30416 Big Data and Databases Final Project

By: Aly Elagroudy
ID: 3085293

Dataset Description and Goal of Analysis:

This is an individual project for the Big Data and Databases course analyzing an IBM sample dataset for a telecommunications company.

The dataset includes the following:

- The customer churn status (whether they left the company or not)
- The various services the customer might have subscribed to (categorical variables)
- Information about contract length, payment methods chosen, monthly and total charges
- Demographic information (Age range, Gender, etc.)

All in all, there are 21 categorical variables in this dataset with each row representing a different customer. The raw data contains 7043 rows (customers). The dataset can be found at: https://www.kaggle.com/blastchar/telco-customer-churn

The goal of this analysis is to try and predict customer churn (target variable) using the other 20 feature variables available per customer. This can help the telecommunications company personalize customer retention programs, better target them, and overall, potentially improve churn rate for the future. Lastly, the Analysis is done using the Knime Analytics Platform and all the steps for the analysis will be commented on in this report.

Preliminary Data Preparation

customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	OnlineBackup	DeviceProtection	TechSupport	StreamingTV	StreamingMovies	Contract	PaperlessBilling	PaymentMethod	MonthlyCharges	TotalCharges	Churn
7590-VHVEG	Female	0	Yes	No	1	No	No phone service	DSL	No	Yes	No	No	No	No	Month-to-month	Yes	Electronic check	29.85	29.85	No
5575-GNVDE	Male	0	No	No	34	Yes	No	DSL	Yes	No	Yes	No	No	No	One year	No	Mailed check	56.95	1889.5	No
3668-QPYBK	Male	0	No	No	2	Yes	No	DSL	Yes	Yes	No	No	No	No	Month-to-month	Yes	Mailed check	53.85	108.15	Yes
7795-CFOCW	Male	0	No	No	45	No	No phone service	DSL	Yes	No	Yes	Yes	No	No	One year	No	Bank transfer (autom...	42.3	1840.75	No
9237-HQITU	Female	0	No	No	2	Yes	No	Fiber optic	No	No	No	No	No	No	Month-to-month	Yes	Electronic check	70.7	151.65	Yes
9305-CDSKC	Female	0	No	No	8	Yes	Yes	Fiber optic	No	No	Yes	No	Yes	Yes	Month-to-month	Yes	Electronic check	99.65	820.5	Yes
1452-KIOVK	Male	0	No	Yes	22	Yes	Yes	Fiber optic	No	Yes	No	No	Yes	No	Month-to-month	Yes	Credit card (automatic)	89.1	1949.4	No
6713-OKOMC	Female	0	No	No	10	No	No phone service	DSL	Yes	No	No	No	No	No	Month-to-month	No	Mailed check	29.75	301.9	No
7892-POOKP	Female	0	Yes	No	28	Yes	Yes	Fiber optic	No	No	Yes	Yes	Yes	Yes	Month-to-month	Yes	Electronic check	104.8	3046.05	Yes
6388-TABGU	Male	0	No	Yes	62	Yes	No	DSL	Yes	Yes	No	No	No	No	One year	No	Bank transfer (autom...	56.15	3487.95	No
9763-GRSKD	Male	0	Yes	Yes	13	Yes	No	DSL	Yes	No	No	No	No	No	Month-to-month	Yes	Mailed check	49.95	587.45	No
7469-LKBCI	Male	0	No	No	16	Yes	No	No	No internet service	No internet ser...	No internet service	No internet service	No internet ser...	No internet service	Two year	No	Credit card (automatic)	18.95	326.8	No
8091-TTVAX	Male	0	Yes	No	58	Yes	Yes	Fiber optic	No	No	Yes	No	Yes	Yes	One year	No	Credit card (automatic)	100.35	5681.1	No
0280-XJGEX	Male	0	No	No	49	Yes	Yes	Fiber optic	No	Yes	Yes	No	Yes	Yes	Month-to-month	Yes	Bank transfer (autom...	103.7	5036.3	Yes
5129-JLPIS	Male	0	No	No	25	Yes	No	Fiber optic	Yes	No	Yes	Yes	Yes	Yes	Month-to-month	Yes	Electronic check	105.5	2686.05	No
3655-SNQYZ	Female	0	Yes	Yes	69	Yes	Yes	Fiber optic	Yes	Yes	Yes	Yes	Yes	Yes	Two year	No	Credit card (automatic)	113.25	7895.15	No
8191-XWSZG	Female	0	No	No	52	Yes	No	No	No internet service	No internet ser...	No internet service	No internet service	No internet ser...	No internet service	One year	No	Mailed check	20.65	1022.95	No
9959-WOFKT	Male	0	No	Yes	71	Yes	Yes	Fiber optic	Yes	No	Yes	No	Yes	Yes	Two year	No	Bank transfer (autom...	106.7	7382.25	No
4190-MFLUW	Female	0	Yes	Yes	10	Yes	No	DSL	No	No	Yes	Yes	No	No	Month-to-month	No	Credit card (automatic)	55.2	528.35	Yes
4183-MYFRB	Female	0	No	No	21	Yes	No	Fiber optic	No	Yes	Yes	No	No	Yes	Month-to-month	Yes	Electronic check	90.05	1862.9	No
8779-QRDMV	Male	1	No	No	1	No	No phone service	DSL	No	No	Yes	No	No	Yes	Month-to-month	Yes	Electronic check	39.65	39.65	Yes
1680-VDCWW	Male	0	Yes	No	12	Yes	No	No	No internet service	No internet ser...	No internet service	No internet service	No internet ser...	No internet service	One year	No	Bank transfer (autom...	19.8	202.25	No
1066-JKSGK	Male	0	No	No	1	Yes	No	No	No internet service	No internet ser...	No internet service	No internet service	No internet ser...	No internet service	Month-to-month	No	Mailed check	20.15	20.15	Yes
3638-WEABW	Female	0	Yes	No	58	Yes	Yes	DSL	No	Yes	No	Yes	No	No	Two year	Yes	Credit card (automatic)	59.9	3505.1	No
6322-HRPFA	Male	0	Yes	Yes	49	Yes	No	DSL	Yes	Yes	No	Yes	No	No	Month-to-month	No	Credit card (automatic)	59.6	2970.3	No
6865-JZNKO	Female	0	No	No	30	Yes	No	DSL	Yes	Yes	No	No	No	No	Month-to-month	Yes	Bank transfer (autom...	55.3	1530.6	No
6467-CHFZW	Male	0	Yes	Yes	47	Yes	Yes	Fiber optic	No	Yes	No	No	Yes	Yes	Month-to-month	Yes	Electronic check	99.35	4749.15	Yes
8665-UTDHZ	Male	0	Yes	Yes	1	No	No phone service	DSL	No	Yes	No	No	No	No	Month-to-month	No	Electronic check	30.2	30.2	Yes
5248-YGIJN	Male	0	Yes	No	72	Yes	Yes	DSL	Yes	Yes	Yes	Yes	Yes	Yes	Two year	Yes	Credit card (automatic)	90.25	6369.45	No
8773-HHUOZ	Female	0	No	Yes	17	Yes	No	DSL	No	No	No	No	Yes	Yes	Month-to-month	Yes	Mailed check	64.7	1093.1	Yes
3841-NFECX	Female	1	Yes	No	71	Yes	Yes	Fiber optic	Yes	Yes	Yes	Yes	No	No	Two year	Yes	Credit card (automatic)	96.35	6766.95	No
4929-XIHVW	Male	1	Yes	No	2	Yes	No	Fiber optic	No	No	Yes	No	Yes	Yes	Month-to-month	Yes	Credit card (automatic)	95.5	181.65	No
6827-IEAUQ	Female	0	Yes	Yes	27	Yes	No	DSL	Yes	Yes	Yes	Yes	No	No	One year	No	Mailed check	66.15	1874.45	No
7310-EGVHZ	Male	0	No	No	1	Yes	No	No	No internet service	No internet ser...	No internet service	No internet service	No internet ser...	No internet service	Month-to-month	No	Bank transfer (autom...	20.2	20.2	No
3413-BMNZE	Male	1	No	No	1	Yes	No	DSL	No	No	No	No	No	No	Month-to-month	No	Bank transfer (autom...	45.25	45.25	No
6234-RAAPL	Female	0	Yes	Yes	72	Yes	Yes	Fiber optic	Yes	Yes	No	Yes	Yes	No	Two year	No	Bank transfer (autom...	99.9	7251.7	No
6047-YHPVI	Male	0	No	No	5	Yes	No	Fiber optic	No	No	No	No	No	No	Month-to-month	Yes	Electronic check	69.7	316.9	Yes
6572-ADKRS	Female	0	No	No	46	Yes	No	Fiber optic	No	No	Yes	No	No	No	Month-to-month	Yes	Credit card (automatic)	74.8	3548.3	No

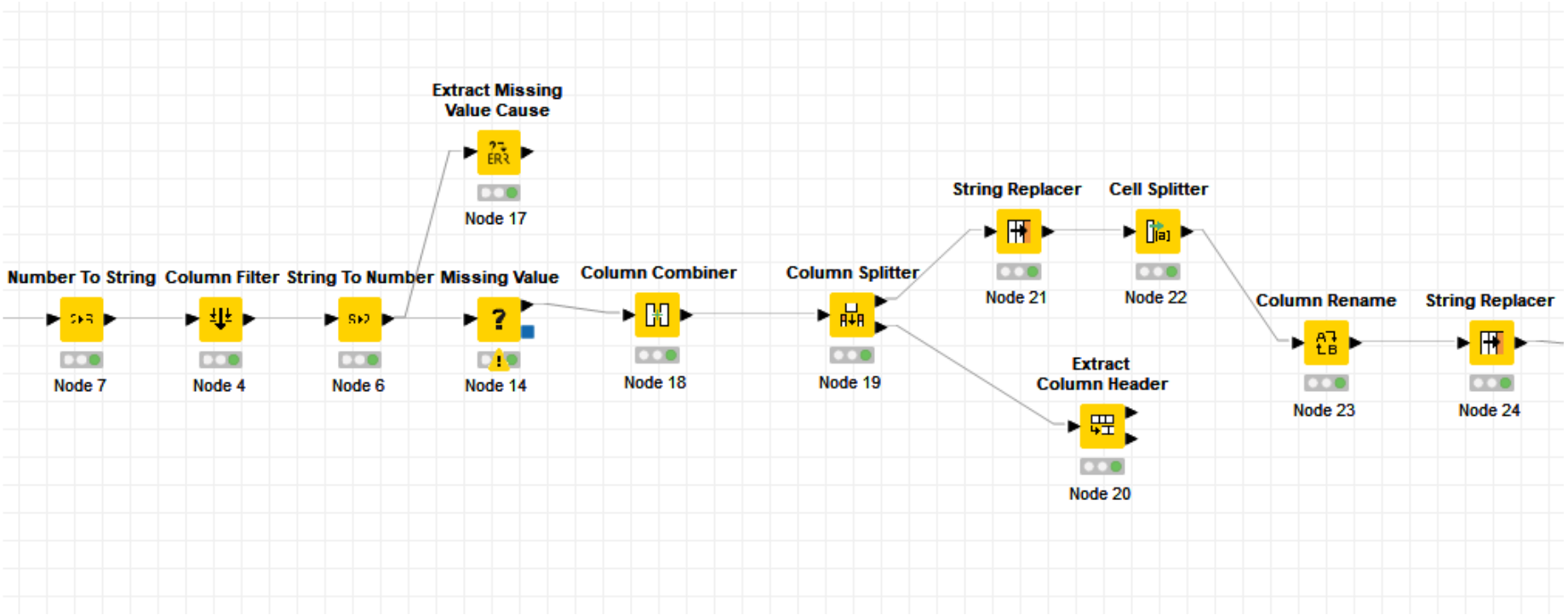
Before any descriptive analysis could be made on the dataset, it had to be prepared in a few ways. The above table is a section of the original dataset.

We can quickly observe a few things that need to be edited:

- The customer ID is an irrelevant and redundant variable
- The SeniorCitizen flag (1 if a senior, 0 if not) should be a string, not an integer
- Notice the Internet Service column contains 3 options: (DSL, Fiber optic, No internet service). It is then followed by 6 variables on internet specific subscriptions with 3 optional responses (Yes, No, No internet Service). The third response option seems redundant given that a variable already exists “Internet Service” with that detail

Preliminary Data Preparation

This is the workflow for the data preparation which addresses the problems given in the above slide.



- Node 7 transforms the Senior Citizen flag to string.
- Node 4 filters the redundant customer ID column.
- Node 14 removes any rows with missing values (11 were found).
- Nodes 18-23 address the problem of having a redundant third response option “No internet service”, given that a variable already exists that catalogs the type of internet service the customer has (The Internet Service column). The “No internet service” response was replaced by a “No” response. This removes any redundancies from the table and allows for a binary classification of the 6 internet subscription variables (Note that “No” and “No internet service” entail the same meaning). This also eases the creation of dummy variables later, which are needed for logistic regression and MLP modelling.
- Node 24 replaces “No phone service” with “No” in the “MultipleLines” variable for the same reason as mentioned above. It is considered redundant information given that the “PhoneService” variable encodes that information already. The benefits are the same as the point above.

