Telco Customer Churn Prediction

Marketing 2505: Marketing Analytics

Group 1: Abir Chakraborty, Binbin Xia, Ritu Ranjani Ravi Shankar, Xiumin Sun

https://www.kaggle.com/blastchar/telco-customer-churn

Who We Are

Telco is a communications service provider (CSP), more precisely a telecommunications service provider (TSP), that provides telecommunications services such as telephony and data communications access. They offer a vast portfolio of products and solutions.

Telco Systems customer base comprises service providers ranging from newcomers to tier 1 internationals.

Business Objective

The **primary objective** is to predict the behavior to retain customers.

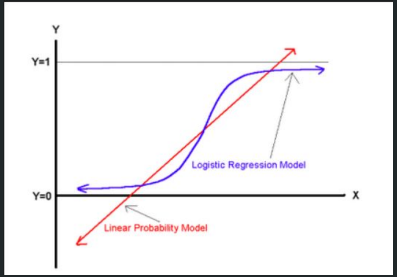
- Analyze the data to understand the reason for customer churn.
- Predict who will be the next person to leave the service.
- Provide suggestions to retain customers.

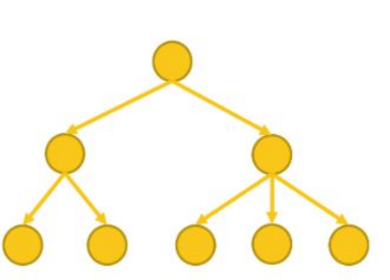
Dataset and its Content

- **Customer ID** – Each customer is identified by a unique customer_id.
- **Demographic Information about Customers** – Gender, Senior/Not, Partner/Dependents.
- **Services availed by Each Customer** – 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.
- **Customers Churn** – Yes/No.

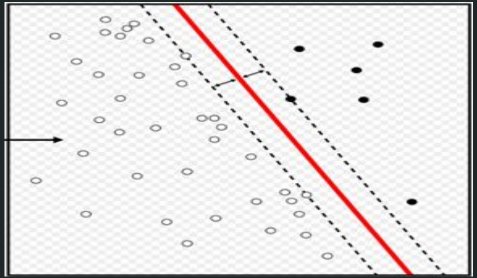
Methods

- Random Forest
- Naive Bayes Classifier
- Binary Logit Model (BLM)
- Survival Analysis
- Support Vector Machine (SVM)





