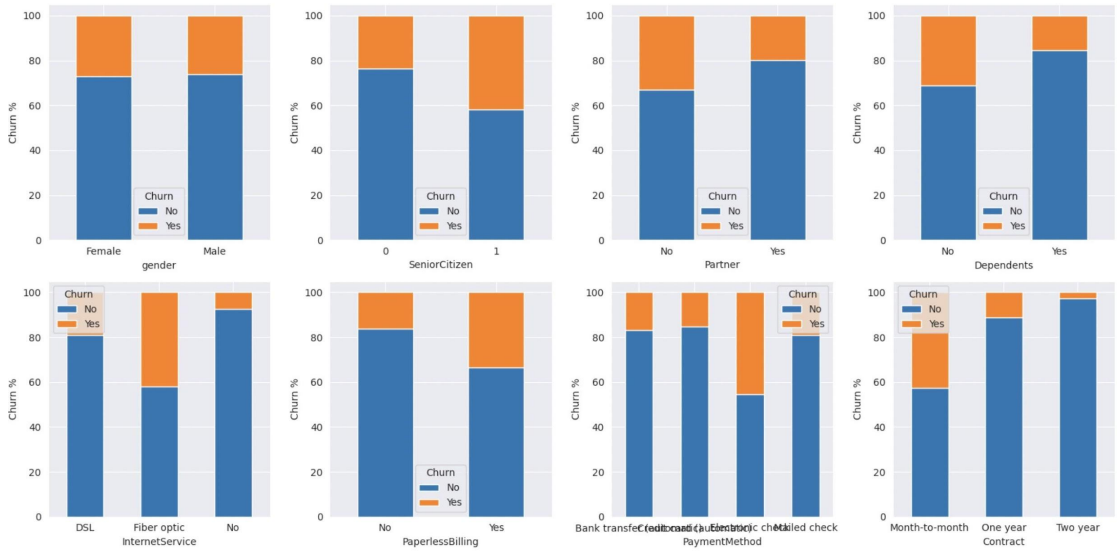


Figure 1. Customer churn analysis demographics data - the gender, senior citizen, partner and dependents attributes, Fiber Optics, Paperless Billing, Payment Method, Contrac

Figure 1: While the demographic information doesn’t clearly convey the Churn % among customers, customers who opted for the company’s FiberOptic Internet Service and Electronic Check Payment Method seemed to churn more than customers who opted for other options in that category.

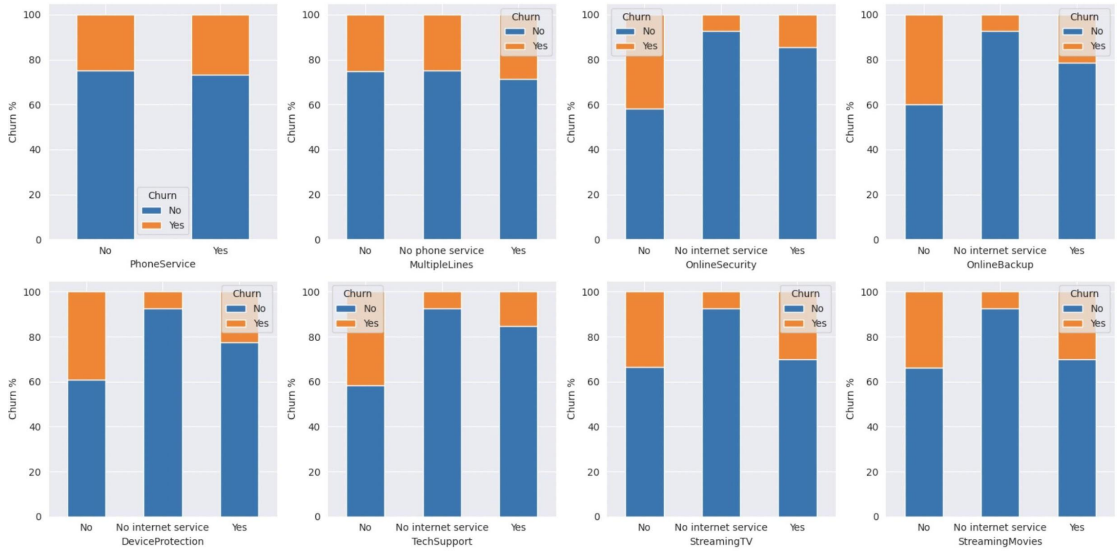


Figure 2. Telecommunication services affecting the customer churn:churn percentage vs type of service

Figure 2: A noticeable pattern from the plots in Figure 2. is that customers who have not opted for the company's internet service have churned less than the ones who opted for it. This leads us to more questions - How good is the company's internet service?

Figure 3: It's pretty evident from Figure 3 that customers are the most likely to churn in the first year. The company can probably attract customers with lucrative deals in the first year, which can help retain them past the first year. Next, we'll look at how each variable plays a role in a Customer's Churn with the help of Survival techniques.

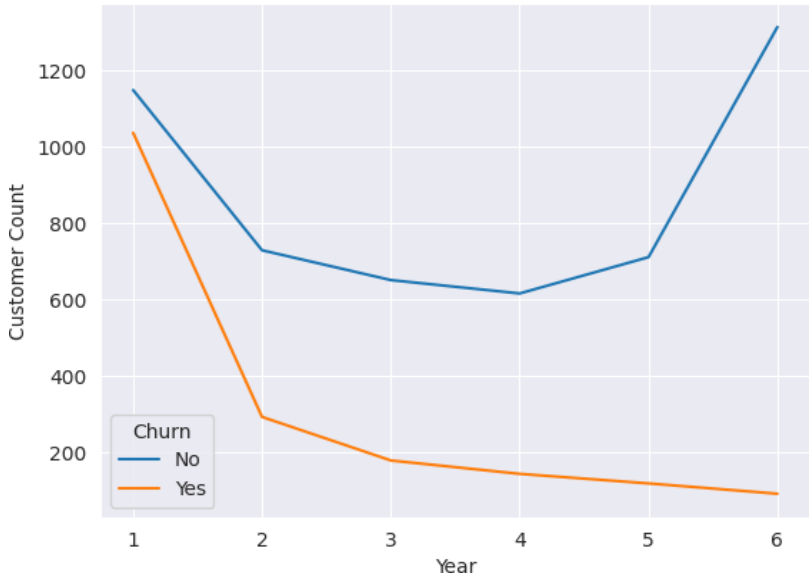


Figure 3. Churn by year

4 Methodology

4.1 Non-Parametric Methods

4.1.1 The Kaplan Meier Curve

The Kaplan-Meier estimate is the simplest way of computing survival over time in spite of all these difficulties associated with subjects or situations. This is the non-parametric method to calculate the survival probability.

$$S(t) = \prod_{i=1}^{t-1}(1 - \frac{d_i}{n_i})$$

The plots below show Kaplan Meier Curves for the customer data stratified by all variables in the dataset. Although it's noticeable for some variables that there is a difference in the levels, we will confirm this later by using log-rank tests anyway.