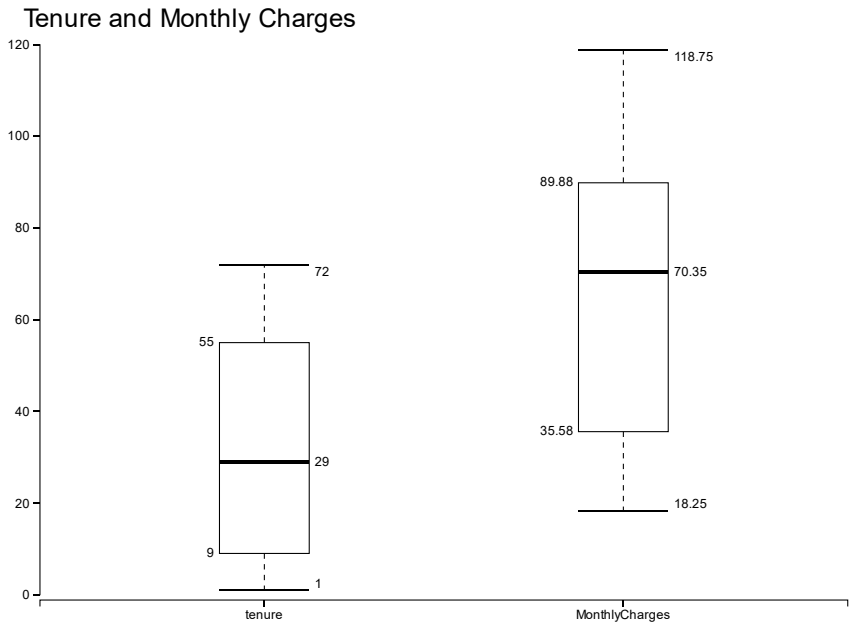
Preliminary Data Preparation

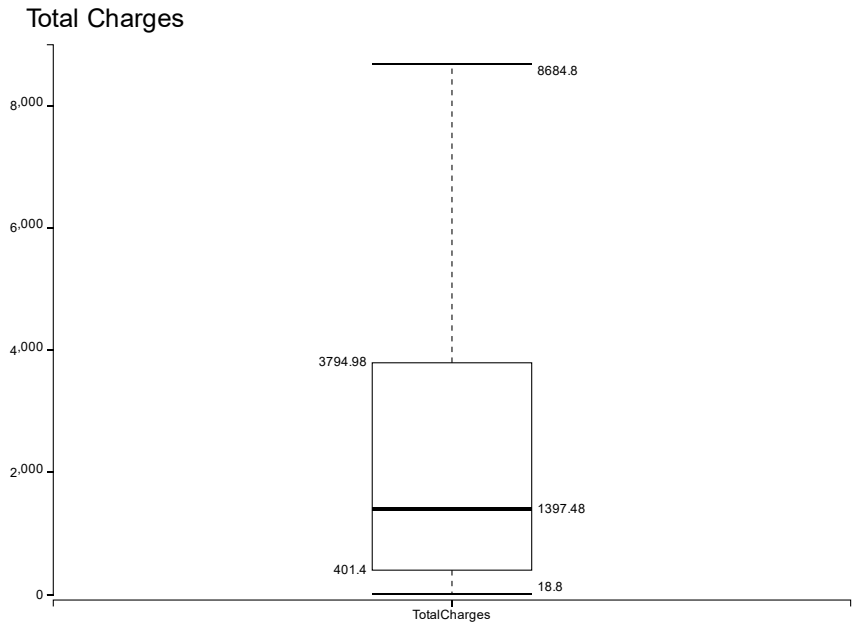
gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	Contract	PaperlessBilling	PaymentMethod	MonthlyCharges	TotalCharges	Online Security	Online Backup	Device Protection	Tech Support	Streaming TV	Streaming Movies	Churn
Female	0	No	No	2	Yes	No	Fiber optic	Month-to-month	Yes	Electronic check	70.7	151.65	No	No	No	No	No	No	Yes
Female	0	No	No	8	Yes	Yes	Fiber optic	Month-to-month	Yes	Electronic check	99.65	820.5	No	No	Yes	No	Yes	Yes	Yes
Male	0	No	Yes	22	Yes	Yes	Fiber optic	Month-to-month	Yes	Credit card (automa...	89.1	1,949.4	No	Yes	No	No	Yes	No	No
Female	0	No	No	10	No	No	DSL	Month-to-month	No	Mailed check	29.75	301.9	Yes	No	No	No	No	No	No
Female	0	Yes	No	28	Yes	Yes	Fiber optic	Month-to-month	Yes	Electronic check	104.8	3,046.05	No	No	Yes	Yes	Yes	Yes	Yes
Male	0	No	Yes	62	Yes	No	DSL	One year	No	Bank transfer (auto...	56.15	3,487.95	Yes	Yes	No	No	No	No	No
Male	0	Yes	Yes	13	Yes	No	DSL	Month-to-month	Yes	Mailed check	49.95	587.45	Yes	No	No	No	No	No	No
Male	0	No	No	16	Yes	No	No	Two year	No	Credit card (automa...	18.95	326.8	No	No	No	No	No	No	No
Male	0	Yes	No	58	Yes	Yes	Fiber optic	One year	No	Credit card (automa...	100.35	5,681.1	No	No	Yes	No	Yes	Yes	No
Male	0	No	No	49	Yes	Yes	Fiber optic	Month-to-month	Yes	Bank transfer (auto...	103.7	5,036.3	No	Yes	Yes	No	Yes	Yes	Yes
Male	0	No	No	25	Yes	No	Fiber optic	Month-to-month	Yes	Electronic check	105.5	2,686.05	Yes	No	Yes	Yes	Yes	Yes	No
Female	0	Yes	Yes	69	Yes	Yes	Fiber optic	Two year	No	Credit card (automa...	113.25	7,895.15	Yes	Yes	Yes	Yes	Yes	Yes	No
Female	0	No	No	52	Yes	No	No	One year	No	Mailed check	20.65	1,022.95	No	No	No	No	No	No	No
Male	0	No	Yes	71	Yes	Yes	Fiber optic	Two year	No	Bank transfer (auto...	106.7	7,382.25	Yes	No	Yes	No	Yes	Yes	No
Female	0	Yes	Yes	10	Yes	No	DSL	Month-to-month	No	Credit card (automa...	55.2	528.35	No	No	Yes	Yes	No	No	Yes
Female	0	No	No	21	Yes	No	Fiber optic	Month-to-month	Yes	Electronic check	90.05	1,862.9	No	Yes	Yes	No	No	Yes	No
Male	1	No	No	1	No	No	DSL	Month-to-month	Yes	Electronic check	39.65	39.65	No	No	Yes	No	No	Yes	Yes
Male	0	Yes	No	12	Yes	No	No	One year	No	Bank transfer (auto...	19.8	202.25	No	No	No	No	No	No	No
Male	0	No	No	1	Yes	No	No	Month-to-month	No	Mailed check	20.15	20.15	No	No	No	No	No	No	Yes
Female	0	Yes	No	58	Yes	Yes	DSL	Two year	Yes	Credit card (automa...	59.9	3,505.1	No	Yes	No	Yes	No	No	No
Male	0	Yes	Yes	49	Yes	No	DSL	Month-to-month	No	Credit card (automa...	59.6	2,970.3	Yes	Yes	No	Yes	No	No	No
Female	0	No	No	30	Yes	No	DSL	Month-to-month	Yes	Bank transfer (auto...	55.3	1,530.6	Yes	Yes	No	No	No	No	No
Male	0	Yes	Yes	47	Yes	Yes	Fiber optic	Month-to-month	Yes	Electronic check	99.35	4,749.15	No	Yes	No	No	Yes	Yes	Yes
Male	0	Yes	Yes	1	No	No	DSL	Month-to-month	No	Electronic check	30.2	30.2	No	Yes	No	No	No	No	Yes
Male	0	Yes	No	72	Yes	Yes	DSL	Two year	Yes	Credit card (automa...	90.25	6,369.45	Yes	Yes	Yes	Yes	Yes	Yes	No
Female	0	No	Yes	17	Yes	No	DSL	Month-to-month	Yes	Mailed check	64.7	1,093.1	No	No	No	No	Yes	Yes	Yes
Female	1	Yes	No	71	Yes	Yes	Fiber optic	Two year	Yes	Credit card (automa...	96.35	6,766.95	Yes	Yes	Yes	Yes	No	No	No
Male	1	Yes	No	2	Yes	No	Fiber optic	Month-to-month	Yes	Credit card (automa...	95.5	181.65	No	No	Yes	No	Yes	Yes	No
Female	0	Yes	Yes	27	Yes	No	DSL	One year	No	Mailed check	66.15	1,874.45	Yes	Yes	Yes	Yes	No	No	No
Male	0	No	No	1	Yes	No	No	Month-to-month	No	Bank transfer (auto...	20.2	20.2	No	No	No	No	No	No	No
Male	1	No	No	1	Yes	No	DSL	Month-to-month	No	Bank transfer (auto...	45.25	45.25	No	No	No	No	No	No	No
Female	0	Yes	Yes	72	Yes	Yes	Fiber optic	Two year	No	Bank transfer (auto...	99.9	7,251.7	Yes	Yes	No	Yes	Yes	No	No
Male	0	No	No	5	Yes	No	Fiber optic	Month-to-month	Yes	Electronic check	69.7	316.9	No	No	No	No	No	No	Yes
Female	0	No	No	46	Yes	No	Fiber optic	Month-to-month	Yes	Credit card (automa...	74.8	3,548.3	No	No	Yes	No	No	No	No
Male	0	No	No	34	Yes	Yes	Fiber optic	Month-to-month	Yes	Electronic check	106.35	3,549.25	No	Yes	Yes	No	Yes	Yes	Yes
Female	0	No	No	11	Yes	Yes	Fiber optic	Month-to-month	Yes	Bank transfer (auto...	97.85	1,105.4	No	No	Yes	No	Yes	Yes	Yes
Male	0	Yes	Yes	10	Yes	No	DSL	One year	No	Mailed check	49.55	475.7	No	Yes	No	No	No	No	No
Female	0	Yes	Yes	70	Yes	Yes	DSL	Two year	Yes	Credit card (automa...	69.2	4,872.35	Yes	Yes	No	No	Yes	No	No

The table above shows the data after being prepared for descriptive analysis and predictive modelling

Univariate Analysis

We start by looking at the distribution of the only quantitative variables in the dataset, which are tenure (amount of time the customer has been with the company), monthly charges, and Total charges using box plots. As seen below, the distributions of monthly charges and tenure are somewhat symmetric. The total charges boxplot is skewed to the right, however no outliers are detected in any variable





This is reaffirmed by the statistics for each variable. The mean and median for tenure and monthly charges are relatively close, but with the Total charges the difference is significant

Column	Min	Mean	Median	Max	Std. Dev.	Skewness	Kurtosis	No. Missing	No. +∞	No. -∞	Histogram
tenure	1	32.4218	29	72	24.5453	0.2377	-1.3878	0	0	0	0 72
MonthlyCharges	18.25	64.7982	70.35	118.75	30.086	-0.2221	-1.2562	0	0	0	15 120
TotalCharges	18.8	2,283.3004	1,397.475	8,684.8	2,266.7714	0.9616	-0.2318	0	0	0	0 8,750

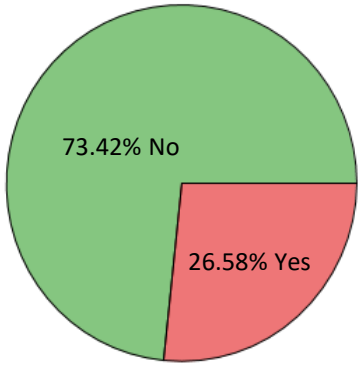
Univariate Analysis

Categorical distributions using pie charts

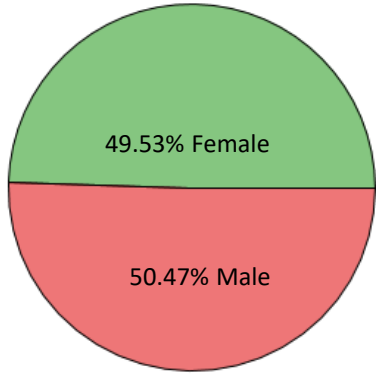
Pie charts are used to determine the relative distribution of the variables.

Note that churners only account for 26.58% of the observations which is somewhat unbalanced when used for modeling and prediction. This will be discussed later.

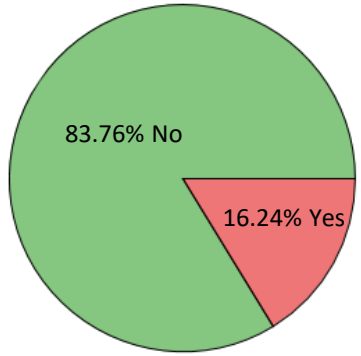
This is just to provide a general idea on how the variables are distributed.



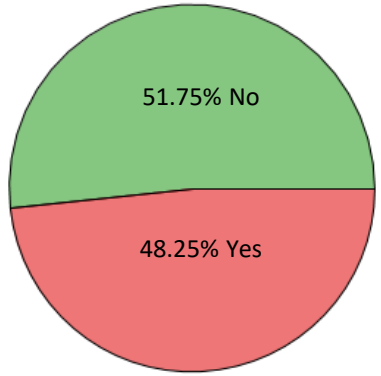
Churners



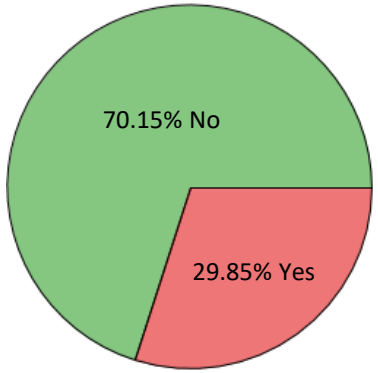
Gender



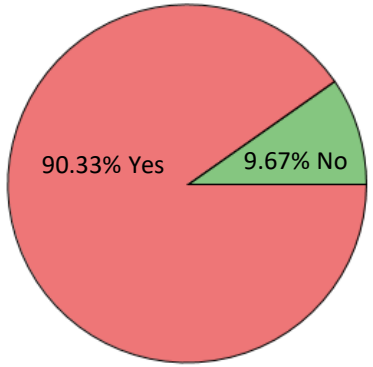
Senior Citizen



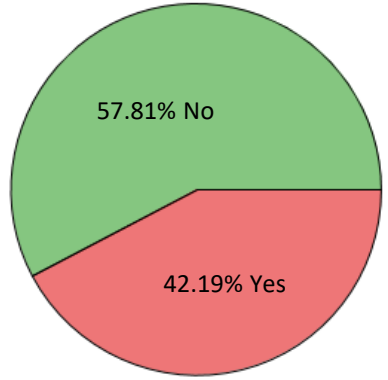
Partner



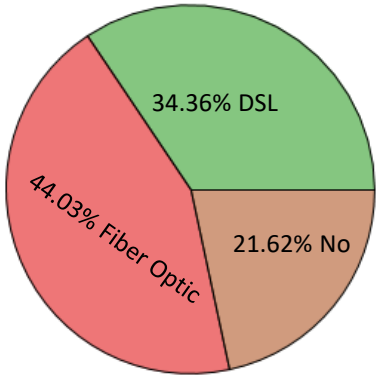
Dependents



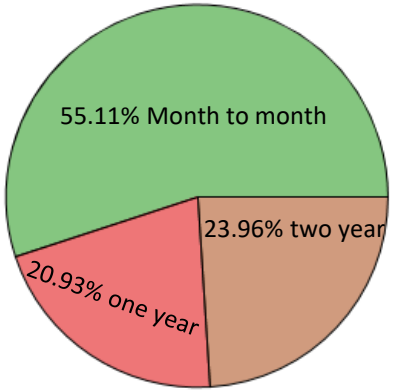
Phone Service



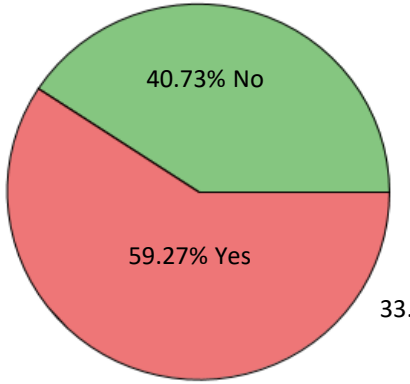
Multiple Lines



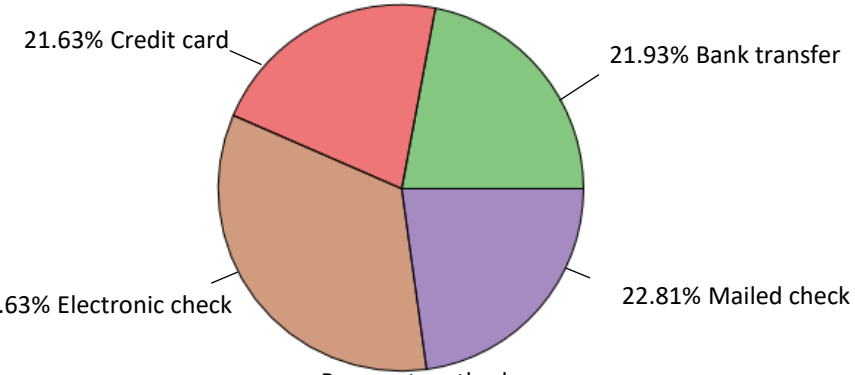
Internet Service



Contract



Paperless billing



Payment methods

Bivariate Analysis

An exploratory analysis of some of the feature variables with each other and with the target variable was made next to better understand some of the relationships before building up predictive models.