$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$



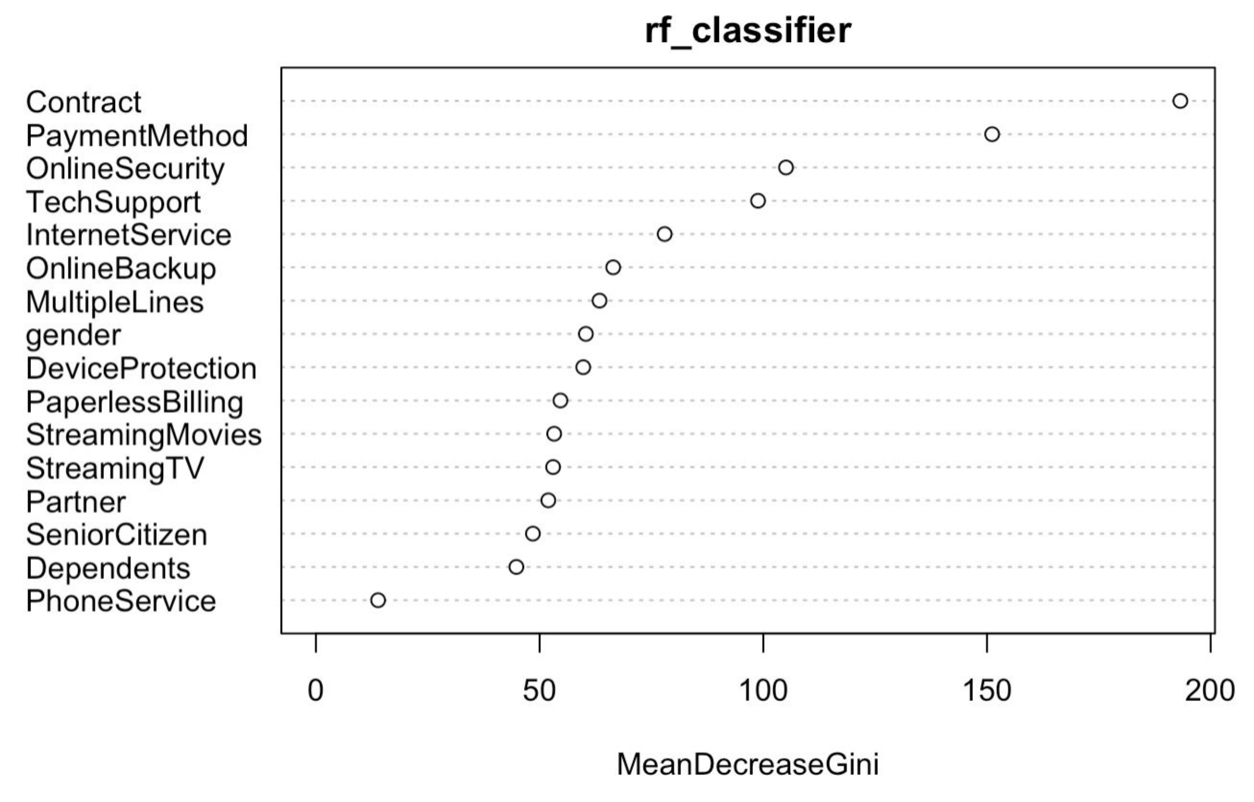
Random Forest

- An ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes

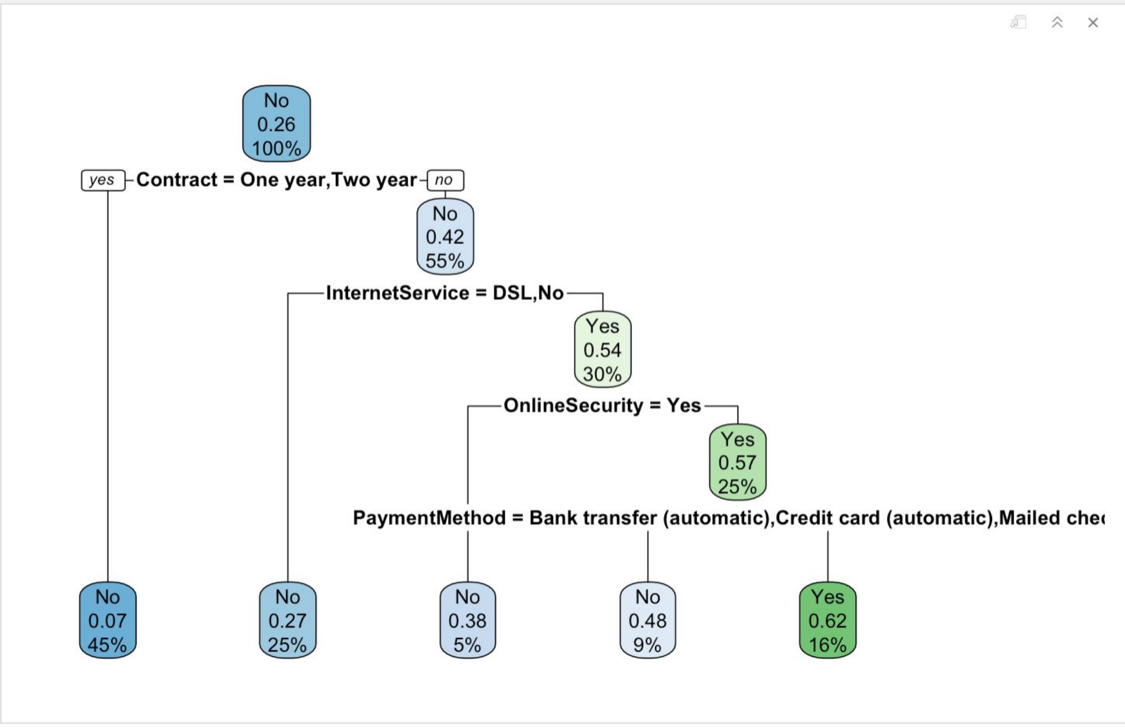
- Using random forest before running other models, in order to get important variables

- Drop off variables with high correlation (Total Charges, Tenure)

Random Forest - Variable Importance Plot



Random Forest Classifier



Important variables for customer churn: Contract, Internet Service, Online Security, and Payment Method.

Naive Bayes Classifier

A family of simple "probabilistic classifiers" based on applying Bayes' theorem with strong (naïve) independence assumptions between the features.

Category	Variables	Likely to Churn
Demographic	Partner and Dependent	-18%
Service	Online Security Service	-39%
	Online Backup Service	-30%
	Device Protection	-28%
	Tech Support	-38%
	Streaming TV/Movies	-15%

Naive Bayes Classifier

Category	Variables	Likely to Churn
Service	Internet Service - Fiber Optic	+30%
Contract	Month-to-Month Contract	+42%
Billing Method	Paperless Billing	+20%
	Electronic Check	+32%

Binary Logistic Model

A statistical technique used to predict the relationship between predictors (our independent variables) and a predicted variable (the dependent variable) where the dependent variable is **binary.**

Choose the model with lowest AIC.

Variables	Estimation	Z Score	P-Value
Senior Citizen	1.006635	3.450	0.0056
Paperless Billing	0.387897	5.300	1.16e-07
Electronic Check	0.389568	4.188	2.81e-05
Online Security Yes	-0.793644	-5.869	4.38e-09

Binary Logistic Model

Variables	Estimation	Z Score	P-Value
Contract One Year	-0.795103	-7.649	2.02e-14
Contract Two Year	-1.602437	-9.159	< 2e-16
Monthly Charges	0.021501	10.908	< 2e-16

Survival Analysis

Estimates the time for a customer to churn from the service.