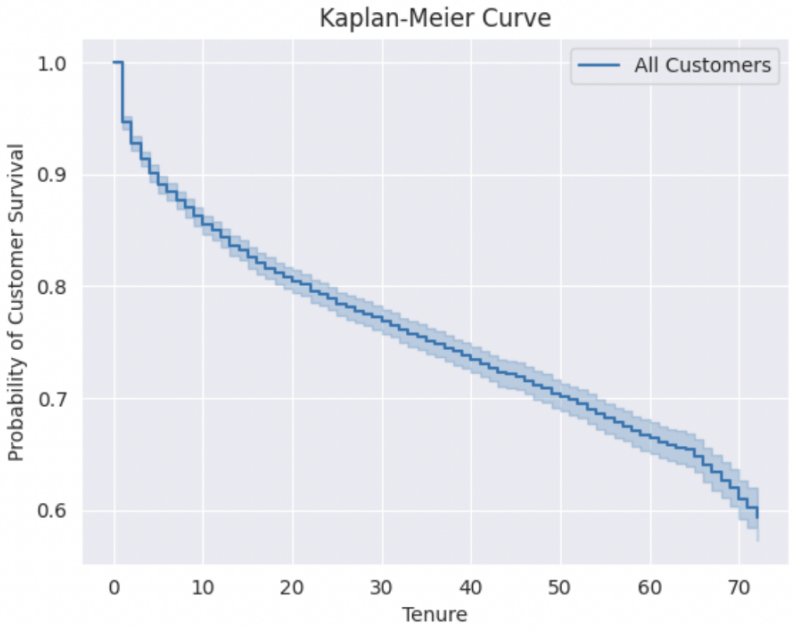


Figure 4. Kaplan Meier Curve for all variables

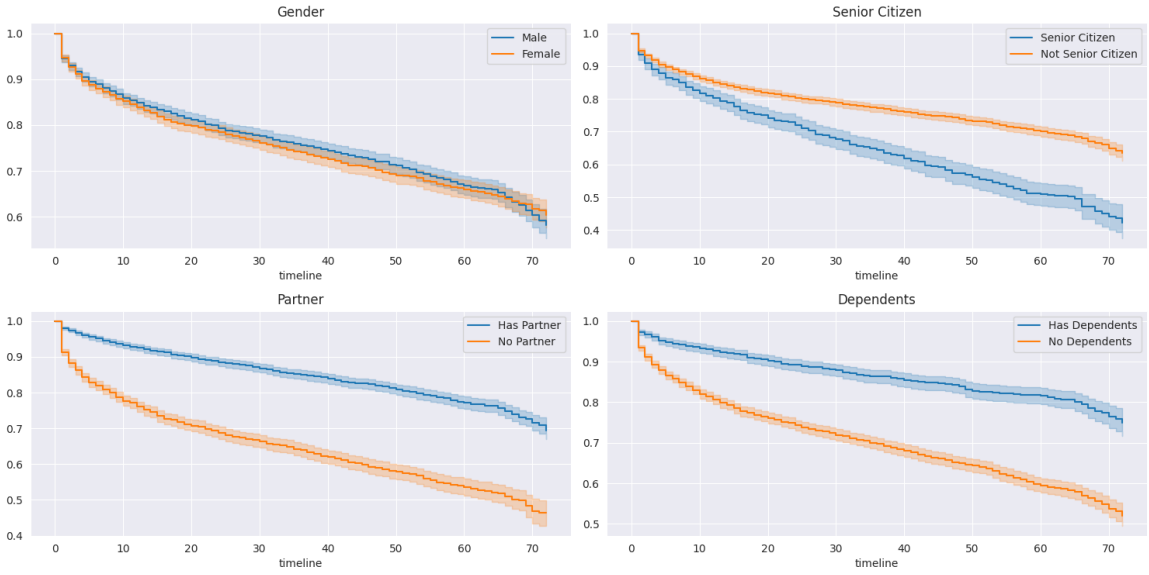


Figure 5. Kaplan Meier Curves for Gender, Senior Citizen, Partner, Dependents

Clearly there is a distinction in survival probabilities between different internet services and the paperless billing options from the above plot

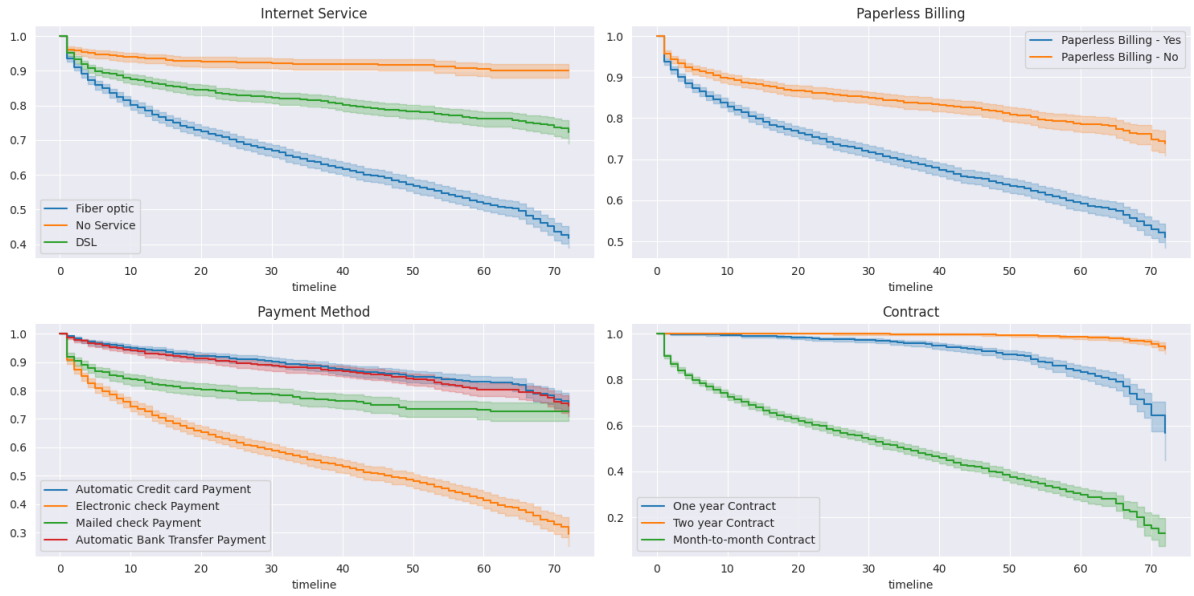


Figure 6. Kaplan Meier Curves for Dependents, Internet Service, Paperless Billing, Payment Method and Contract

Visually, Gender doesn't seem to impact survival probabilities. On the other hand, customer's relationship status and dependents seem to have a difference in churning rates. We can confirm this later using log-rank tests.