For analyzing categorical variables with each other, a crosstabulation is made to study if the relationships are significant. This is however not enough given that the large dataset (7032 rows) would result in small p-values for variables that may not be that important. That is why the Cramer's V score is then used to assess the strength of the relationship.

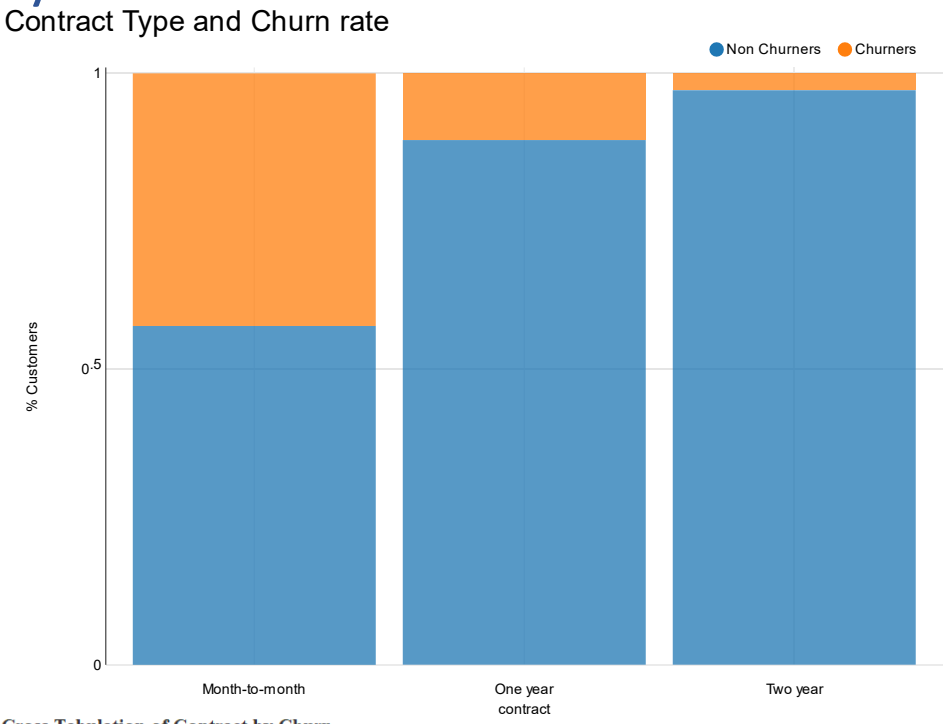
First is crosstab analysis of some of the features with the target (churn)

Contract Type and Churn rate

The contract type seems to easily indicate how churners are distributed. On a two-year contract, most customers are loyal with only a 3% churn rate. This increases when moving down to a one-year contract, with the rate increasing to 11%. Finally, we see a very large increase in churn rate to 43% when moving to a month-to-month contract. Most churners are concentrated in this contract type with almost 89% being in that group.

This significant relationship is backed by a Cramer's V score of 0.29 which indicates a high strength of the relationship

The following stacked bar chart gives a visual understanding of how churners are distributed



Cross Tabulation of Contract by Churn

Frequency Row Percent Column Percent	No	Yes	Total
Month-to-month	2,220	1,655	3,875
	57.2903%	42.7097%	
	42.9983%	88.55%	
One year	1,306	166	1,472
	88.7228%	11.2772%	
	25.2954%	8.8818%	
Two year	1,637	48	1,685
	97.1513%	2.8487%	
	31.7064%	2.5682%	
Total	5,163	1,869	7,032

☑ Frequency ☐ Expected ☐ Deviation ☐ Percent ☑ Row Percent ☑ Column Percent ☐ Cell Chi-Square

Max rows: 10

Max columns: 10

Statistics for Table of Contract by Churn

Statistic	DF	Value	Prob
Chi-Square	2	1,179.5458	7.33E-257

Total sample size: 7032.0

Bivariate Analysis

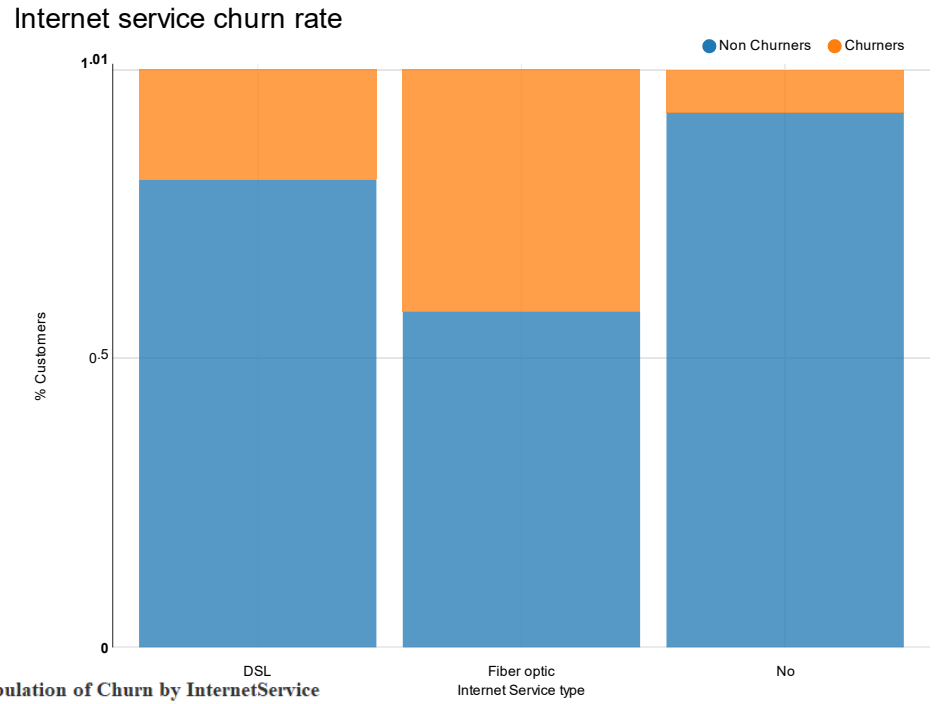
Internet Service and Churn Rate

Churn rate and internet service type are also tied significantly. We see that the DSL service is mostly filled with non churners with almost 81% being loyal customers.

For Fiber optic, that loyalty decreases to 58%.

Surprisingly however, if the customer does not have an internet subscription, they are very likely to be non churners with almost 93% of people without an internet connection being loyal. This is not very intuitive given that the general assumption is that the more services one uses the more loyal they become, however this is not the case here

The Cramer's V which is 0.228 is also relatively high. All this information is visualized with the stacked bar chart.



Cross Tabulation of Churn by InternetService

Frequency Row Percent Column Percent	DSL	Fiber optic	No	Total
No	1,957	1,799	1,407	5,163
	37.9043%	34.8441%	27.2516%	
	81.0017%	58.1072%	92.5658%	
Yes	459	1,297	113	1,869
	24.5586%	69.3954%	6.046%	
	18.9983%	41.8928%	7.4342%	
Total	2,416	3,096	1,520	7,032

☑ Frequency
☐ Expected
☐ Deviation
☐ Percent
☑ Row Percent
☑ Column Percent
☐ Cell Chi-Square

Max rows: 10

Max columns: 10

Statistics for Table of Churn by InternetService

Statistic	DF	Value	Prob
Chi-Square	2	728.6956	5.83E-159

Total sample size: 7032.0

Bivariate Analysis

Payment Method and Churn Rate

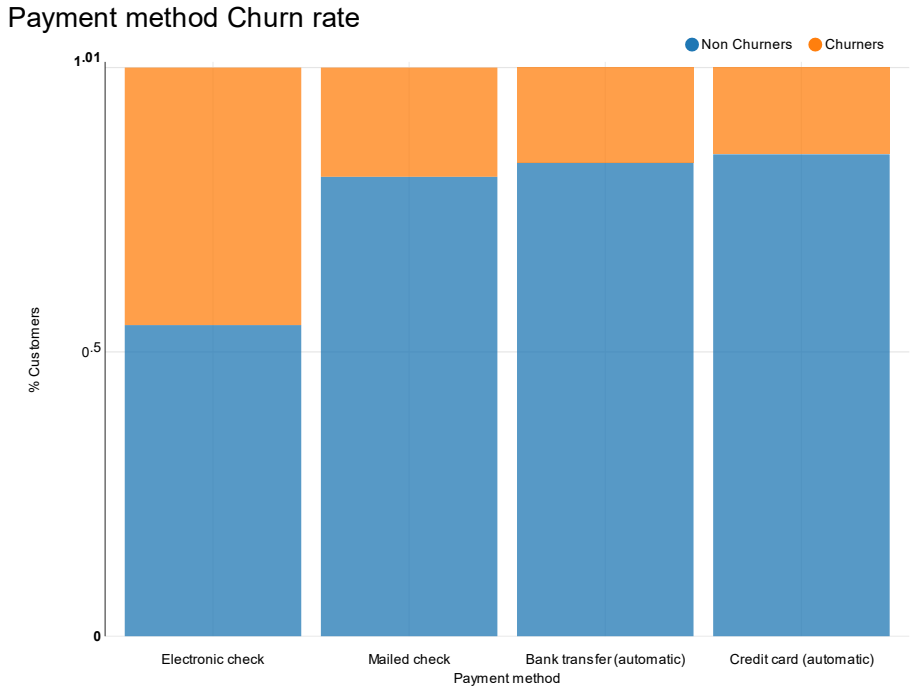
The last significant categorical feature variable we will cover in depth is the payment method used by the customer.

We can easily see from the table that the non-churners are almost uniformly distributed across the four types of payment methods used.

For the churners however, 57% of them are using electronic checks, with the rest being distributed evenly on the 3 other payment methods. Electronic payment is also the one with the most churners as a percentage of total customers using that method, accounting for 45%.

The Cramer's V has a medium/high strength of 0.214

The next page shows off the stacked bar chart for visualizing the distribution



Cross Tabulation of Churn by PaymentMethod

Frequency Row Percent Column Percent	Bank transfer (automatic)	Credit card (automatic)	Electronic check	Mailed check	Total
No	1,284	1,289	1,294	1,296	5,163
	24.8693%	24.9661%	25.0629%	25.1017%	
	83.2685%	84.7469%	54.7146%	80.798%	
Yes	258	232	1,071	308	1,869
	13.8042%	12.4131%	57.3034%	16.4794%	
	16.7315%	15.2531%	45.2854%	19.202%	
Total	1,542	1,521	2,365	1,604	7,032

Frequency
Expected
Deviation
Percent
Row Percent
Column Percent
Cell Chi-Square

Max rows: 10

Max columns: 10

Statistics for Table of Churn by PaymentMethod

Statistic	DF	Value	Prob
Chi-Square	3	645.4299	1.43E-139

Total sample size: 7032.0

Bivariate Analysis

Some of the insignificant categorical variables when cross-tabulated with churn included Phone service and Gender, which had a Cramer's V score of 0.006 and 0.008 respectively.

There are a lot more categorical variables that could be cross-tabulated with each other and analyzed, however that is too impractical given the large number of variables in the dataset, so we will move on to other analysis types.

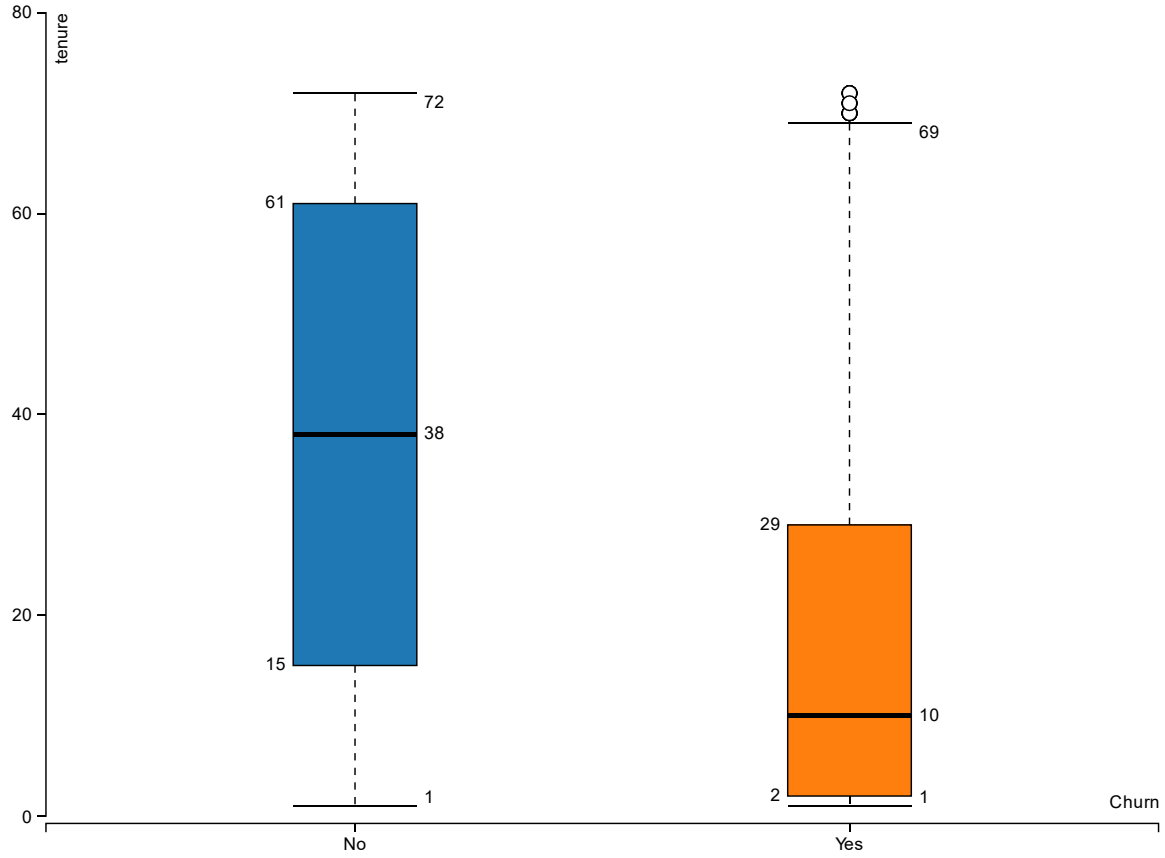
We will next analyze the quantitative variables in the dataset (total charges, monthly charges, tenure) against churn using a conditional box plot to see if any visual patterns are noticed.