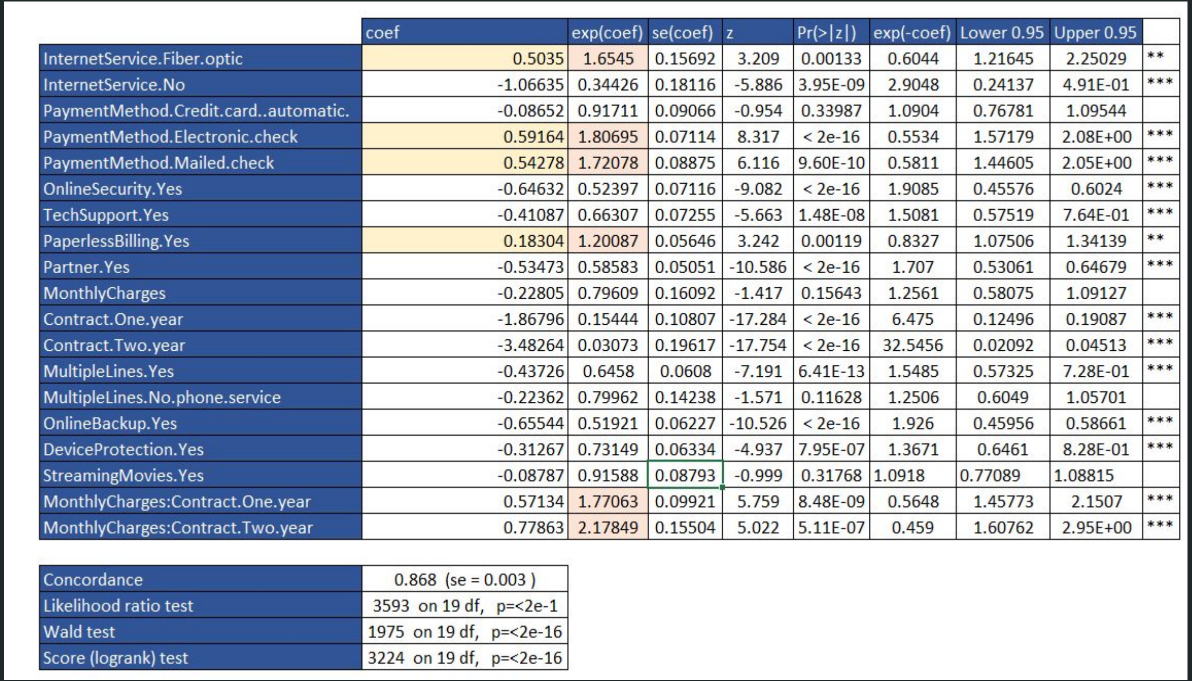
- Cox Proportional Hazard Regression Model

  The proportional hazards model is similar to multiple regression, here the unique effect of a unit increase in a covariate is multiplicative with respect to the hazard rate.

- Kaplan-Meier Estimate

  The Kaplan-Meier Survival Curve gives the probability of surviving in a given length of time.

Cox Proportional Hazard Estimate



	coef	exp(coef)	se(coef)	z	Pr(>|z|)	exp(-coef)	Lower 0.95	Upper 0.95	
InternetService.Fiber.optic	0.5035	1.6545	0.15692	3.209	0.00133	0.6044	1.21645	2.25029	**
InternetService.No	-1.06635	0.34426	0.18116	-5.886	3.95E-09	2.9048	0.24137	4.91E-01	***
PaymentMethod.Credit.card..automatic.	-0.08652	0.91711	0.09066	-0.954	0.33987	1.0904	0.76781	1.09544	
PaymentMethod.Electronic.check	0.59164	1.80695	0.07114	8.317	< 2e-16	0.5534	1.57179	2.08E+00	***
PaymentMethod.Mailed.check	0.54278	1.72078	0.08875	6.116	9.60E-10	0.5811	1.44605	2.05E+00	***
OnlineSecurity.Yes	-0.64632	0.52397	0.07116	-9.082	< 2e-16	1.9085	0.45576	0.6024	***
TechSupport.Yes	-0.41087	0.66307	0.07255	-5.663	1.48E-08	1.5081	0.57519	7.64E-01	***
PaperlessBilling.Yes	0.18304	1.20087	0.05646	3.242	0.00119	0.8327	1.07506	1.34139	**
Partner.Yes	-0.53473	0.58583	0.05051	-10.586	< 2e-16	1.707	0.53061	0.64679	***
MonthlyCharges	-0.22805	0.79609	0.16092	-1.417	0.15643	1.2561	0.58075	1.09127	
Contract.One.year	-1.86796	0.15444	0.10807	-17.284	< 2e-16	6.475	0.12496	0.19087	***
Contract.Two.year	-3.48264	0.03073	0.19617	-17.754	< 2e-16	32.5456	0.02092	0.04513	***
MultipleLines.Yes	-0.43726	0.6458	0.0608	-7.191	6.41E-13	1.5485	0.57325	7.28E-01	***
MultipleLines.No.phone.service	-0.22362	0.79962	0.14238	-1.571	0.11628	1.2506	0.6049	1.05701	
OnlineBackup.Yes	-0.65544	0.51921	0.06227	-10.526	< 2e-16	1.926	0.45956	0.58661	***
DeviceProtection.Yes	-0.31267	0.73149	0.06334	-4.937	7.95E-07	1.3671	0.6461	8.28E-01	***
StreamingMovies.Yes	-0.08787	0.91588	0.08793	-0.999	0.31768	1.0918	0.77089	1.08815	
MonthlyCharges:Contract.One.year	0.57134	1.77063	0.09921	5.759	8.48E-09	0.5648	1.45773	2.1507	***
MonthlyCharges:Contract.Two.year	0.77863	2.17849	0.15504	5.022	5.11E-07	0.459	1.60762	2.95E+00	***

Concordance	0.868 (se = 0.003)
Likelihood ratio test	3593 on 19 df, p=<2e-1
Wald test	1975 on 19 df, p=<2e-16
Score (logrank) test	3224 on 19 df, p=<2e-16

** Feature selection performed using STEPAIC() method.

Interpretation:

1) Compared to those with InternetService.DSL, customers with InternetService.FibreOptic are at a higher risk of churning, i.e they are 1.65 times more likely to churn for a unit change in the covariate.
2) Compared to automatic transfers, electronic and mailed checks are at higher risk of churning with hazard rate of 1.80 and 1.72 respectively.
3) Compared to customers who didn't opt for PaperlessBilling, those who opted are at a higher risk of churning with a hazard rate of 1.20

Model Fit:

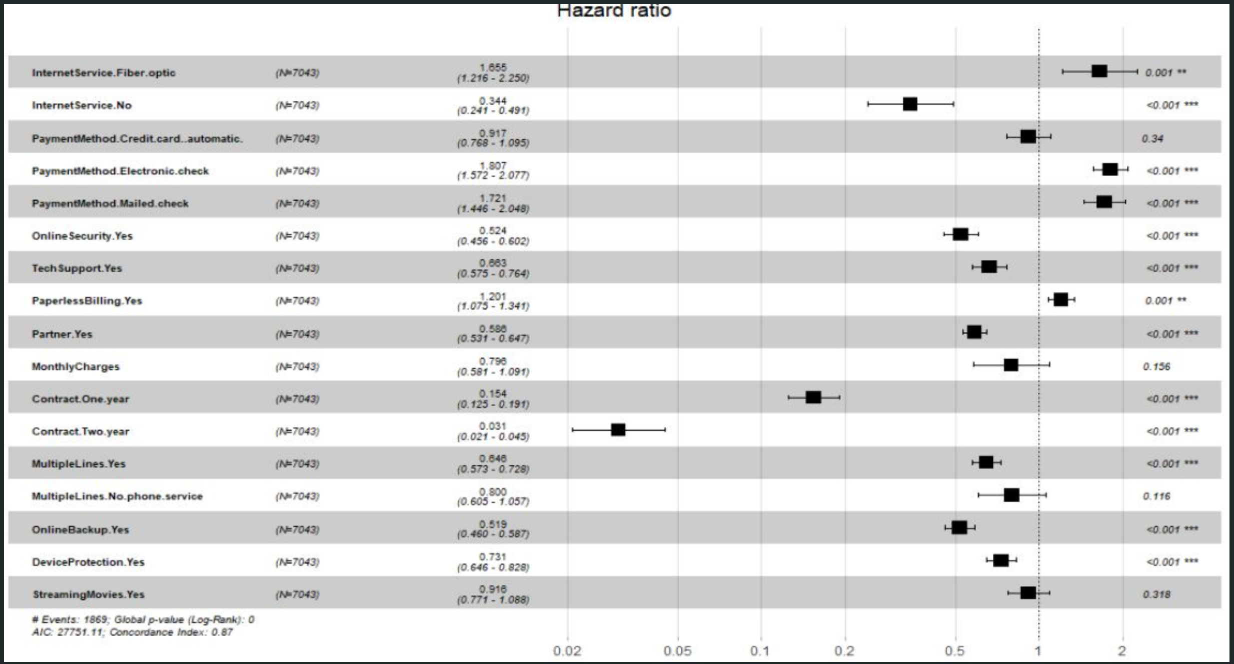
Almost every element in the model is significant.

Concordance index = 0.86.

**c-index = 1 is perfect concordance.

Likelihood, Wald and logrank test are all significant.