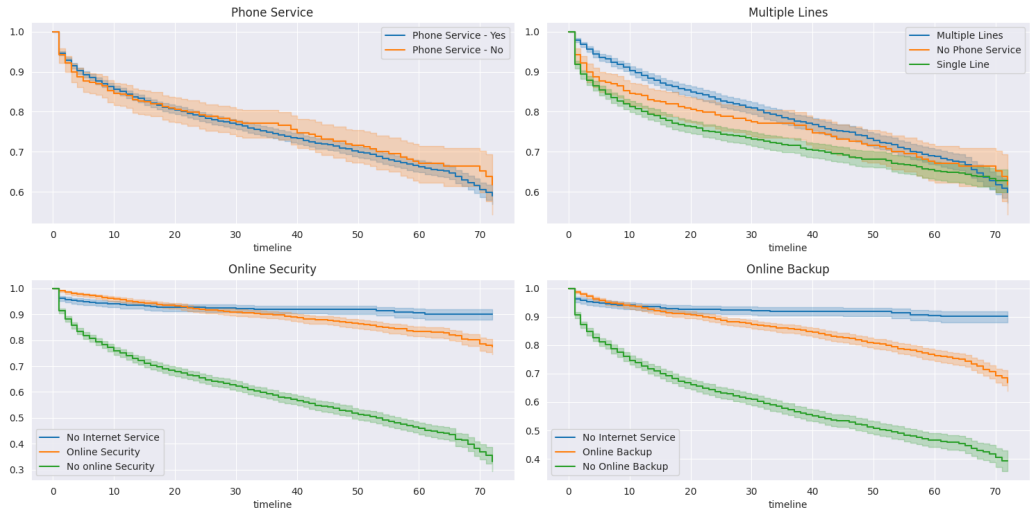


Figure 7. Kaplan Meier Curves for Phone Service, Multiple Lines, Online Security, Online Backup

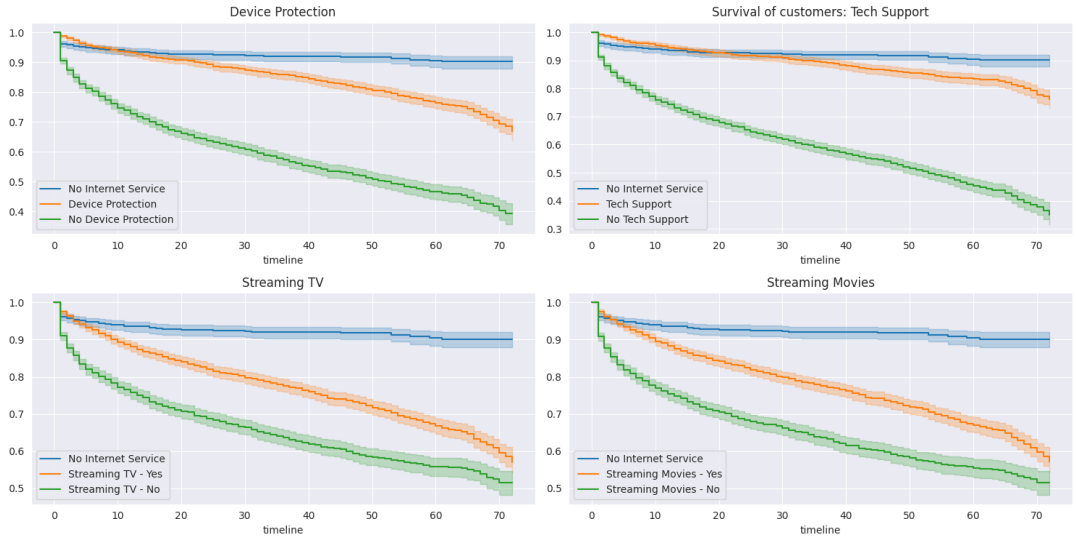


Figure 8. Kaplan Meier Curves for Device Protection, Tech Support, Stream- ing TV and Streaming Movies

Again, from the above plots some variables seem to have a difference in their survival probabilities, whereas some do not. Log-rank tests are more accurate in proving the significance of these variables.

4.12 The Log-rank Test

In the nonparametric log-rank test, which in statistical analysis is used for comparing survival distributions, higher values of the log-rank statistic mean greater dissimilarity between distributions and those utilizing likelihood-based measures.

Class	test_statistic	p-value
Gender	0.53	0.47
Senior Citizen	109.49	<0.005
Partner	423.54	<0.005
Dependents	232.70	<0.005
Phone Service	0.43	0.51
Mutiple Lines	30.97	<0.005
Internet Service	520.12	<0.005
Device Protection	1013.86	<0.005
Online Security	821.34	<0.005
Online Backup	1763.51	<0.005
Tech Support	989.56	<0.005
Streaming TV	368.31	<0.005
Streaming Movies	378.43	<0.005
Contract	2352.87	<0.005
Payment Method	865.24	<0.005
Paperless Billing	189.51	<0.005

Table 1. Summary of Log-Rank Tests for all variables

We are essentially testing the following hypothesis:

H_0 : There is no difference in the levels of each variable

H_a : There is significant difference between the levels of the variable

A p-value less than 0.05 confirms that there is evidence to support the claim that there is significant difference between the levels of the variables.

Except for the variables highlighted in the table, the rest are all significant at the level '0.05'.

For example,

Gender: Visually inspecting the Kaplan Meier curve stratified by Gender, it's quite noticeable that there is not much difference in survival probabilities between the two genders. The log-rank test confirms this result as the p-value is 0.47, indicating the difference in survival probabilities is not statistically significant between the genders.