Tenure and Churn Rate

As we can see, churners are more likely to have low tenure with the company than non churners. The median for churners is 10 months, while for non churners it is 38 months.

Non churners have a symmetric distribution about the tenure, but churners have a right tail indicating that most churn during the first months.

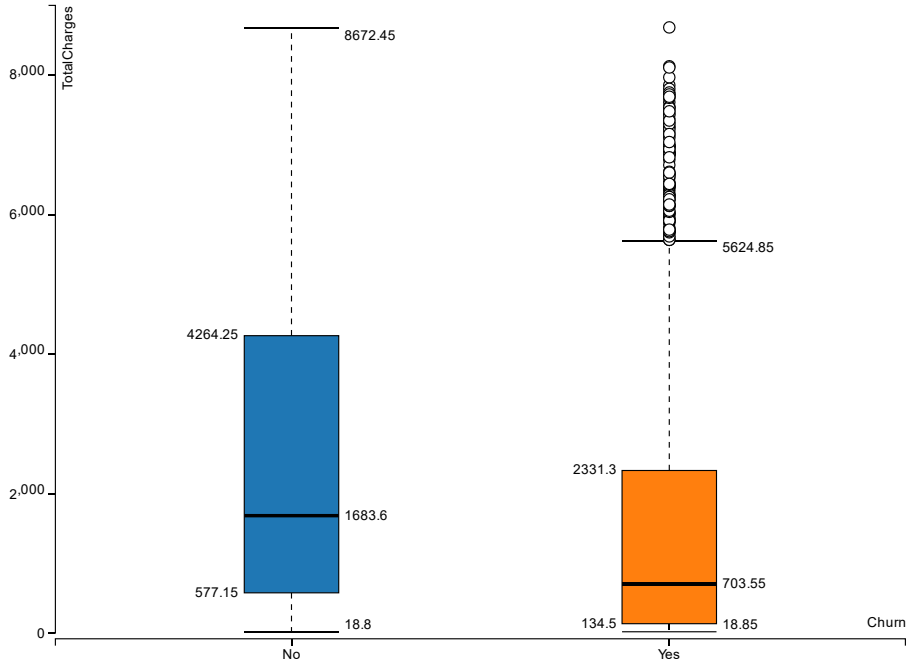
This is very intuitive given that customers generally are less likely to leave a company that they are familiar with, while new customers might still be hesitant on choosing a specific company are thus more volatile.



Bivariate Analysis

Total charges and Churn Rate

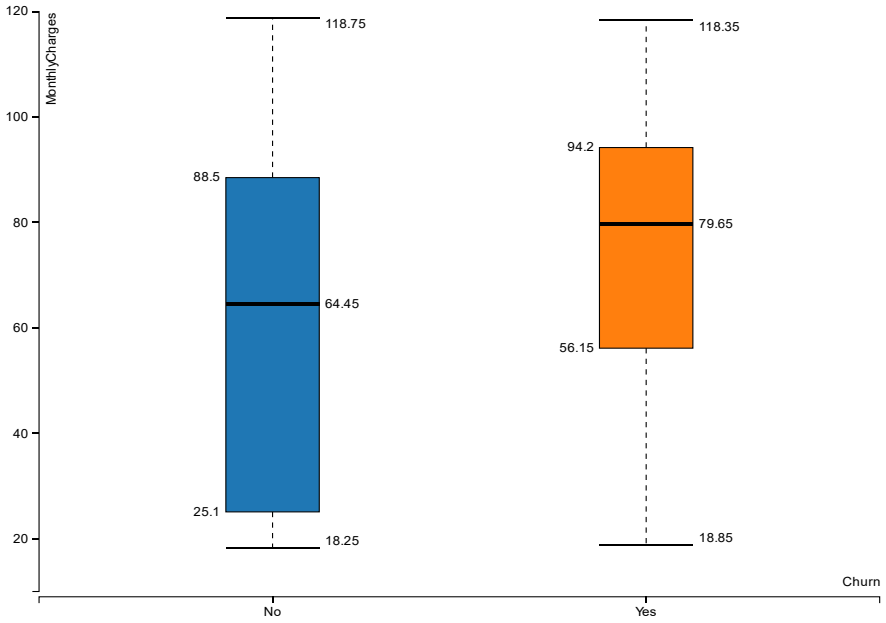
It is also seen that churners are more likely to have fewer total charges than non churners. This is expected given that the total time with the company (tenure) strongly determines churn rate, and the more time a customer spends with the company, the additional charges they incur.
Tenure and Total charges are heavily correlated as will be seen later



Monthly charges and Churn Rate

Monthly charges seem to have a less apparent effect on the churn rate distribution than tenure and total charges. It seems however, that unlike total charges, monthly charges of none churners are less than that of churners.

It can be hypothesized that this is because of the pricing scheme of the company which increases charges on month-to-month contracts (where most churners are grouped) while decreasing monthly charges on a yearly contract type.



Bivariate Analysis

One-way ANOVA tables complement the box plots, with all three quantitative variables being significant with respect to churn. The group tables are given below. Note that all the p-values for the ANOVA analysis are 0

	Group	N	Missing	Missing Group	Mean	Std. Deviation	Std. Error	CI (Lower Bound)	CI (Upper Bound)	Minimum	Maximum
MonthlyCharges	No	5163	0	0	61.3074	31.0946	0.4327	60.459	62.1558	18.25	118.75
MonthlyCharges	Yes	1869	0	0	74.4413	24.6661	0.5706	73.3223	75.5603	18.85	118.35
MonthlyCharges	Total	7032	0	0	64.7982	30.086	0.3588	64.0949	65.5015	18.25	118.75

	Group	N	Missing	Missing Group	Mean	Std. Deviation	Std. Error	CI (Lower Bound)	CI (Upper Bound)	Minimum	Maximum
TotalCharges	No	5163	0	0	2,555.3441	2,329.457	32.4193	2,491.7886	2,618.8997	18.8	8,672.45
TotalCharges	Yes	1869	0	0	1,531.7961	1,890.823	43.7367	1,446.0181	1,617.5741	18.85	8,684.8
TotalCharges	Total	7032	0	0	2,283.3004	2,266.7714	27.0314	2,230.3108	2,336.2901	18.8	8,684.8

	Group	N	Missing	Missing Group	Mean	Std. Deviation	Std. Error	CI (Lower Bound)	CI (Upper Bound)	Minimum	Maximum
tenure	No	5163	0	0	37.65	24.0769	0.3351	36.9931	38.3069	1	72
tenure	Yes	1869	0	0	17.9791	19.5311	0.4518	17.0931	18.8652	1	72
tenure	Total	7032	0	0	32.4218	24.5453	0.2927	31.848	32.9956	1	72

The difference in Means for churners and non-churners is substantial in the tenure and Total charges variables, but is less pronounced in the Monthly Charges (however they are all significant)

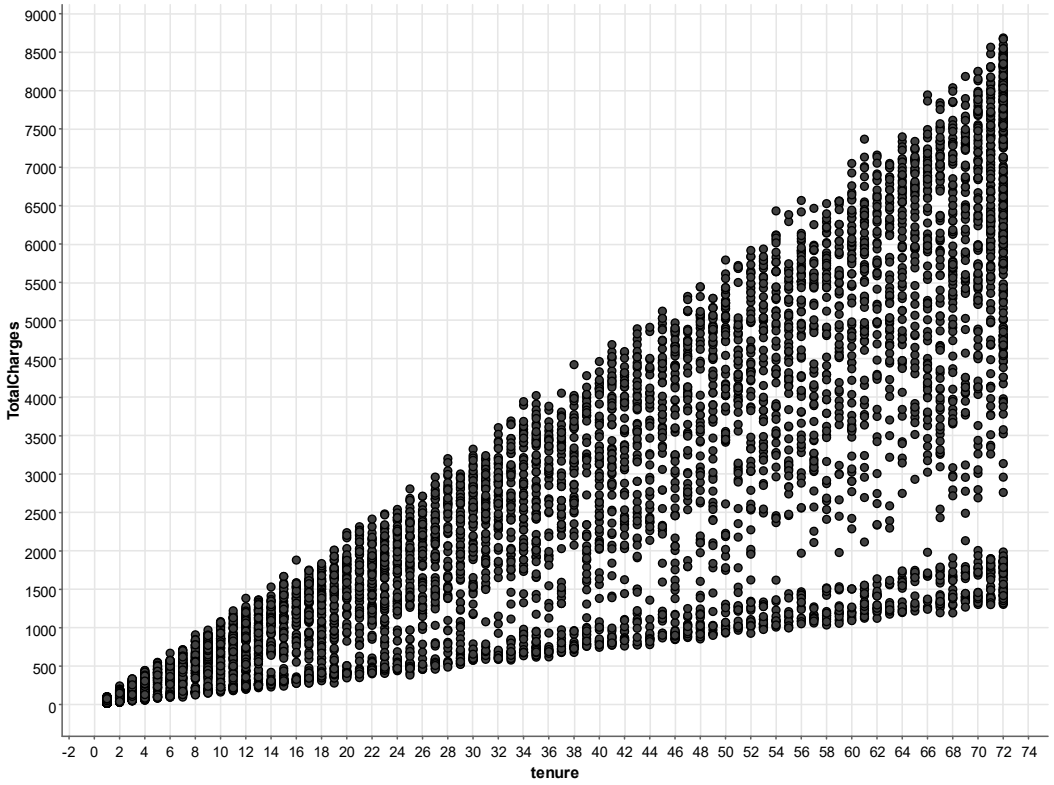
Bivariate Analysis

Lastly, we look at the relationships of quantitative variables with one another.

Tenure and Total charges are heavily related. The correlation coefficient for a linear relationship model is 0.83, which is very high. This might be a result of collinearity between both variables given that they are intuitively linked. (The longer the person is a customer the more they pay).

The relationship between Total charges, monthly charges, and tenure is assumed to be: Total charges = Monthly charges * tenure. We can test that assumption by creating a new variable (tenure * monthly charges) and seeing if it correlates with Total charges.

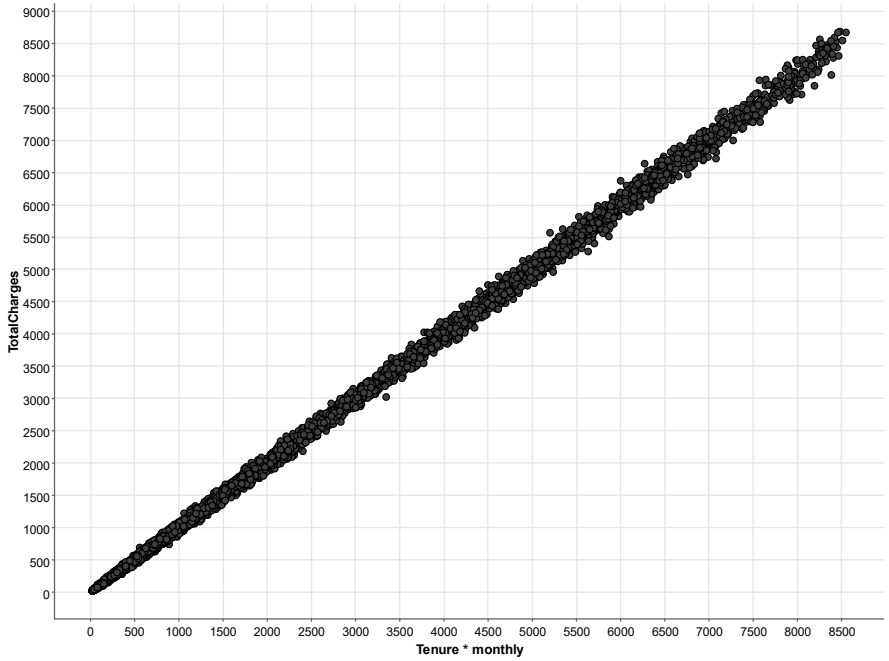
The results are given on the next page



S First col...	S Second...	D Correlation value	D p value	I Degree...
tenure	TotalCharges	0.825880460933204	0.0	7030

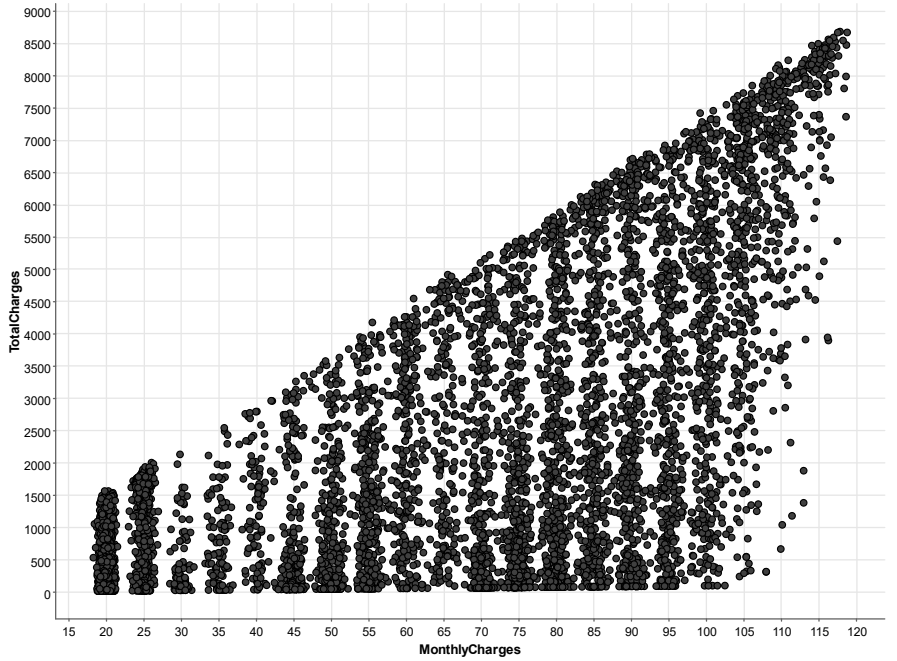
Bivariate Analysis

The total charges variable is almost exactly equal to the new variable created. This is represented by an almost perfect straight line and a correlation coefficient that rounds up to 1. This can have unfavorable effects on the model accuracy used for prediction, as well as being inefficient due to there existing redundant variables that do not add any explanation towards classifying the target variable. This will be addressed during feature selection



S First col...	S Second col...	D Correlation value	D p value	I Degree...
TotalCharges	Tenure * monthly	0.9995598572867...	0.0	7030

The relationship between monthly charges and total charges is also relatively strong. The linear correlation coefficient is 0.65 which suggests that monthly charges do play a role in determining total charges, with the rest of the explained variance coming from tenure.