Visualize Cox Proportional Hazard using ggforest

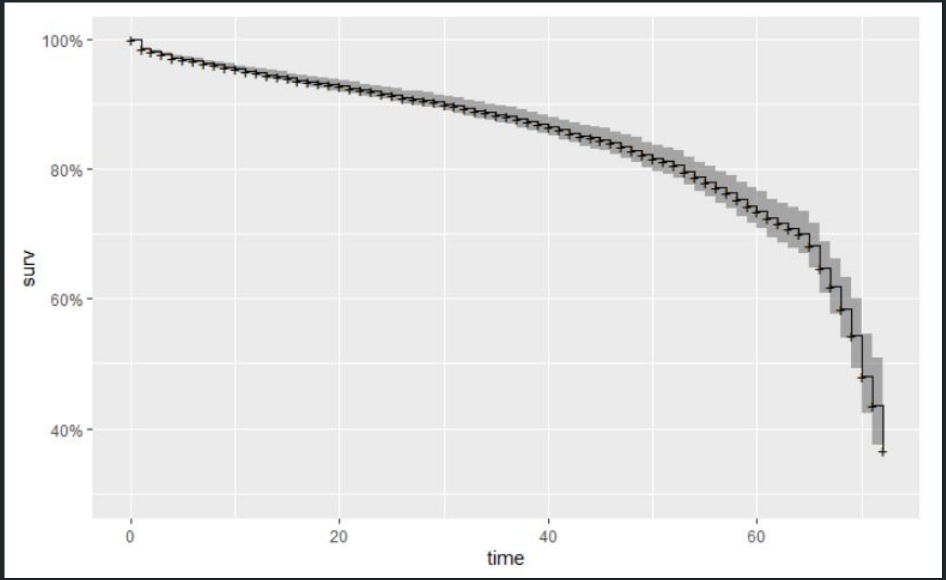


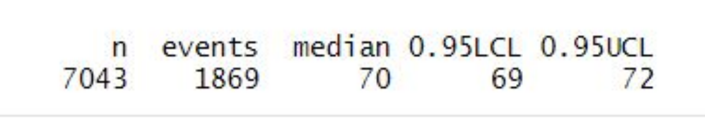
InternetService.Fiber.optic	(N=7043)	1.655 (1.216 - 2.250)	0.001 **
InternetService.No	(N=7043)	0.344 (0.241 - 0.491)	<0.001 ***
PaymentMethod.Credit.card..automatic.	(N=7043)	0.917 (0.768 - 1.095)	0.34
PaymentMethod.Electronic.check	(N=7043)	1.807 (1.572 - 2.077)	<0.001 ***
PaymentMethod.Mailed.check	(N=7043)	1.721 (1.446 - 2.048)	<0.001 ***
OnlineSecurity.Yes	(N=7043)	0.524 (0.456 - 0.602)	<0.001 ***
TechSupport.Yes	(N=7043)	0.663 (0.575 - 0.764)	<0.001 ***
PaperlessBilling.Yes	(N=7043)	1.201 (1.075 - 1.341)	0.001 **
Partner.Yes	(N=7043)	0.586 (0.531 - 0.647)	<0.001 ***
MonthlyCharges	(N=7043)	0.796 (0.581 - 1.091)	0.156
Contract.One.year	(N=7043)	0.154 (0.125 - 0.191)	<0.001 ***
Contract.Two.year	(N=7043)	0.031 (0.021 - 0.045)	<0.001 ***
MultipleLines.Yes	(N=7043)	0.646 (0.573 - 0.728)	<0.001 ***
MultipleLines.No.phone.service	(N=7043)	0.800 (0.605 - 1.057)	0.116
OnlineBackup.Yes	(N=7043)	0.519 (0.460 - 0.587)	<0.001 ***
DeviceProtection.Yes	(N=7043)	0.731 (0.646 - 0.828)	<0.001 ***
StreamingMovies.Yes	(N=7043)	0.916 (0.771 - 1.088)	0.318

Events: 1869; Global p-value (Log-Rank): 0
AIC: 27751.11; Concordance Index: 0.87

Using this model, we can understand how each attribute of subscription and customer demographic influence the risk of churning.

Visualize Cox Proportional Hazard using Survfit()



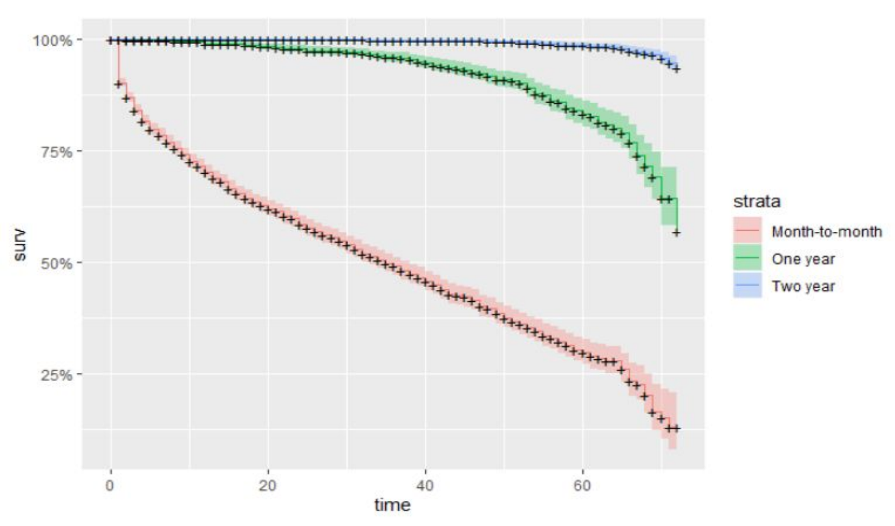


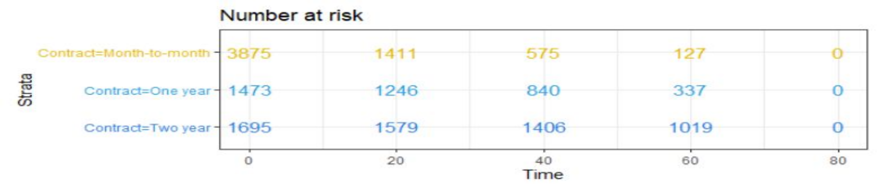
```
   n  events  median 0.95LCL 0.95UCL
7043    1869      70      69      72
```

Out of 7043 observations, 1869 churned. We see from the Coxph graph that, at 72 months, the survival rate of the customers is almost 38%.

Kaplan-Meier Estimate

Single Curve against Contract





Number at risk

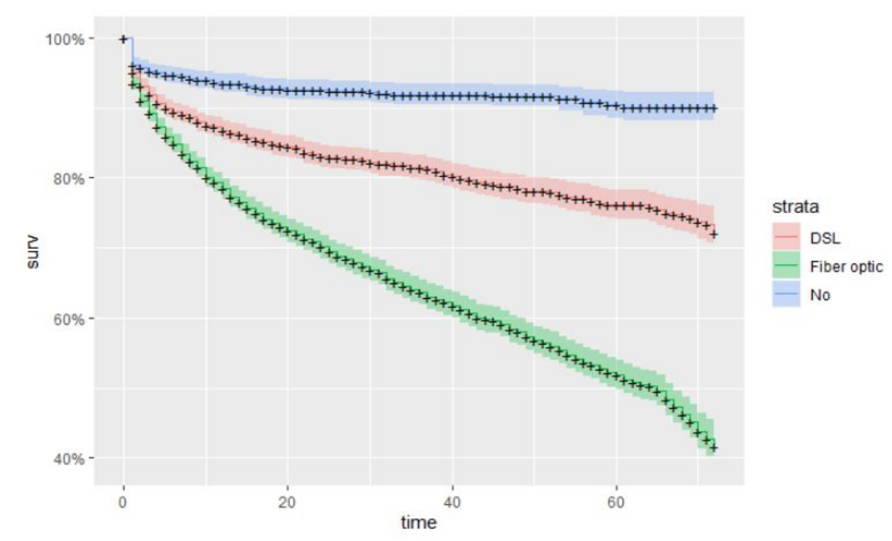
Strata	0	20	40	60	80
Contract=Month-to-month	3875	1411	575	127	0
Contract=One year	1473	1246	840	337	0
Contract=Two year	1695	1579	1406	1019	0

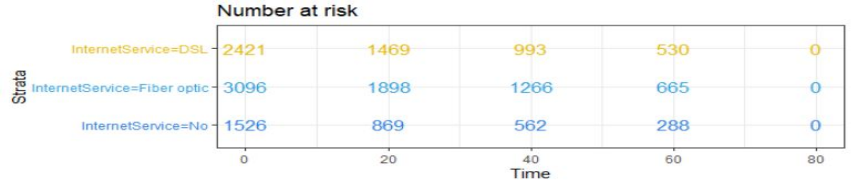
Time

Customers in month-to-month contract have lower chances of surviving than one/year contracts.