Partner: The Kaplan Meier plot stratified by whether the customer has a Partner or not, suggests there is a difference in survival probabilities based on the customer's relationship status. The log- rank test gives a statistically significant p-value, indicating evidence of a difference in survival probabilities based on a customer's relationship status.

Phone Service: The p-value obtained from the log-rank tests above suggests that the difference in survival probabilities between customers who have opted for the company's Phone Service feature or not, is not statistically significant. Visually inspecting the above Kaplan Meier plot also confirms this.

4.2 Semi - Parametric Methods

4.2.1 Cox Proportional Hazards Model

The semi-parametric Cox proportional hazards model is widely used to see the effect of different features.

Hazard Ratio - Models the effect of explanatory covariates on risk. The model's strong proportionality assumption is supported by this parametric function, which states that at any given time, a subject's hazard functions will continue to be proportional to one another or to the baseline in the same ratio.

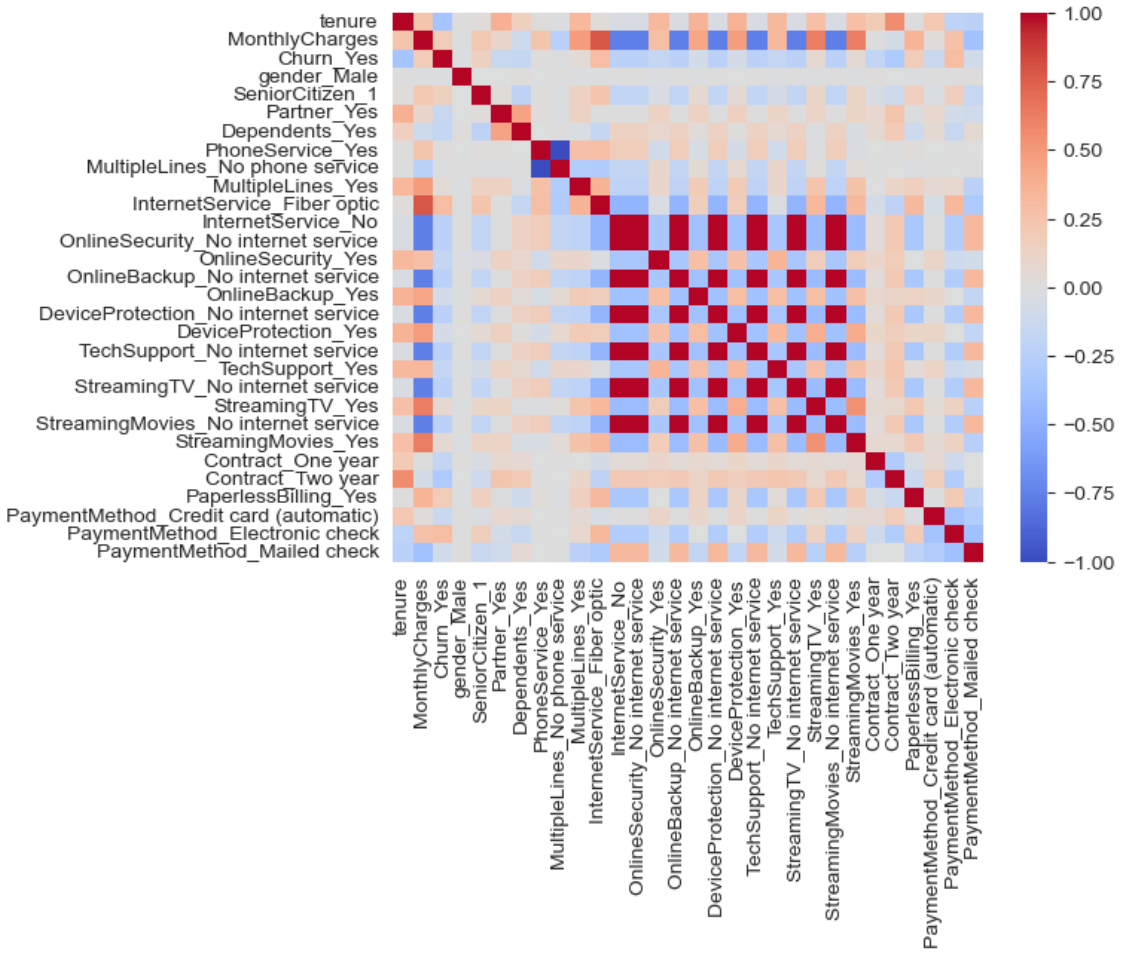


Figure 9. Correlation Heat Map

From the correlation plot above, it is quite evident that there's an extra factor for the variables associated with the company's internet services. This leads to a perfect correlation between some variables. For example, OnlineSecurity has three levels: 'Yes,' 'No,' and 'No Internet service.' The latter level is redundant and can be quoted as a 'No'. to reduce complexity. Similar changes can be done on other variables having redundant levels. The correlation plot after making these changes looks as given below. A significant reduction in multicollinearity.