S First colu...	S Second...	D Correlation value	D p value	I Degree...
MonthlyCharges	TotalCharges	0.6510648032262...	0.0	7030

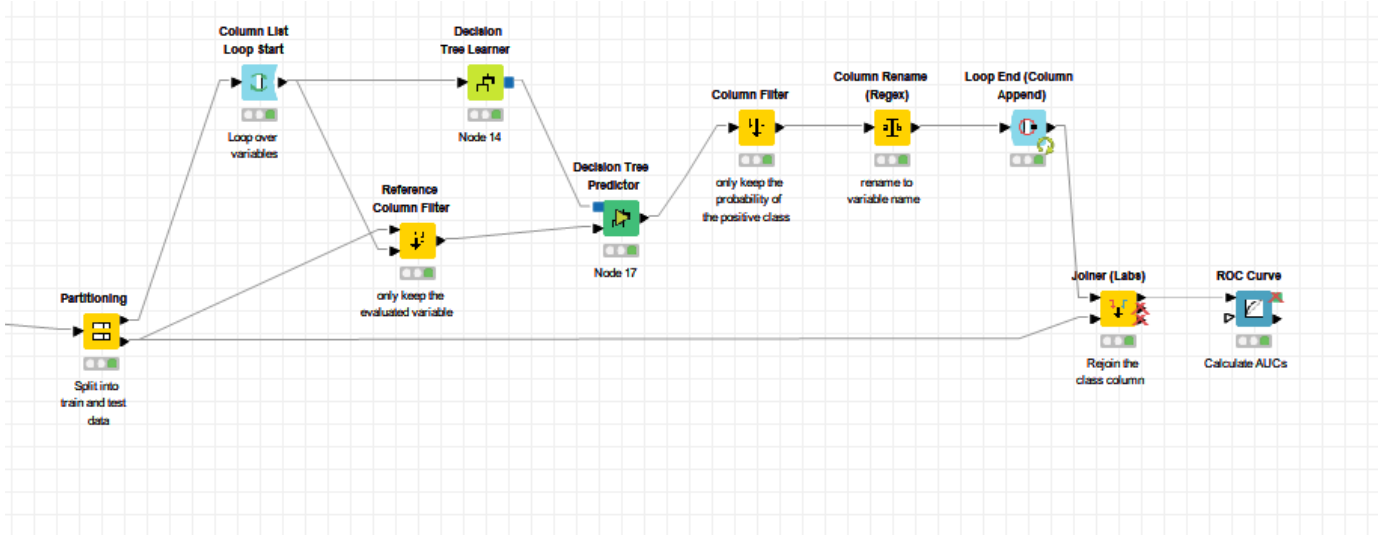
Explanatory Variable Impact Evaluation

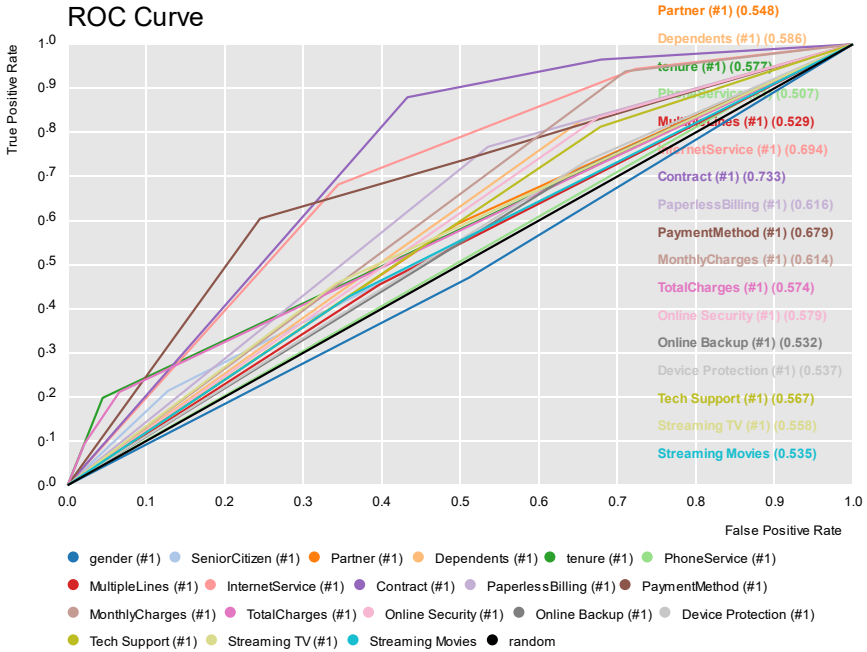
We move on to a general summary of how the feature variables impact the ability of a simple decision tree to predict the customers who churn. This is done through a loop in Knime which iteratively swaps the variables to be used for the decision tree learner one at a time separately. This is done to quickly understand the predicting power at hand if only one variable was to be analyzed.

The impact of the variables is measured by the area under the ROC curve, which is constructed for each one-variable model in the loop.

The table shows the AUC for each variable. Contract, payment method and Internet service have high AUCs, and thus can be used to predict the target churn variable very well, while when Phone service, gender, and Multiple lines are used by the decision tree, they are as good as random guessing. This helps map out the relative importance of these variables as we go into data modelling.

All of this is in line with the ANOVA and cross tabulation analysis findings discussed earlier.

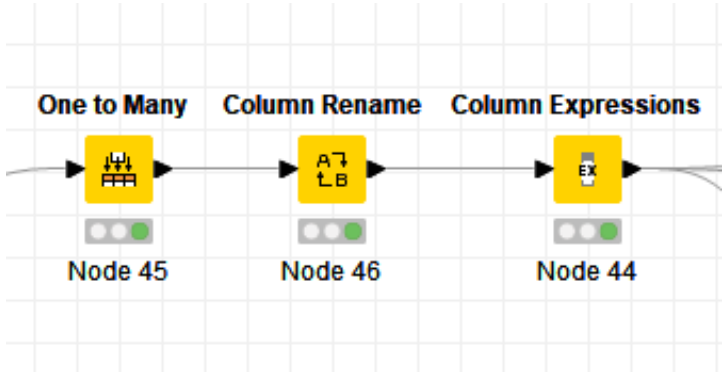




Row ID	Area Und...
Contract (#1)	0.733
InternetService (#1)	0.694
PaymentMethod (#1)	0.679
PaperlessBilling (#1)	0.616
MonthlyCharges (#1)	0.614
Dependents (#1)	0.586
Online Security (#1)	0.579
tenure (#1)	0.577
TotalCharges (#1)	0.574
Tech Support (#1)	0.567
Streaming TV (#1)	0.558
Partner (#1)	0.548
SeniorCitizen (#1)	0.543
Device Protection (#1)	0.537
Streaming Movies	0.535
Online Backup (#1)	0.532
MultipleLines (#1)	0.529
PhoneService (#1)	0.507
gender (#1)	0.48

Data Preparation for MLP and Logistic Regression

The last data preparation step needed before modeling is creating flags for all the non-binary categorical variables (Internet Service, Contract, Payment Method) which are interpreted as 1 or 0 integers. This is needed for logistic regression and multi layer perceptron modelling as they cannot weigh categorical string variables without one hot encoding first. A section of the output table after preparation is shown below



gen...	Senior...	Part...	Dependents	tenure	PhoneS...	Multi...	Pap...	MonthlyCharges	TotalCharges	Churn	Online ...	Onli...	Devic...	Tech Support	Streami...	Streami...	DSL	Fiber optic	No intern...	Month-to-month	One year	Two year	Electroni...	Maile...	Bank tra...	Credit card ...
0	0	1	1	1	0	0	1	29.85	29.85	No	0	1	0	0	0	0	1	0	0	1	0	0	1	0	0	0
1	0	0	0	34	1	0	0	56.95	1,889.5	No	1	0	1	0	0	0	1	0	0	0	1	0	0	1	0	0
1	0	0	0	2	1	0	1	53.85	108.15	Yes	1	1	0	0	0	0	1	0	0	1	0	0	0	1	0	0
1	0	0	0	45	0	0	0	42.3	1,840.75	No	1	0	1	1	0	0	1	0	0	0	1	0	0	0	1	0
0	0	0	0	2	1	0	1	70.7	151.65	Yes	0	0	0	0	0	0	0	1	0	1	0	0	1	0	0	0
0	0	0	0	8	1	1	1	99.65	820.5	Yes	0	0	1	0	1	1	0	1	0	1	0	0	1	0	0	0
1	0	0	0	22	1	1	1	89.1	1,949.4	No	0	1	0	0	1	0	0	1	0	1	0	0	0	0	0	1
0	0	0	0	10	0	0	0	29.75	301.9	No	1	0	0	0	0	0	1	0	0	1	0	0	0	1	0	0
0	0	1	1	28	1	1	1	104.8	3,046.05	Yes	0	0	1	1	1	1	0	1	0	1	0	0	1	0	0	0
1	0	0	0	62	1	0	0	56.15	3,487.95	No	1	1	0	0	0	0	1	0	0	0	1	0	0	0	1	0
1	0	1	1	13	1	0	1	49.95	587.45	No	1	0	0	0	0	0	1	0	0	1	0	0	0	1	0	0
1	0	0	0	16	1	0	0	18.95	326.8	No	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	1
1	0	1	1	58	1	1	0	100.35	5,681.1	No	0	0	1	0	1	1	0	1	0	0	1	0	0	0	0	1
1	0	0	0	49	1	1	1	103.7	5,036.3	Yes	0	1	1	0	1	1	0	1	0	1	0	0	0	0	1	0
1	0	0	0	25	1	0	1	105.5	2,686.05	No	1	0	1	1	1	1	0	1	0	1	0	0	1	0	0	0
0	0	1	1	69	1	1	0	113.25	7,895.15	No	1	1	1	1	1	1	0	1	0	0	0	1	0	0	0	1
0	0	0	0	52	1	0	0	20.65	1,022.95	No	0	0	0	0	0	0	0	0	1	0	1	0	0	1	0	0
1	0	0	0	71	1	1	0	106.7	7,382.25	No	1	0	1	0	1	1	0	1	0	0	0	1	0	0	1	0
0	0	1	1	10	1	0	0	55.2	528.35	Yes	0	0	1	1	0	0	1	0	0	1	0	0	0	0	0	1
0	0	0	0	21	1	0	1	90.05	1,862.9	No	0	1	1	0	0	1	0	1	0	1	0	0	1	0	0	0
1	1	0	0	1	0	0	1	39.65	39.65	Yes	0	0	1	0	0	1	1	0	0	1	0	0	1	0	0	0
1	0	1	1	12	1	0	0	19.8	202.25	No	0	0	0	0	0	0	0	0	1	0	1	0	0	0	1	0
1	0	0	0	1	1	0	0	20.15	20.15	Yes	0	0	0	0	0	0	0	0	1	1	0	0	0	1	0	0
0	0	1	1	58	1	1	1	59.9	3,505.1	No	0	1	0	1	0	0	1	0	0	0	0	1	0	0	0	1
1	0	1	1	49	1	0	0	59.6	2,970.3	No	1	1	0	1	0	0	1	0	0	1	0	0	0	0	0	1
0	0	0	0	30	1	0	1	55.3	1,530.6	No	1	1	0	0	0	0	1	0	0	1	0	0	0	0	1	0
1	0	1	1	47	1	1	1	99.35	4,749.15	Yes	0	1	0	0	1	1	0	1	0	1	0	0	1	0	0	0
1	0	1	1	1	0	0	0	30.2	30.2	Yes	0	1	0	0	0	0	1	0	0	1	0	0	1	0	0	0
1	0	1	1	72	1	1	1	90.25	6,369.45	No	1	1	1	1	1	1	1	0	0	0	0	1	0	0	0	1
0	0	0	0	17	1	0	1	64.7	1,093.1	Yes	0	0	0	0	1	1	1	0	0	1	0	0	0	1	0	0
0	1	1	1	71	1	1	1	96.35	6,766.95	No	1	1	1	1	0	0	0	1	0	0	0	1	0	0	0	1
1	1	1	1	2	1	0	1	95.5	181.65	No	0	0	1	0	1	1	0	1	0	1	0	0	0	0	0	1
0	0	1	1	27	1	0	0	66.15	1,874.45	No	1	1	1	1	0	0	1	0	0	0	1	0	0	1	0	0
1	0	0	0	1	1	0	0	20.2	20.2	No	0	0	0	0	0	0	0	0	1	1	0	0	0	0	1	0
1	1	0	0	1	1	0	0	45.25	45.25	No	0	0	0	0	0	0	1	0	0	1	0	0	0	0	1	0
0	0	1	1	72	1	1	0	99.9	7,251.7	No	1	1	0	1	1	0	0	1	0	0	0	1	0	0	1	0
1	0	0	0	5	1	0	1	69.7	316.9	Yes	0	0	0	0	0	0	0	1	0	1	0	0	1	0	0	0
0	0	0	0	46	1	0	1	74.8	3,548.3	No	0	0	1	0	0	0	0	1	0	1	0	0	0	0	0	1
1	0	0	0	34	1	1	1	106.35	3,549.25	Yes	0	1	1	0	1	1	0	1	0	1	0	0	1	0	0	0

Feature Selection

To better understand which features are redundant and which are not we perform a basic feature selection.

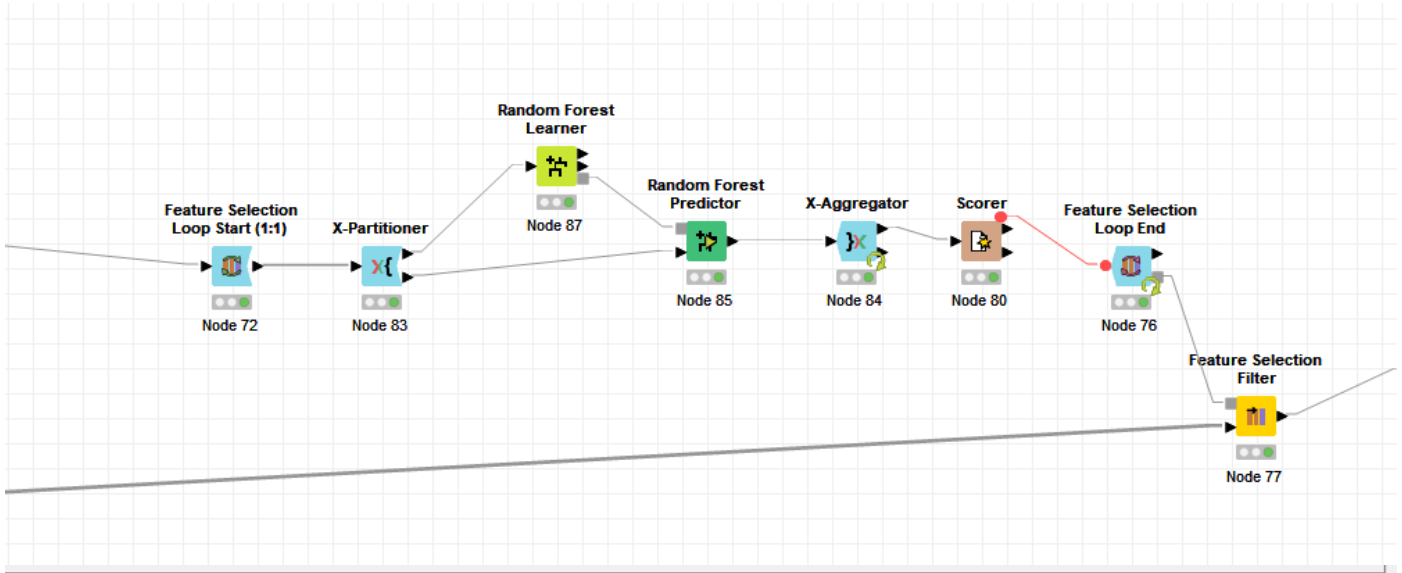
A basic backward feature selection loop is used in Knime to measure the impact of removing certain features. The model starts off with all the features, The cross-validation accuracy score is measured, and then the loop iteratively removes a variable from the model and keeps measuring the accuracy score. A Random Forest Learner is used to predict the target in the loop.