Kaplan-Meier Estimate

Survival curve against Internet Service





Number at risk

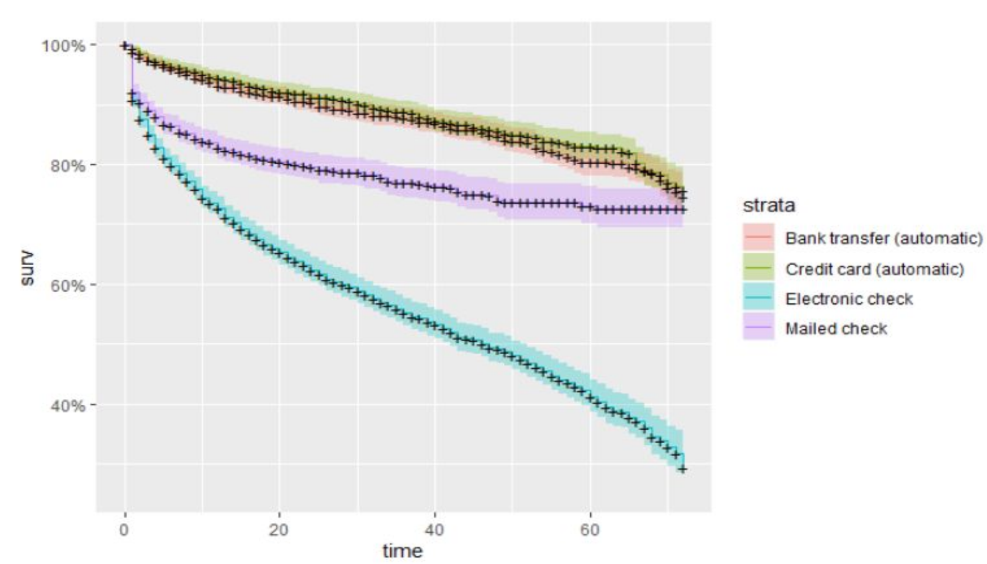
Strata	0	20	40	60	80
InternetService=DSL	2421	1469	993	530	0
InternetService=Fiber optic	3096	1898	1266	665	0
InternetService=No	1526	869	562	288	0

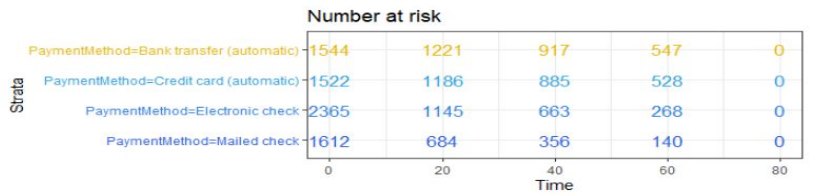
Time

There were 3096 customers at the beginning of the study and after 20 months, only 1898 customers remained in the Fibre Optic internet service

Kaplan-Meier Estimate

Survival curve against Payment Method




Number at risk

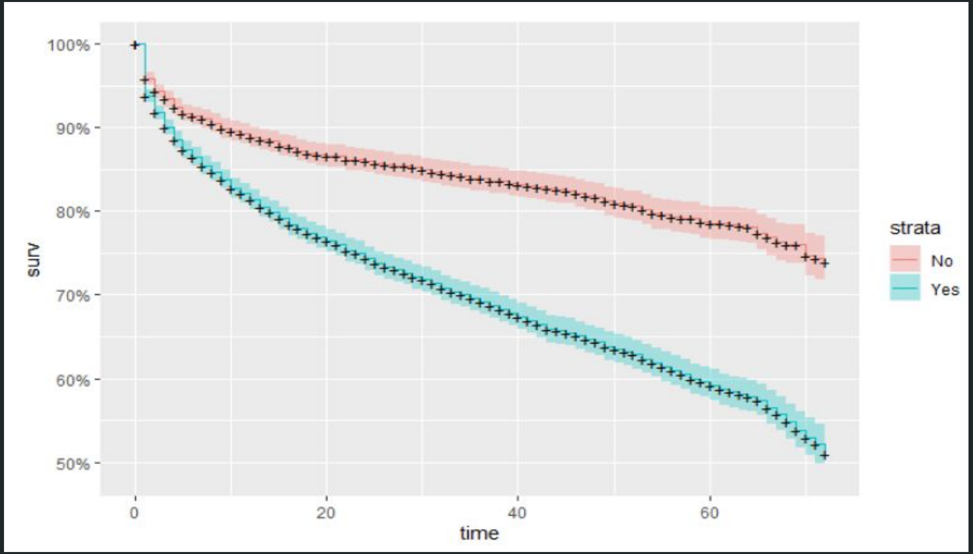
Strata	0	20	40	60	80
PaymentMethod=Bank transfer (automatic)	1544	1221	917	547	0
PaymentMethod=Credit card (automatic)	1522	1186	885	528	0
PaymentMethod=Electronic check	2365	1145	663	268	0
PaymentMethod=Mailed check	1612	684	356	140	0

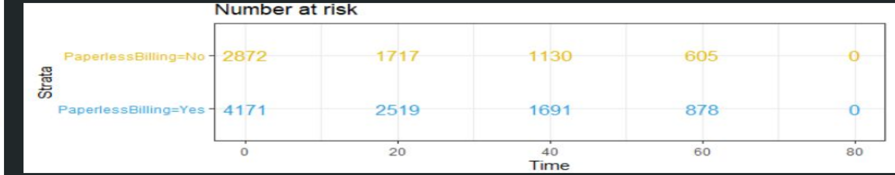
Time

At 20th month only 65% of the customers who paid through Electronic check survived.