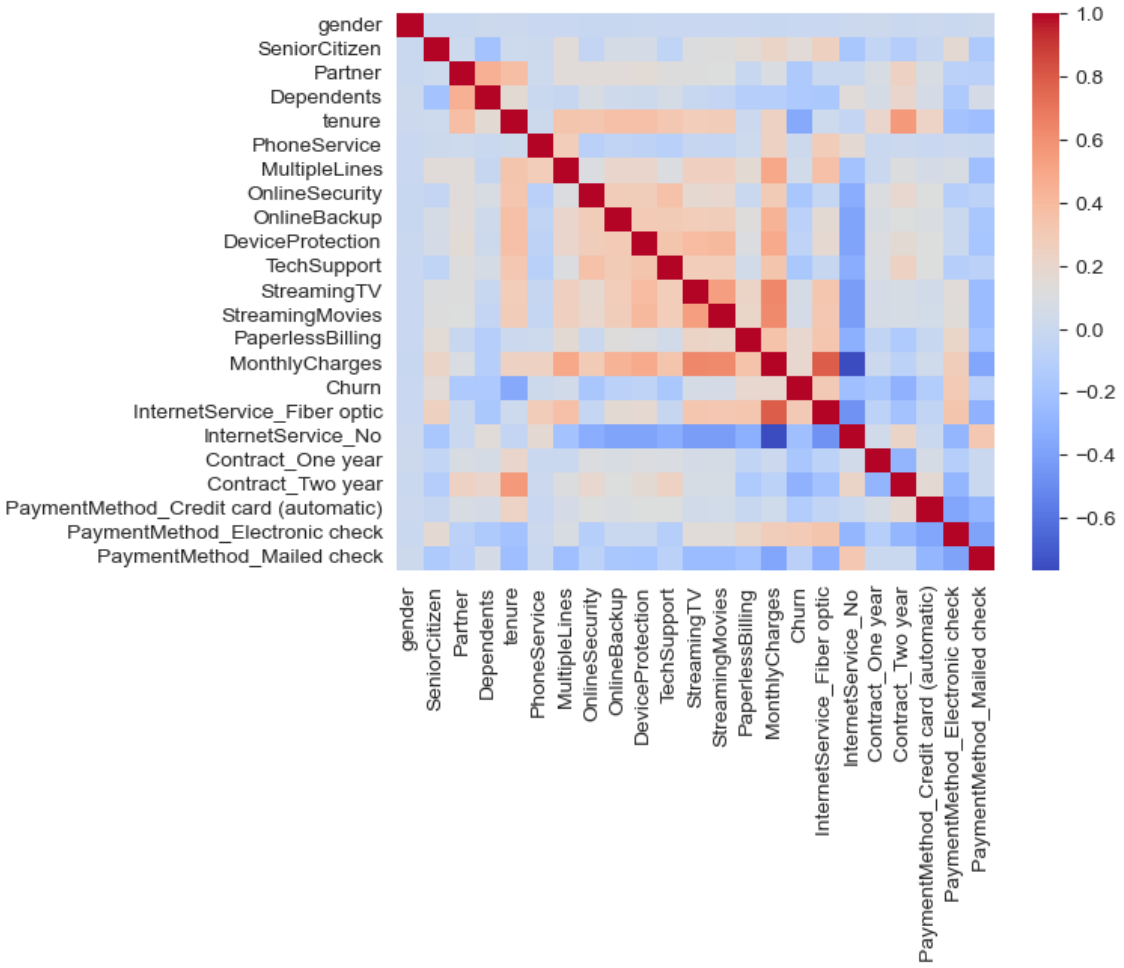


Figure 10. Heatmap with reduced multicollinearity

Concordance	0.87
Partial AIC	27811.24
log-likelihood ratio test	3538.06 on 21 df
-log2(p) of ll-ratio test	inf

Table 2. Summary of Cox proportional Model

Concordance Index: A concordance index of 0.87 indicates the model's ability to rank the survival times of customers accurately. The c-index is more interested in the order of the predictions than the predictions themselves.

For example, let A B have true Churn in months 3 and 8, respectively. If the model predicts that A B churn in months 3 and 8, respectively - the concordance index would be 1. Also, if the model predicts that A B would churn in months 15 and 16, respectively, the

con- cordance index would still be 1. As mentioned above, the c - index is only interested in the order of the predictions rather than the actual survival predictions.

Hazard Ratios: From the plot below, all variables having negative log(HR) indicate that the probability of the event(Churn in our case) happening for that variable is less than the baseline model. For example, customers subscribed to two-year contracts are exp(log(HR)) times more likely to Churn than customers on a monthly plan(baseline model). In other terms, customers on two- year contracts are 0.04 times more likely to Churn than customers on monthly plans.

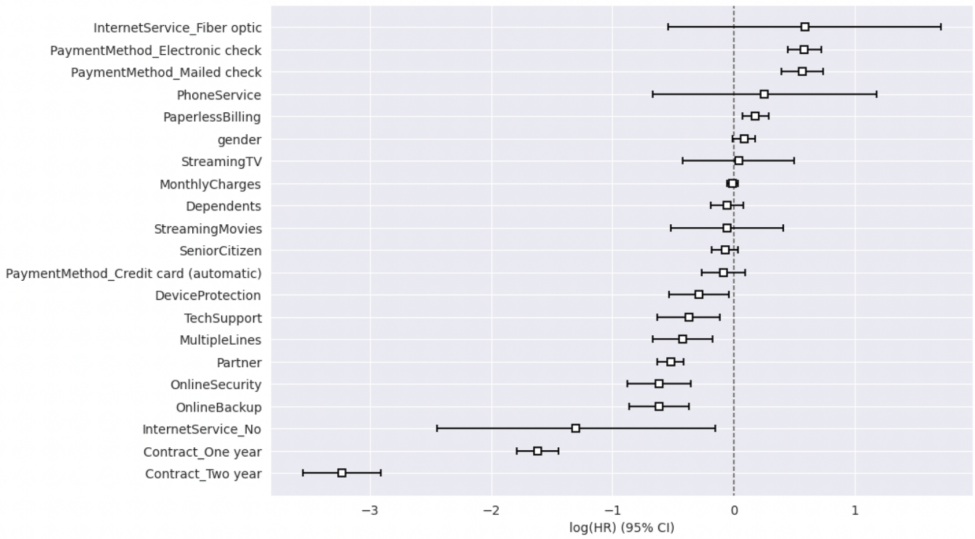


Figure 11.Variable coefficients plot

Similarly, for Internet Service, the baseline model is InternetService = DSL. From the plot above, customers who have opted for the company's Fiber Optic are at risk of churning more than the ones subscribed to the company's fiber optic since the log(HR) is positive for this case. Customers who have not opted for any internet service are at a lesser risk of churning. This confirms the inference made in EDA that the company's internet service is probably not up to the customer's expectation.

4.2.2 Penalized Cox Models

Penalized Cox Model is another approach for the survival data with a large number of covariates. Though Cox Model is appealing and provides valuable insights, in the case of many features, it is not efficient because, internally, it tries to invert a matrix that becomes non-singular due to correlations among features. Hence alternative methods are explored, and Penalized Cox is the one that includes the regression models and computes LASSO and RIDGE on the data.

RIDGE

Ridge works by adding an L2 penalty to shrink the coefficients to zero.

The coefficient change for varying α is inspected. Figure.17 shows that if the penalty has a large weight, all the coefficients shrink at most to zero. As the penalty's weight is decreased, the coefficients' value increases. The feature Contract and Internet Service separate themselves from the remaining indicating how important they are for prediction.

The Ridge or the L2 regularization technique reduces the values of coefficients on the model. From the above plot, unless the penalty is huge(10^3), the coefficients are not reduced to zero.

$$\arg\max_{\beta} \quad \log \mathrm{PL}(\beta) - \frac{\alpha}{2} \sum_{j=1}^{p} \beta_j^2,$$

The coefficient change for varying α is inspected. Figure.17 shows that if the penalty has a large weight, all the coefficients shrink at most to zero. As the penalty's weight is decreased, the coefficients' value increases. The feature Contract and Internet Service separate themselves from the remaining indicating how important they are for prediction.