Looking at the table we see that decreasing the number of variables in the dataset increases accuracy, even though not significantly.

The 14 feature variables (along with churn) in the table are selected for further predictive modelling.

The benefits are efficiency in modelling at the least (with little to no impact on performance), and a potential for increased performance at the most.

If managerial input dictates the need to analyze certain other variables that are not included in this selection, we can easily choose another efficient allocation and model with those variables instead.



Accuracy	Nr. of features
0.803	15
0.802	12
0.802	9
0.802	11
0.801	18
0.801	13
0.8	10
0.8	17
0.8	16
0.8	14
0.799	19
0.799	20
0.799	21
0.798	24
0.798	22
0.797	23
0.796	7
0.796	8
0.796	25
0.796	6
0.794	26
0.791	5
0.791	4
0.788	3
0.787	2
0.734	1

I	gender
I	SeniorCitizen
I	Partner
I	Dependents
I	tenure
I	PhoneService
I	MultipleLines
I	PaperlessBilling
D	MonthlyCharges
D	TotalCharges
S	Churn
I	Online Security
I	Online Backup
I	Device Protection
I	Tech Support
I	Streaming TV
I	Streaming Movies
I	DSL
I	Fiber optic
I	No internet service
I	Month-to-month
I	One year
I	Two year
I	Electronic check
I	Mailed check
I	Bank transfer (automatic)
I	Credit card (automatic)

Modelling

Costs of Misclassification

This project is relevant to the telecommunications industry in the sense that they need to be able to better classify the potential churners by either targeting them more effectively by providing them better contracts or services, or just analyzing the market shifts they operate in. Either way, a cost is associated to both the false positive and false negative predictions in a model.

In our situation, a False positive entails classifying a non churner as a churner. The cost associated would be the unneeded provision of resources to target the customer try to make them stay, even though they had no intention of leaving.

A false negative however, entails classifying a churner as a non churner. The cost associated is the potential loss of future business from that customer that might have changed their mind if the company targeted them better if they knew they were potential churners.

The False Positive Rate (1- Specificity) will be used to measure the first case, while the False Negative Rate (1 - Sensitivity) will be used to measure the latter case.

Given that we have no information as to which misclassification costs are more severe on a financial level to the company, it is assumed that both are equally undesirable.

That is why a model that maximizes both sensitivity and specificity will be sought after. The cut off point used for the ROC curves for the model assessments in this case is the Max Youden's index as it provides a good balance between both measures.

The models will also be evaluated based on the Area under the ROC curves. Due to the imbalance of the target variable, the accuracy measures will be considered after the 3 previous measures. Precision is also compared (at the Max Youden's index) in order to see how confident the models are at assessing positive predictions.

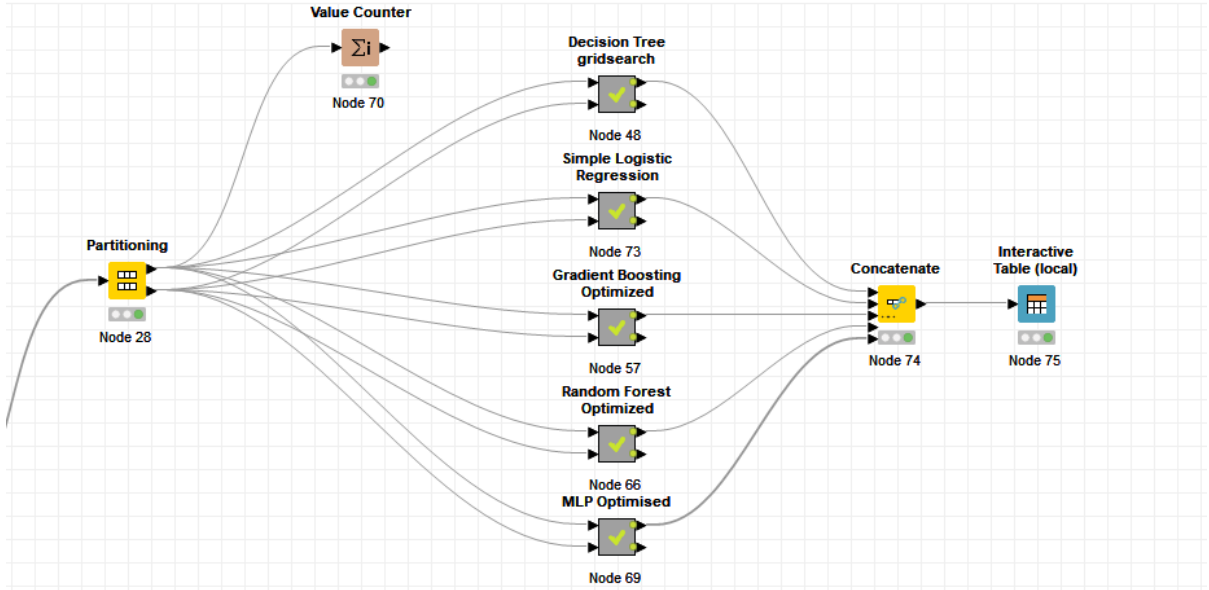
Modelling

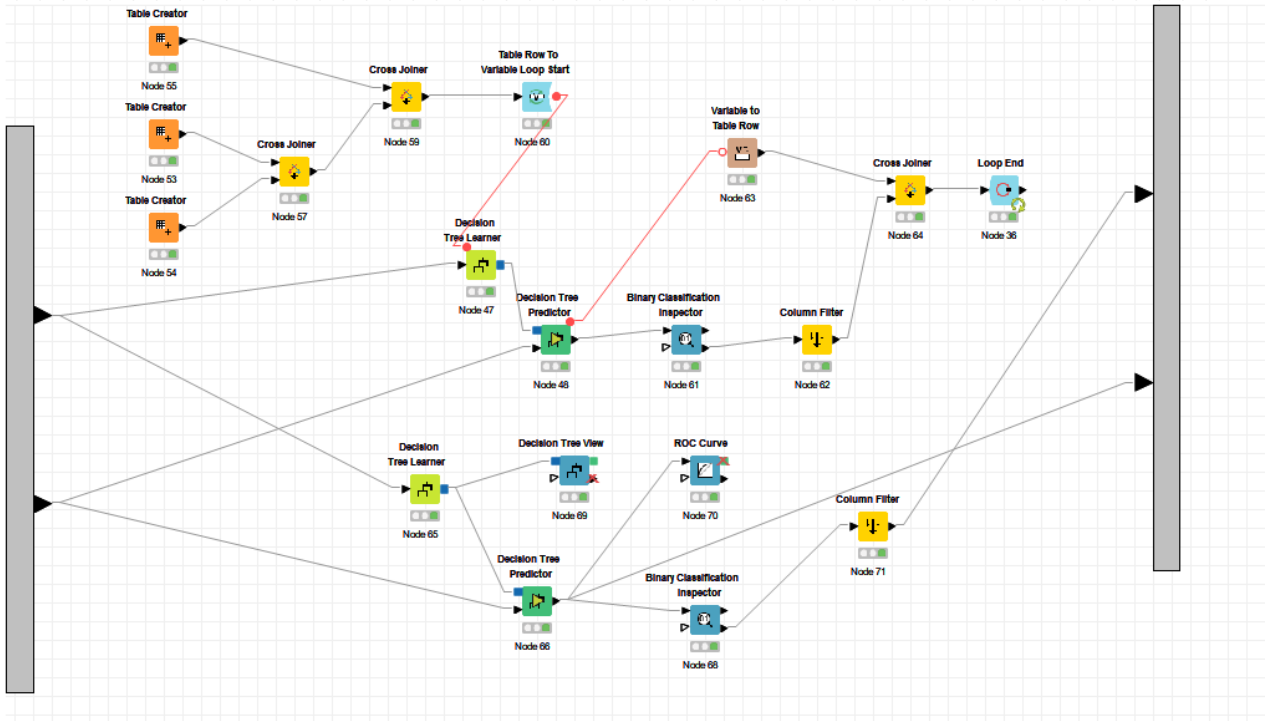
The models used are Decision Trees, Logistic Regression, Gradient Boosting Trees, Random Forests and a Multi Layer Perceptron.

Each model will be discussed in detail followed by a final comparison between all of them in the end.

The train test split used was 80%/20% due to the large number of observations, 20% would be considered enough to test on.

Also note that no balancing of the training set was made as it was found that balancing the training set before modelling resulted in worse, or at most unchanged performance across all models (Balancing was done using both SMOTE and Equal sized sampling techniques but yielded worse results)





Decision Trees

Decision trees are used for a very fast model fit and prediction as well as very easy interpretability of the features, which is perfect when intuitive decision making must be backed up by easily readable evidence.

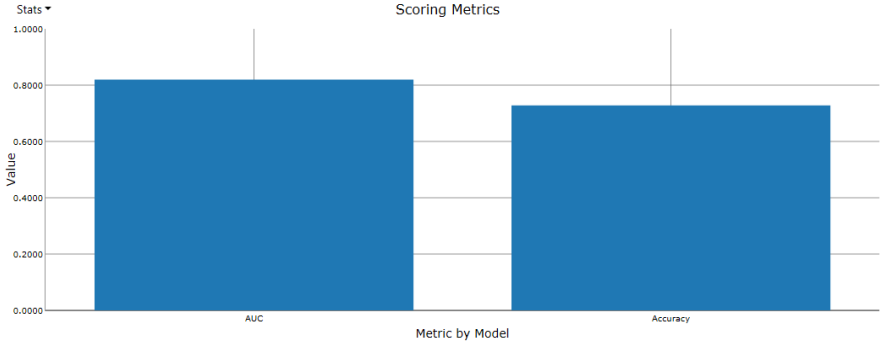
A grid search is used for parameter optimization. The split quality measure, pruning method, and minimum number records per node are the parameters tuned.

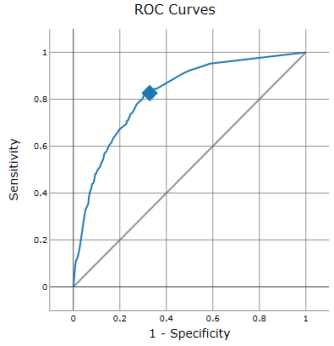
Decision Trees

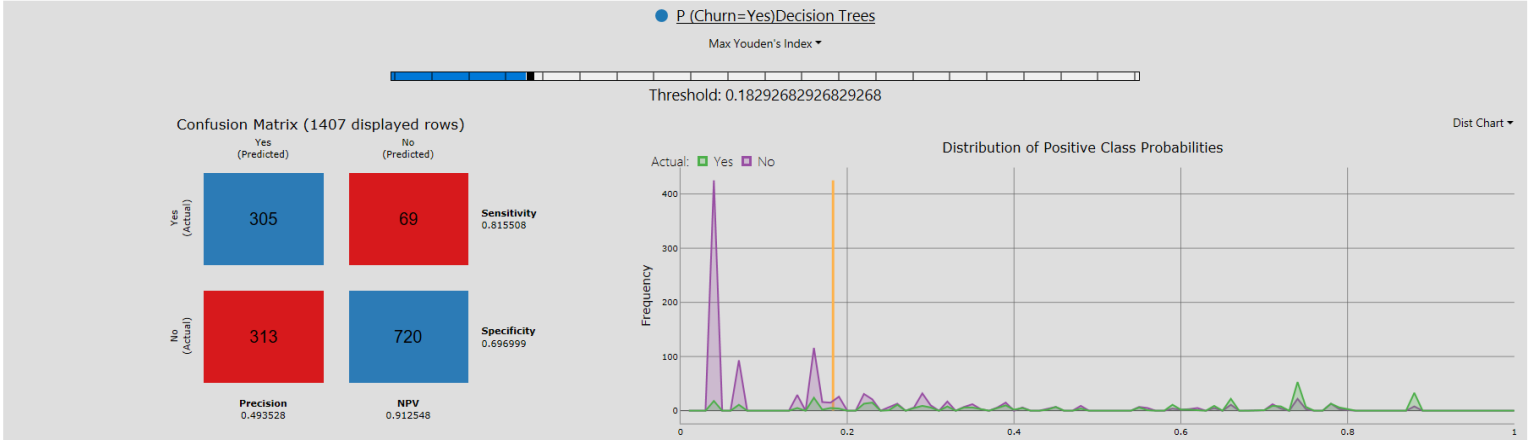
The ROC and Tree view are shown below. The threshold chosen to balance sensitivity and specificity is 0.183. We can see in the tree view that the month-to-month variable is very heterogenous with respect to churn, accounting for 1495 churners. The Fiber optic variable is then used to further try and homogenize the leaves, followed by total charges.

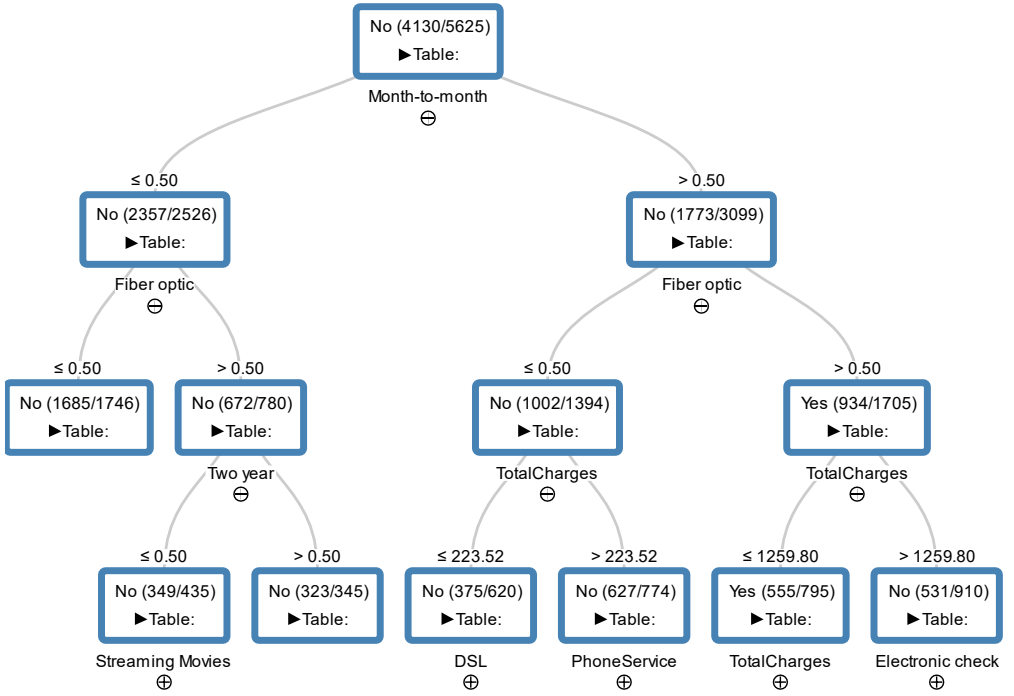
The month-to-month variable being the first root split is in line with the bivariate analysis findings that concluded that the contract type was a key variable in determining churners from none churners. The Fiber optic split after also makes sense given that it also was a variable with a high degree of heterogeneity.