Kaplan-Meier Estimate

Survival curve against Paperless Billing





Number at risk

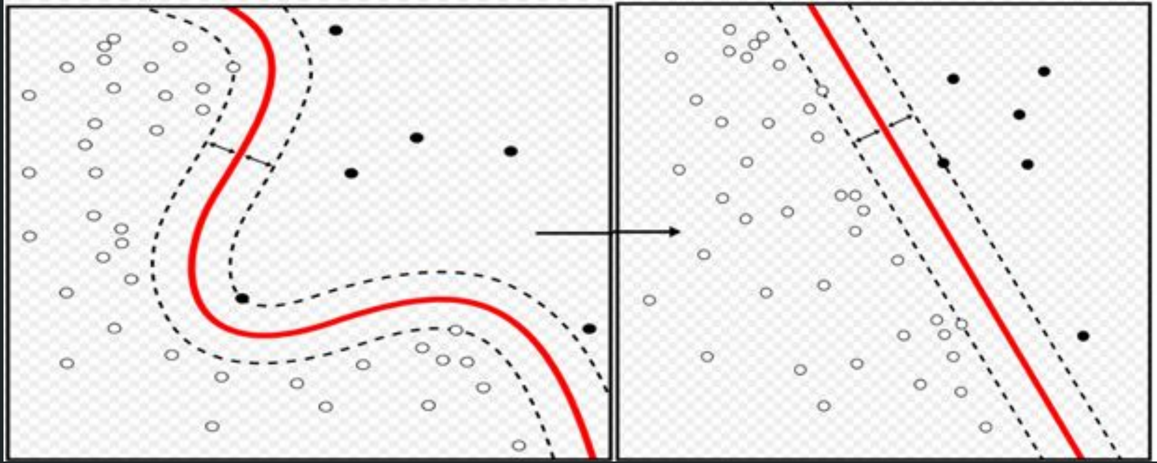
Strata	0	20	40	60	80
PaperlessBilling=No	2872	1717	1130	605	0
PaperlessBilling=Yes	4171	2519	1691	878	0

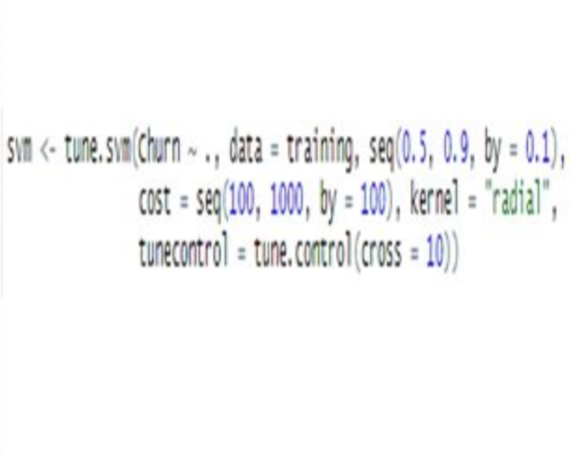
Time

At 20th month only 75% of the customers who availed Paperless Billing survived.

Support Vector Machine Model

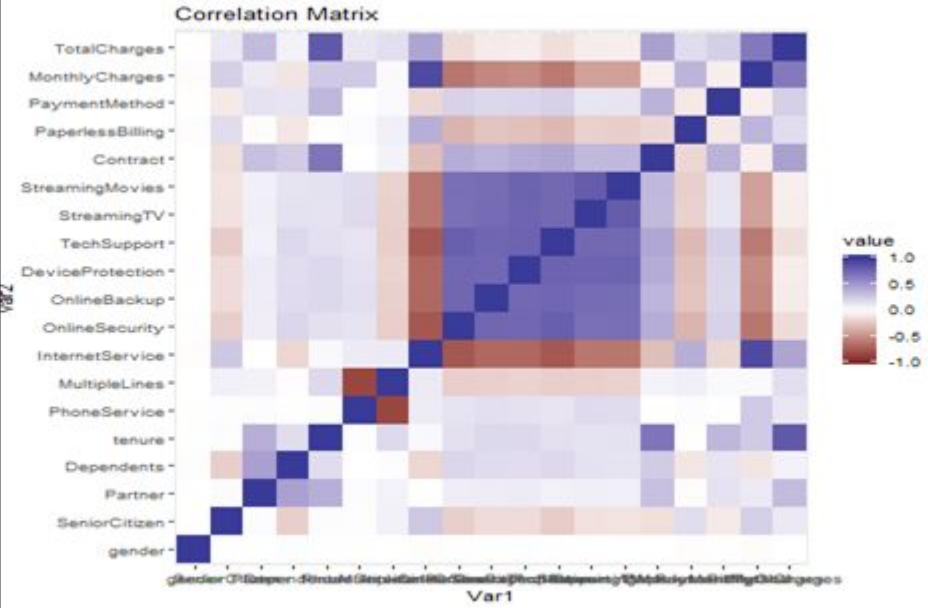
Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier (although methods such as Platt scaling exist to use SVM in a probabilistic classification setting)