The Ridge or the L2 regularization technique reduces the values of coefficients on the model. From the above plot, unless the penalty is huge(10^3), the coefficients are not reduced to zero.

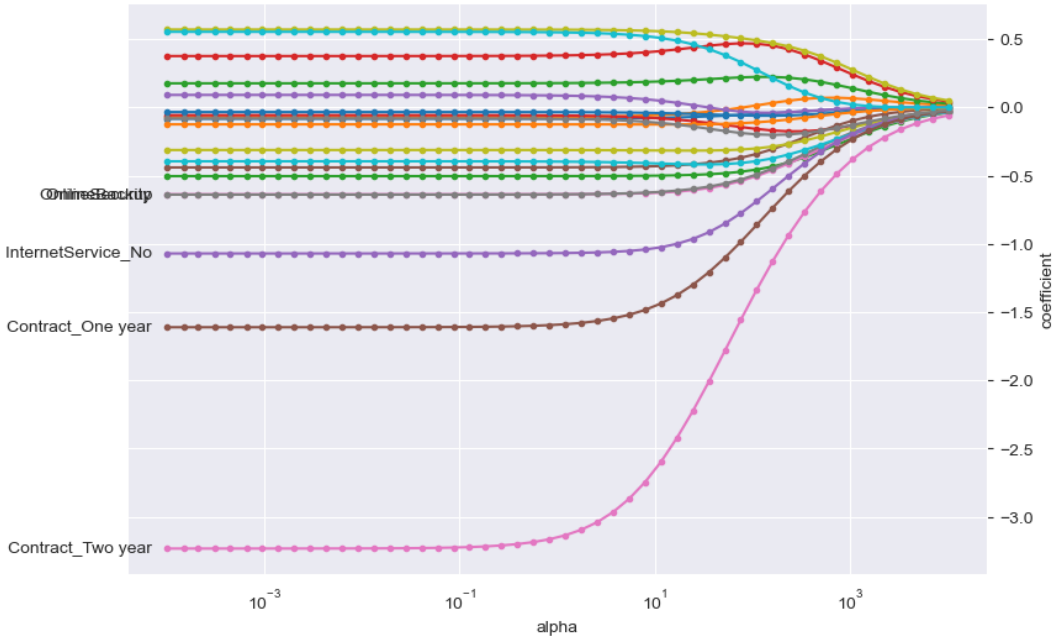


Figure 12. RIDGE - Coefficient Shrinkage Plot

LASSO

Lasso shrinks the coefficients towards 'zero' and also performs variable selection operations as well. As alpha increases from left to right, the variables are shrunk towards zero sets to zero. Lasso is mathematically represented as follows:

$$\arg\max_{\beta} \quad \log \mathrm{PL}(\beta) - \alpha \sum_{j=1}^{p} |\beta_j|.$$

The number of variables cannot be controlled in Lasso, but α is determined to choose the correct subset and work on it. Figure.18 shows that the LASSO penalty, indeed, is for a small selected subset of features. The pink line tends to be non-zero.
Also similar to the ridge, here the features stand out to be important for churn prediction. The plot Figure gives us the value of coefficients of the best model chosen with the Cross-Validation technique. Important variables are clearly noticeable in the below plot! We will see if other models choose the same variables as well.