Accuracy is 73%, AUC was 0.82, both stand to improve. The model (at that threshold) also slightly favors sensitivity over specificity, which means that false positives are more common. Precision is 0.49, which means that given a positive prediction, we can almost randomly guess if that prediction is correct or not and get the same result.









Logistic Regression

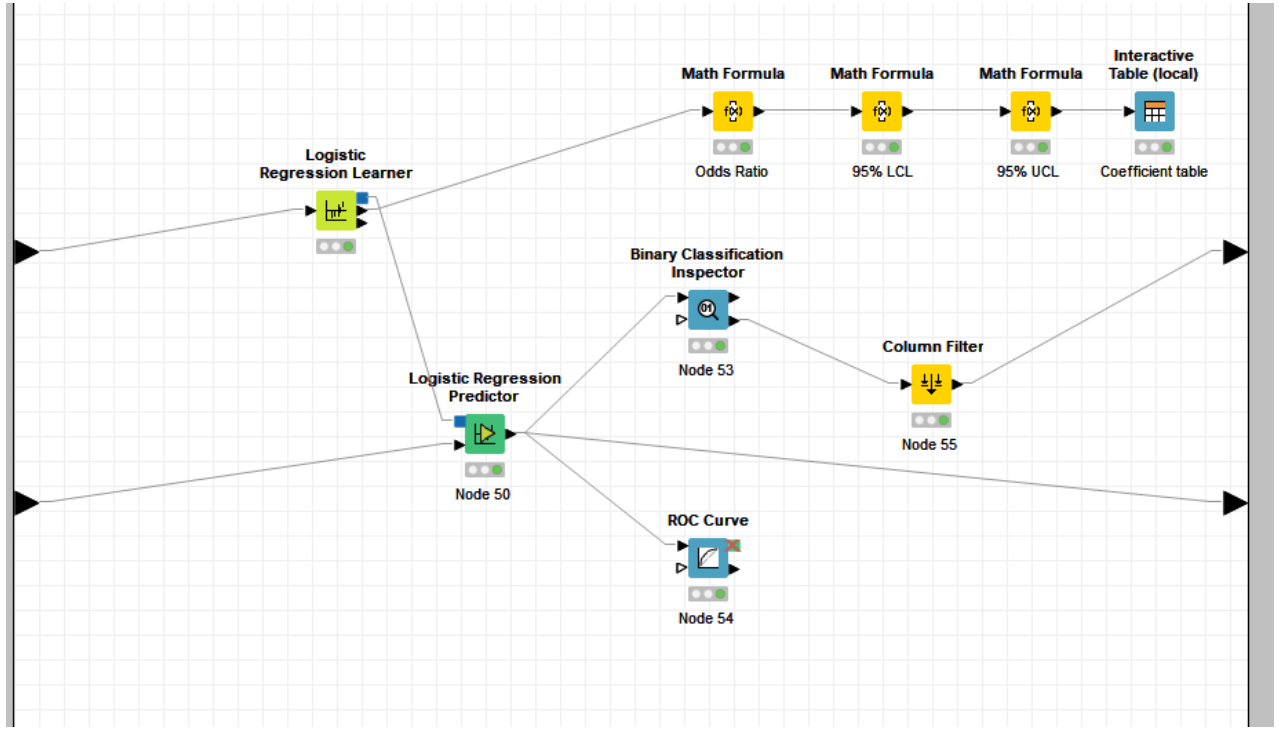
A very simple logistic regression model is implemented as well. An iterative least squares method was used. It was found that regularization using both Lasso and Ridge regression did not yield better results (even after parameter optimization) and were not needed for automatic feature selection because of the feature selection loop implemented after data preparation.

Using a non regularized regression also allows the coefficients for the predictors to not be affected by the normalization of the dataset, which is a needed step before a logistic regression model can be used.

We can see in the table that all the variables are significant with p-values close to 0. The constant is -2.389 which means that when all other variables are 0, the probability to churn is much lower than to stay with the company (P(churn) = 8%).

If we look at the fiber optic or month to month predictors, the coefficients are positive which shows us that they both have positive effects on the probability to churn when the customer belongs in those groups.

The two-year category in contract has a negative coefficient of -0.967 which means that the probability to churn decreases when the customer enters that contract.

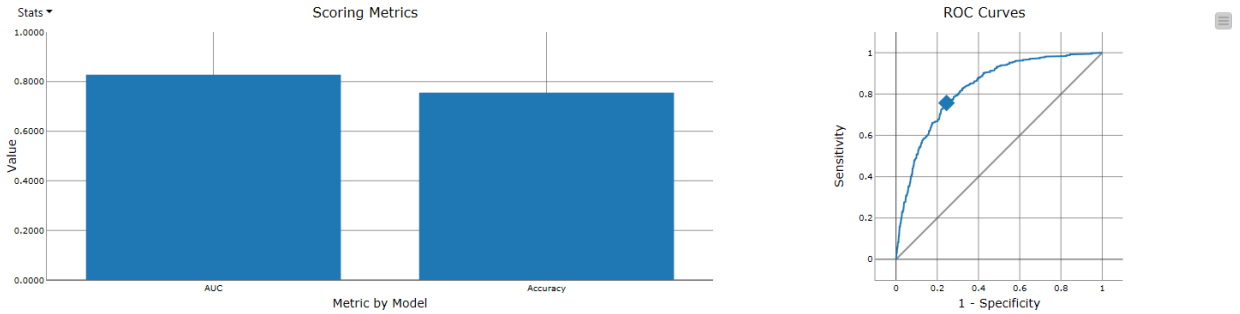


Logit	Variable	Coeff.	Std. Err.	z-score	P>|z|	odds_r...	low_95%	upp_95%
Yes	SeniorCitizen	0.31	0.091	3.386	0.001	1.363	0.13	0.489
Yes	Dependents	-0.235	0.078	-3.021	0.003	0.791	-0.387	-0.082
Yes	PhoneService	-0.387	0.138	-2.806	0.005	0.679	-0.658	-0.117
Yes	MultipleLines	0.23	0.087	2.632	0.008	1.259	0.059	0.401
Yes	PaperlessBilling	0.309	0.082	3.787	0	1.362	0.149	0.469
Yes	TotalCharges	-0	0	-10.712	0	1	-0	-0
Yes	Online Security	-0.43	0.094	-4.575	0	0.651	-0.614	-0.246
Yes	Streaming Movies	0.379	0.084	4.494	0	1.461	0.214	0.544
Yes	DSL	0.84	0.147	5.719	0	2.317	0.552	1.129
Yes	Fiber optic	2.013	0.149	13.464	0	7.483	1.72	2.306
Yes	Month-to-month	0.85	0.115	7.405	0	2.34	0.625	1.075
Yes	Two year	-0.967	0.199	-4.854	0	0.38	-1.357	-0.576
Yes	Electronic check	0.424	0.085	4.975	0	1.528	0.257	0.591
Yes	Mailed check	0.21	0.108	1.94	0.052	1.234	-0.002	0.423
Yes	Constant	-2.389	0.213	-11.215	0	0.092	-2.807	-1.972

The logistic regression AUC is 0.827, with an accuracy of 0.76 and a better balance between specificity and sensitivity than the decision trees model. Precision is also slightly better at 0.53

The threshold used for this balance is 0.3, meaning the model classifies any probability above 0.3 as a churner.

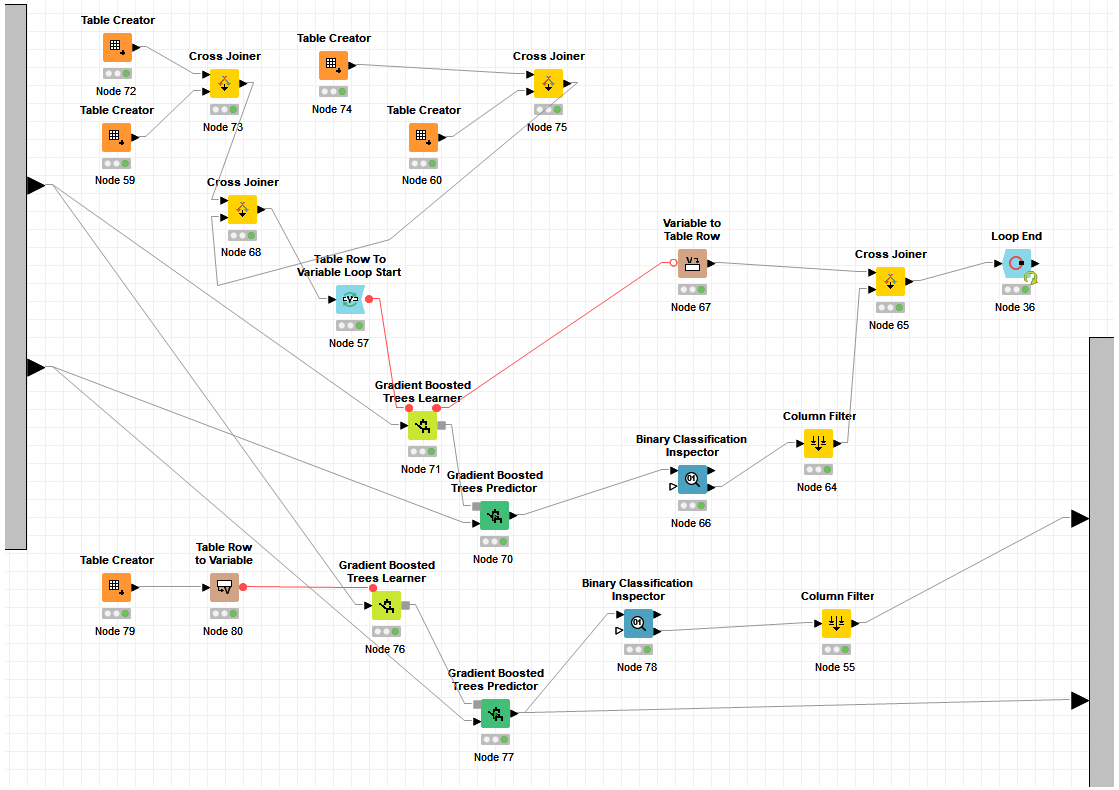
This is an improvement on all fronts from the parameter optimized decision tree model, while also maintaining the ability of interpreting the results and understanding them.



Gradient Boosting:

Gradient Boosting along with the two last models are somewhat of a 'Black Box', in the sense that we cannot interpret the results like we have in decision trees or logistic regression, we can only take them at face value.

The Knime workflow below shows a grid searched gradient boosting loop optimizing learning rate, number of models, minimum node size, and maximum tree level.

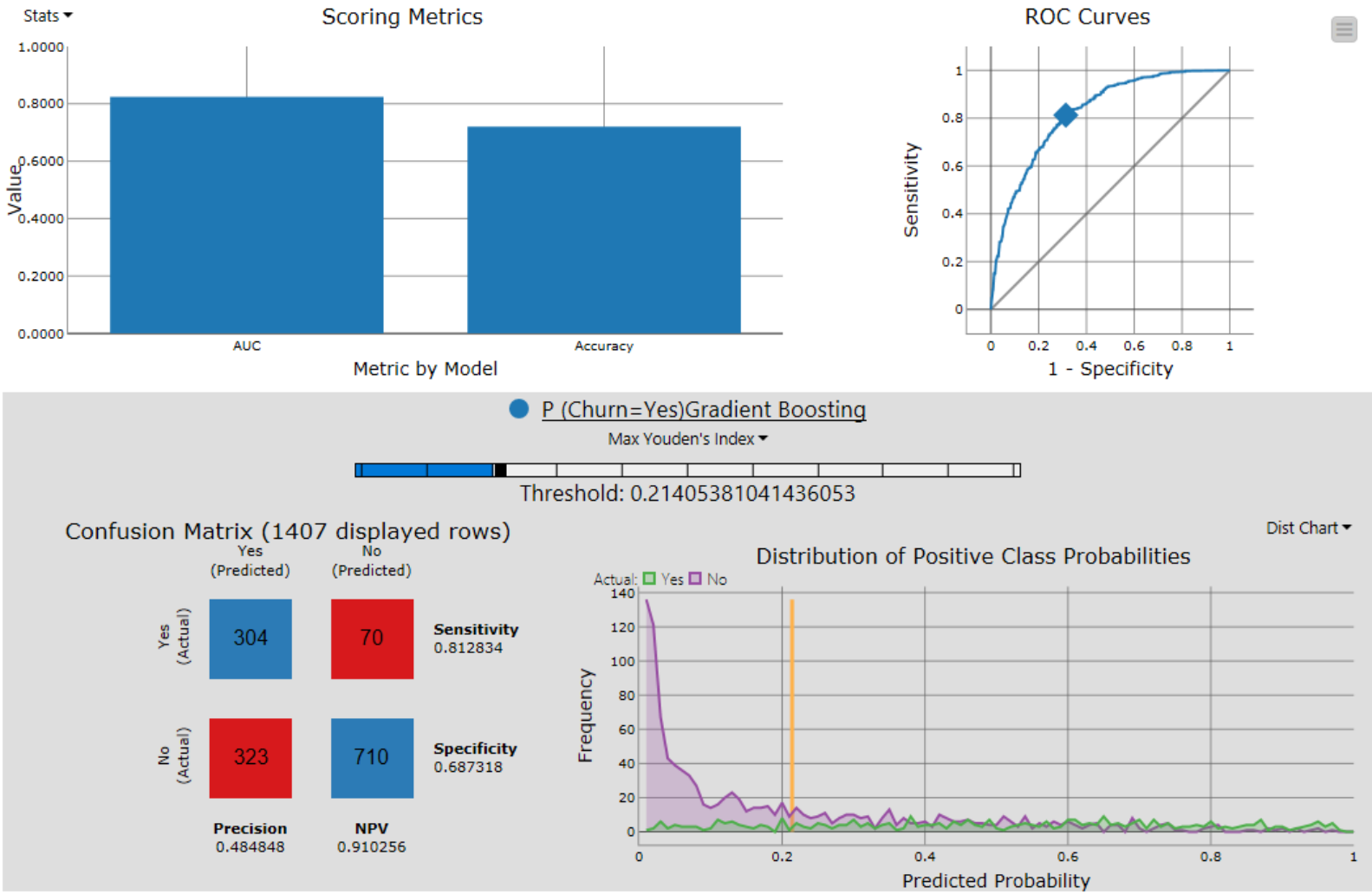


Gradient Boosting:

Being a black box, we can only hope to optimize the parameters and the result should be a more accurate model. This is the trade off between interpretability and ability to accurately predict churn rates. If the performance of these models are substantially better than the logistic regression and decision tree models, we can overlook this and apply the model. Only performance would tell.