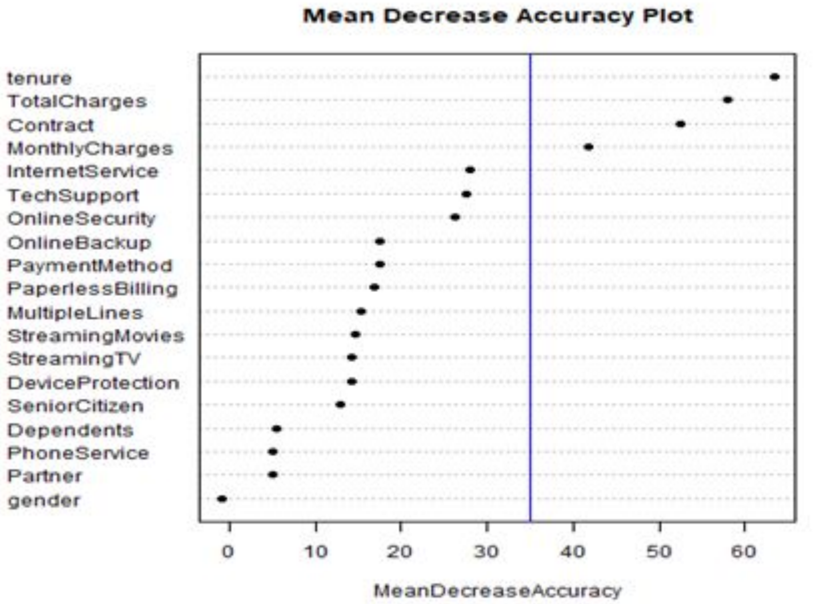




```r
svm <- tune.svm(Churn ~ ., data = training, seq(0.5, 0.9, by = 0.1),
                cost = seq(100, 1000, by = 100), kernel = "radial",
                tunecontrol = tune.control(cross = 10))
```

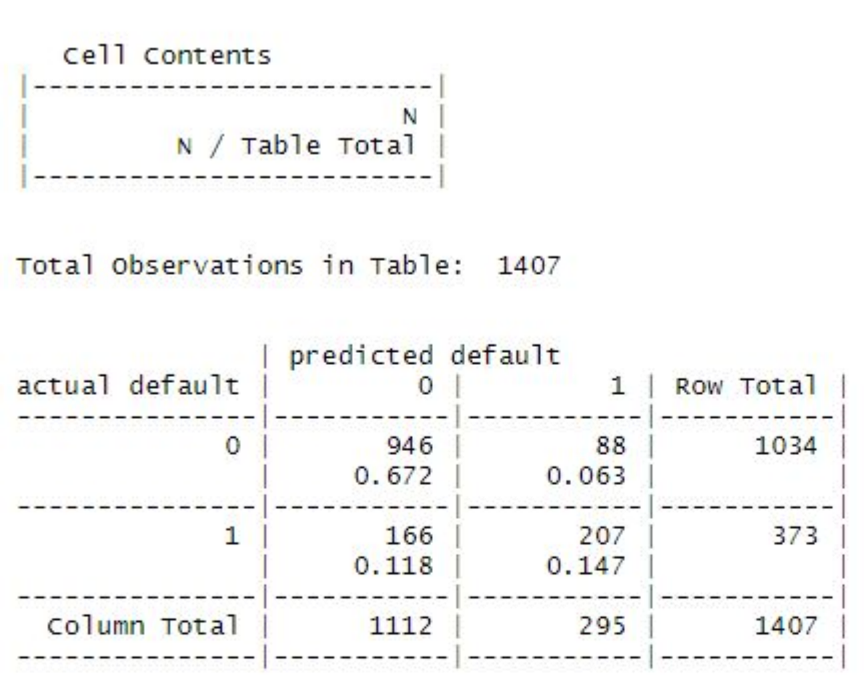
Support Vector Machine Model





Confusion Matrix: a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known

Support Vector Machine Model



```
   Cell Contents
|-------------------------|
|                       N |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  1407


               | predicted default
actual default |         0 |         1 | Row Total |
---------------|-----------|-----------|-----------|
             0 |       946 |        88 |      1034 |
               |     0.672 |     0.063 |           |
---------------|-----------|-----------|-----------|
             1 |       166 |       207 |       373 |
               |     0.118 |     0.147 |           |
---------------|-----------|-----------|-----------|
  Column Total |      1112 |       295 |      1407 |
---------------|-----------|-----------|-----------|
```

True positives (TP): These are cases in which we predicted yes (they have the disease), and they do churn.
True negatives (TN): We predicted no, and they don't churn.
False positives (FP): We predicted yes, but they don't actually churn . (Also known as a "Type I error.")

False negatives (FN): We predicted no, but they actually do churn. (Also known as a "Type II error.")

Summary From Models

Churn More	Internet - Fiber Optic Payment Method - Electronic/Mailed Contract - Month-to-Month Paperless Billing - Yes
Churn Less	Online Backup - Yes Online Security - Yes Multiple Line - Yes Streaming Movies - Yes

Action Points

- Create campaigns and promotions to convert Month-to-Month customers into yearly contract customers.
- Create Ads to target families with Partners/Dependents as large families tend to churn less.
- Promote automatic payment options.
- Lower the cost of Fiber Optic Internet Service/ Improve service quality.

Thank You!!!!