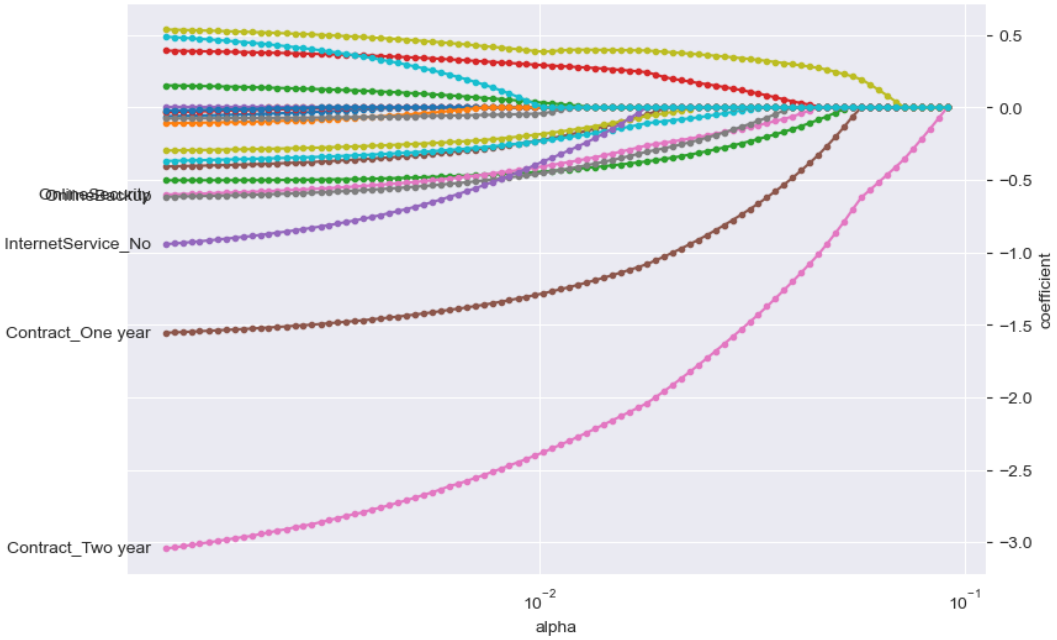


Figure 13. LASSO - Coefficient Shrinkage Plot

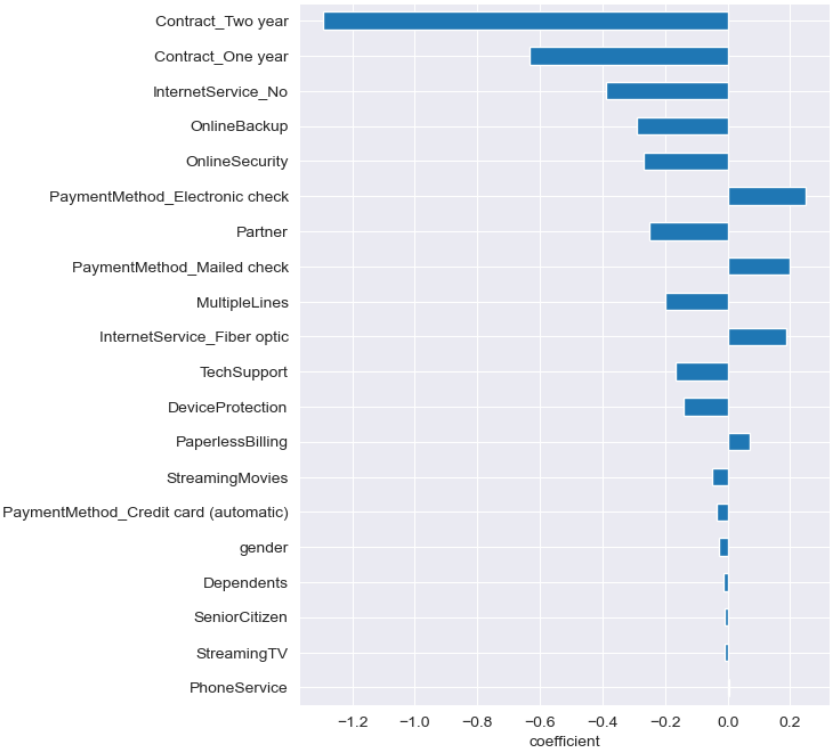


Figure 14. Coefficients for the features

The above figure gives us the value of coefficients of the best model chosen with the Cross-Validation technique. Important variables are clearly noticeable from the above plot! Contract, Internet Service, Payment Methods are some of the most important factors responsible for customer churning.
We will see if other models choose the same variables as well.

4.3 Parametric Methods

4.3.1 Accelerated Failure Time Model

Accelerated Failure Time model is a parametric method that provides an alternative to the commonly used Cox - PH model. In contrast to Cox Proportional, the AFT model assumes that the effect of a covariate is to accelerate or decelerate the life course of an event by some constant.

The AIC value for the Weibull model is less than the Exponential Model. Therefore, we will use that for parametric modeling.

A concordance index or c-index of 0.9 means that the model predicts the survival times with high accuracy. It indicates the model's ability to correctly rank the order of survival times of customers, such that those with shorter survival times are predicted to churn earlier, and those with higher survival times would stay with the company for longer.

Let us look at the variable - Contract.
Customers essentially register with the company on (a) a Monthly contract, (b) a One-Year contract, or (c) a Two-years contract.

1) Customers who enroll in the Two-Year option have an estimated survival time of exp(coef) = 16.9 times the ones who opted for the Monthly option(baseline model).

2) Similarly, customers who enrolled in One-Year contracts have an estimated survival time of exp(coef) = 4.8 times the Monthly customers. Therefore, the company should lure customers into opting for longer contracts in order to get a higher CLV(Customer Lifetime Value).

3) Another interesting service to look at is the company's Internet Service option. Customers who have NOT opted for the company's internet service have a positive coefficient of 1.5. This means that Customers who have opted Internet_Service = (No) have a 4.5 times better survival chance than the ones who have opted Internet_Service = (Yes). This probably means that the company's internet service is not up to the customer's expectations.

Concordance	0.9
AIC	17877.3
log-likelihood ratio test	3322.0 on 21 df
-log2(p) of ll-ratio test	inf

Table 3. Weibull Summary: Concordance Index, AIC, Ratio test values

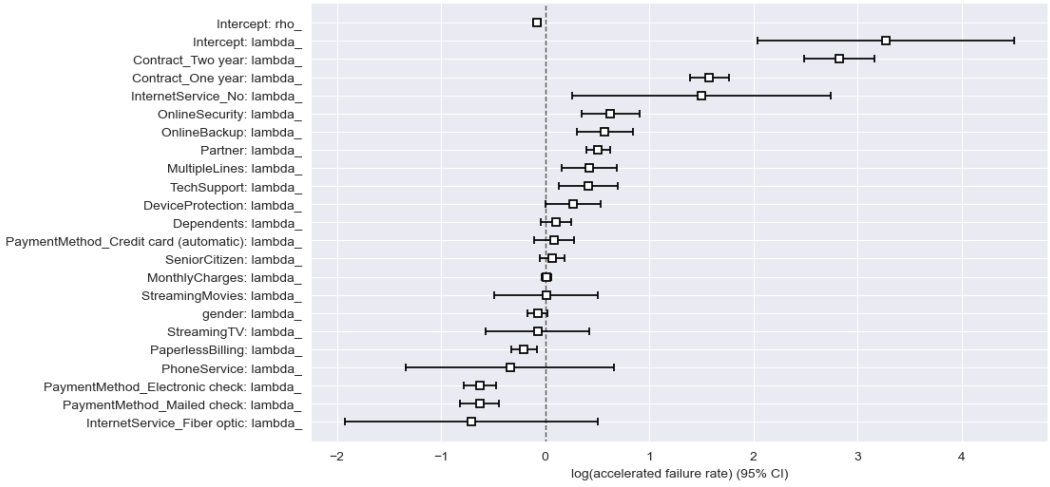


Figure 15. Plot of variable ranking based on log(accelerated failure rate)

5.1 Machine Learning Techniques

5.1.1 Random Survival Forest

Random Survival Forests (RSFs), unlike traditional regression methods such as the Cox proportional hazards model, can handle complex, nonlinear relationships between predictors and survival outcomes, making them a powerful tool in analyzing time-to-event data.

These work by creating an ensemble of decision trees, where each tree is grown on a randomly sampled subset of the predictors and the observations. This randomness helps to reduce overfitting and improve the accuracy of the model. The trees in the forest are then combined to make predictions about the risk of an event occurring at a given time and the probability of survival beyond that time.

```
y_val = np.rec.fromarrays([E_val.astype(bool), T_val], names=['event', 'time'])
c_index = rsf.score(X_val, y_val)
print("Concordance index on test set: {:.3f}".format(c_index))
```

```
Concordance index on test set: 0.845
```

Figure 16. C-index for test set

```
ci_rsf_trn = concordance_index(T_trn, -rsf.predict(X_trn), E_trn)
ci_rsf_val = concordance_index(T_val, -rsf.predict(X_val), E_val)

print(f'Concordance index of Random survival forest: train: {ci_rsf_trn:.3f}, valid: {ci_rsf_val:.3f}')
```

```
Concordance index of Random survival forest: train: 0.900, valid: 0.845
```

Figure 17. C-index for train and valid set

Figure 16 & 17 show how well the model performed on the testing and training data respectively. A concordance index of 85.5% is achieved on the testing data. This is ~5% less than the parametric model.

Figure 18 displays the important features and scores based on the Random Survival Forest model. Similar to what was inferred in the previous models, Contract, Payment Method and Monthly Charges seem to influence customer churn the most in this method as well.

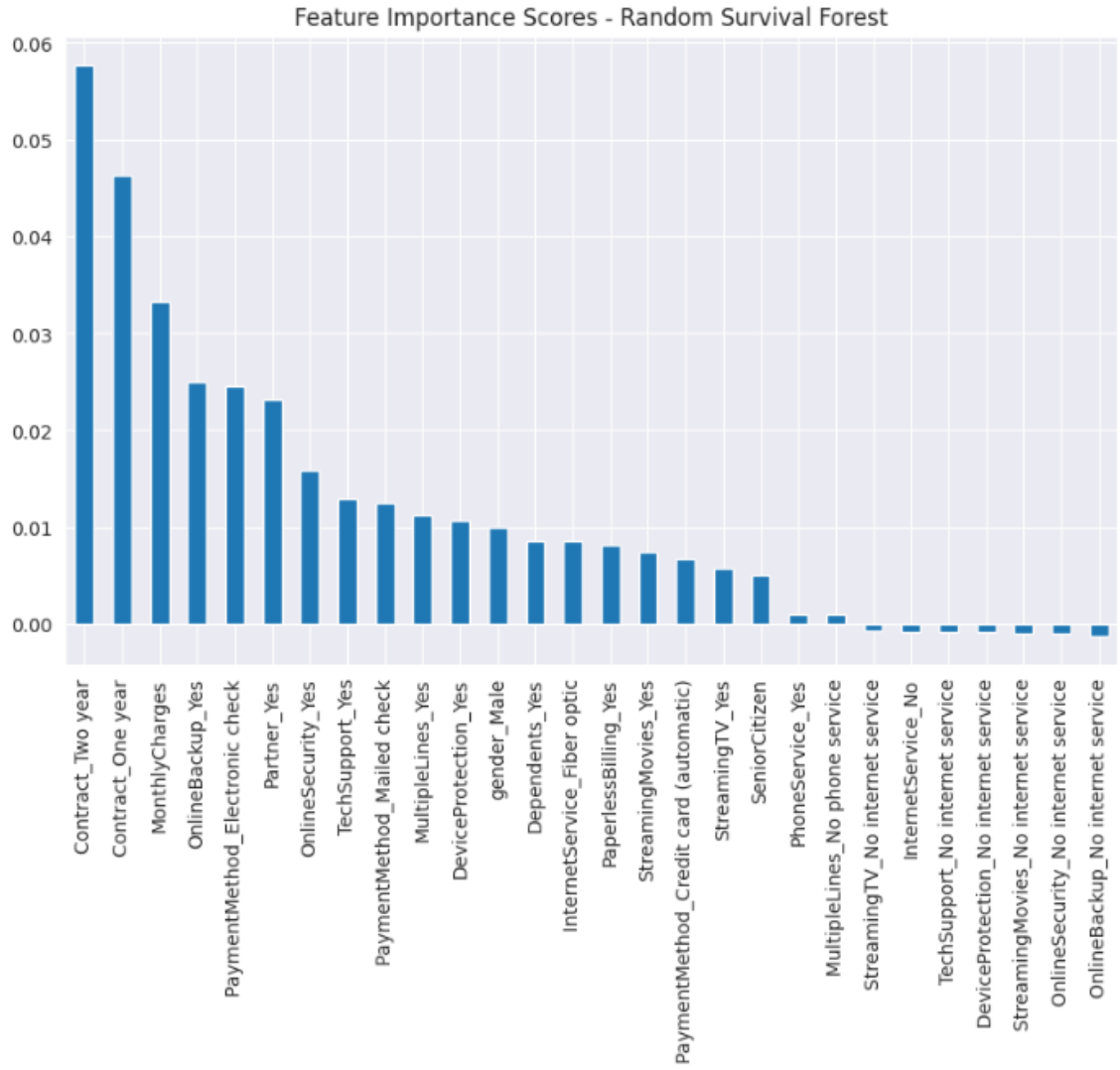


Figure 18. The feature importance scores for the RSF model

5. Result

The customer churn rate is investigated exhaustively using different survival methods. In terms of pure concordance index, the Weibull(or the AFT) model yielded the best c-index at 90%, followed by cox-proportional hazards model at 87% and random survival forest at 84.5%.

Every statistical method, parametric, non-parametric, or semi- parametric, has pros and cons. The non-parametric methods do not make any assumptions about the data distribution and hence are the closest to the real-life situation, and their plots are easy to in- terpret. The downside to this approach is that hazard ratios cannot be estimated. Another drawback is that it can only include categorical variables.

Parametric methods are conducted based on the assumption that hazard is constant over time. The Weibull model in this project allows hazards to increase/decrease over time. Parametric methods can be used to estimate the survival probabilities and hazard ratios.

Lastly, semi-parametric methods are a mix of parametric and non-parametric methods. These methods assume that hazards can fluctuate with time(similar to non-parametric methods).

6. Conclusion and Scope

6.1 Conclusion

From the above analysis - Customers who have opted for longer-term contracts have better survival probabilities than those who opted for shorter contracts. Payment Method is another influential variable for customer churn. Customers who have not opted for automatic payments are at a higher risk of churning than the ones who've scheduled automatic payments. Additionally, the company's internet service seems unsatisfactory for the customers, and those subscribed to the internet service and the network service have a significantly higher probability of Churning.

6.2 Scope

Nested modeling can be performed based on the importance of variables to keep only the statistically significant variables without affecting the concordance index and reducing the dimensionality simultaneously.

The data can further be divided into two parts - based on whether the customer has opted for the internet connection or not. This also facilitates analysis on a deeper level based on what exact service(of the ones who opted for internet service) is unsatisfactory to the customers.

For example, of customers who have internet service if they happen to stream a lot of movies and have low survival probabilities - it could mean that the company's internet service is slow.

References

[1] Ozer Çelik and Usame O Osmanoglu. Comparing to techniques used in customer churn analysis.
Journal of Multidisciplinary Developments, 4(1):30–38, 2019.

[2] Bingquan Huang, Mohand Tahar Kechadi, and Brian Buckley. Customer churn prediction in telecommunications. *Expert Systems with Applications*, 39(1):1414–1425, 2012.

[3] Junxiang Lu. Predicting customer churn in the telecommunications industry—-an application of survival analysis modeling using sas. *SAS User Group International (SUGI27) Online Proceedings*, 114, 2002.

[4] Melik Masarifoglu and Ali Hakan Buyuklu. Applying survival analysis to telecom churn data.
American Journal of Theoretical and Applied Statistics, 8(6):261–275, 2019.