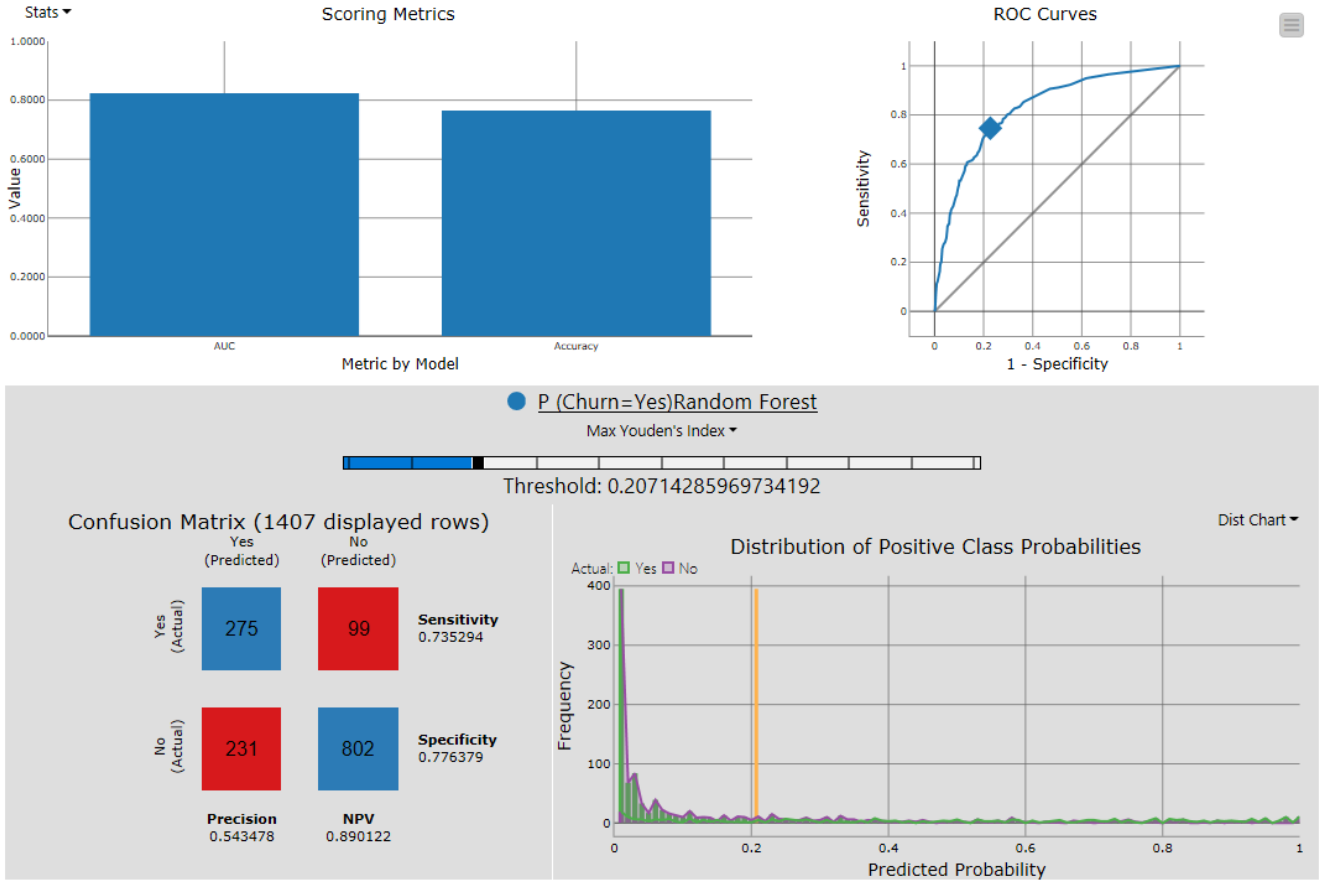
The results, however, are disappointing. After 480 iterations using different combinations of hyper parameters, the AUC for gradient boosting trees was 0.824, accuracy 0.72, with sub par specificity. Precision is also worse than logistic regression at 0.48.

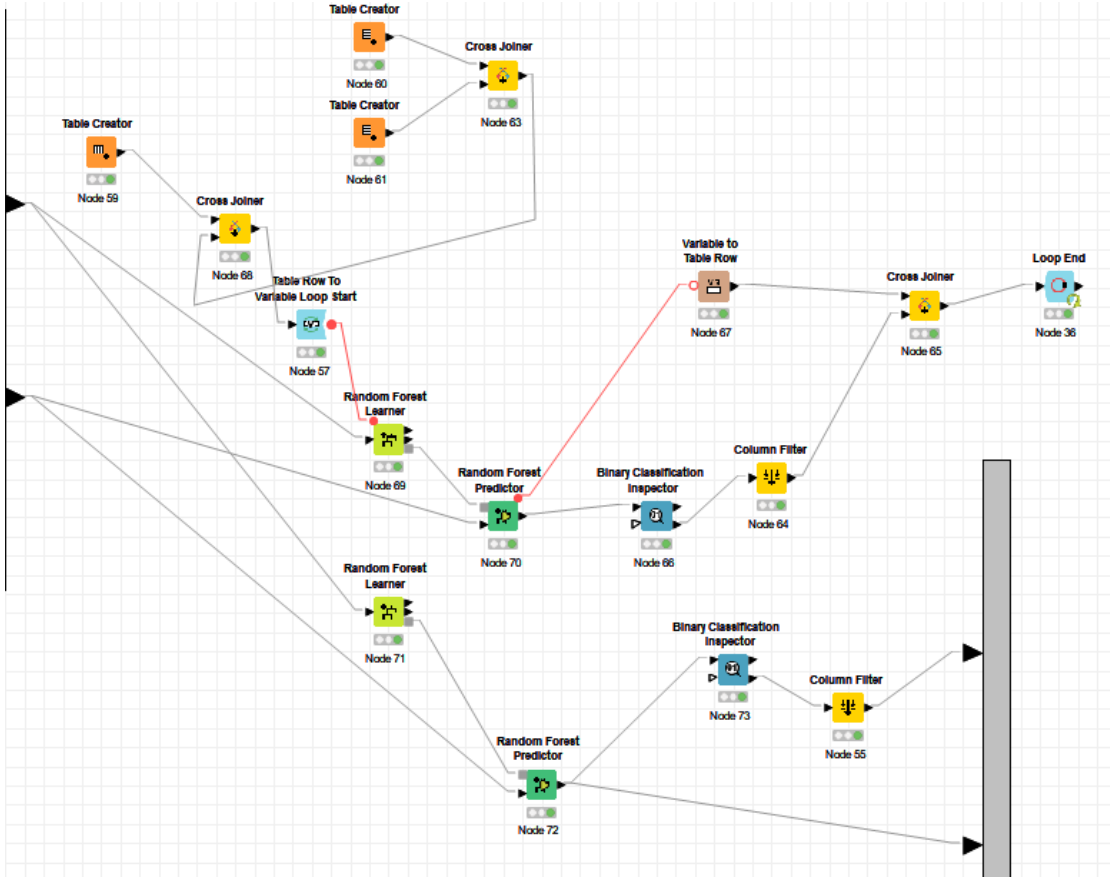
Gradient boosting has failed at all fronts compared to decision trees and logistic regression, and there exists no reason to implement this model in a business or scientific context.



Random Forest

The Random Forest model showed similar disappointing results just like the gradient boosted trees model. AUC score was 0.826, Accuracy 0.74, precision 0.54. It performs similarly to the logistic regression, however its very complicated to interpret and not worth the implementation over a logistic regression, even after hyperparameter tuning, as we lose simplicity and gain nothing in return.



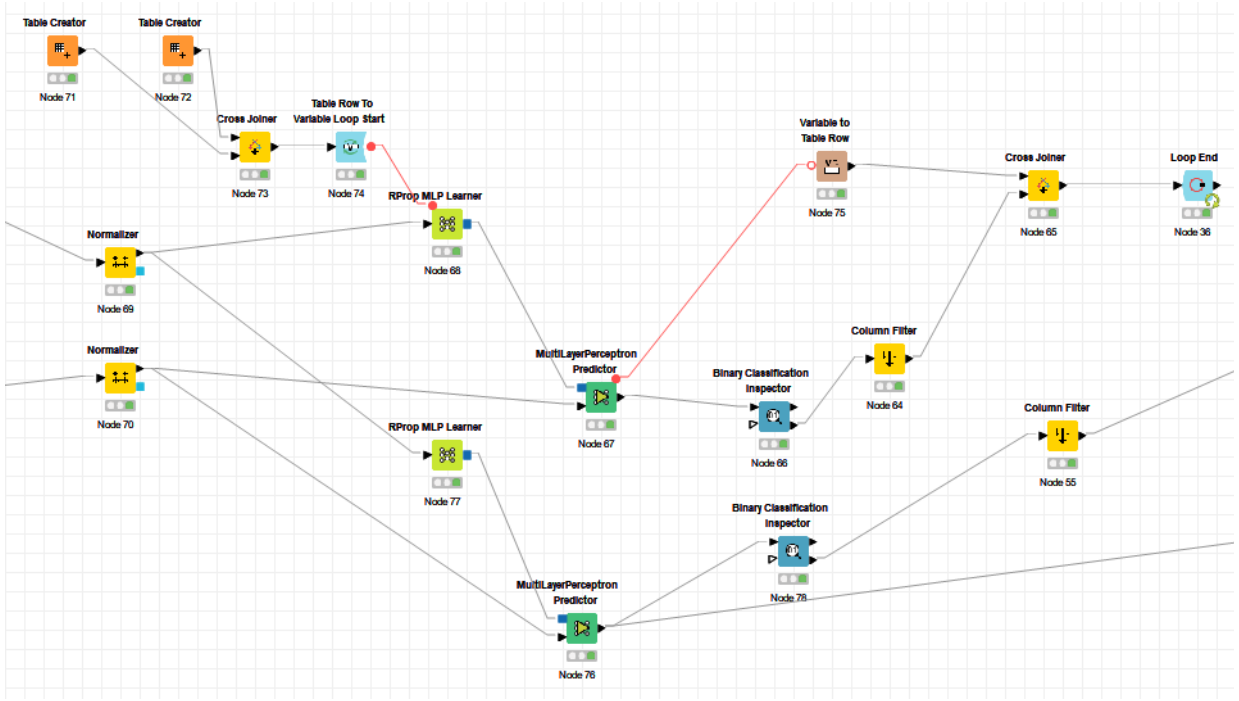


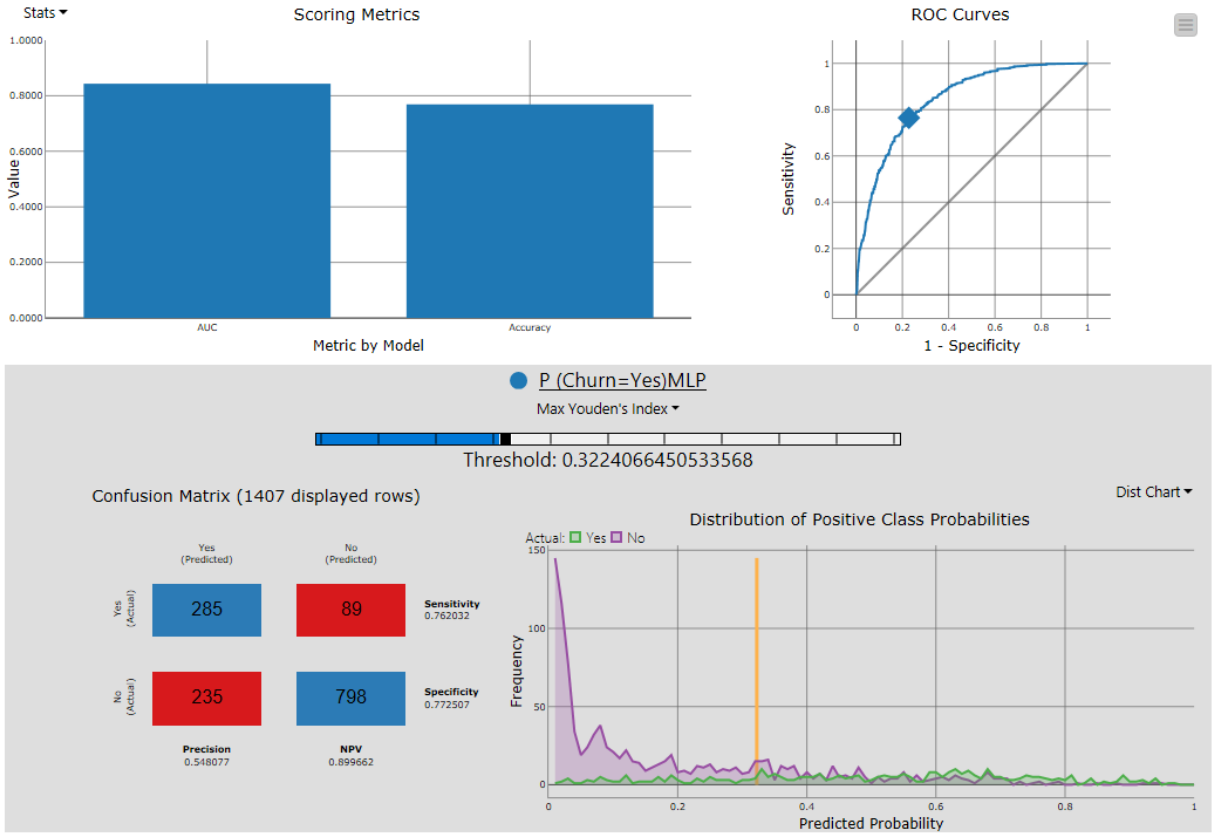
Multi Layer Perceptron

The last ‘Black box’ model to try is the multi layer perceptron. The parameters optimized are number of layers and number of neurons per layer. 44 iterations are made using a grid search, and the results are somewhat better than Gradient Boosting and Random Forests.

The AUC score is 0.843, the highest yet. Accuracy is 0.77 and both specificity and sensitivity are perfectly balanced. Precision is also the best in all the models found at 0.55. It improved on every front compared to logistic regression.

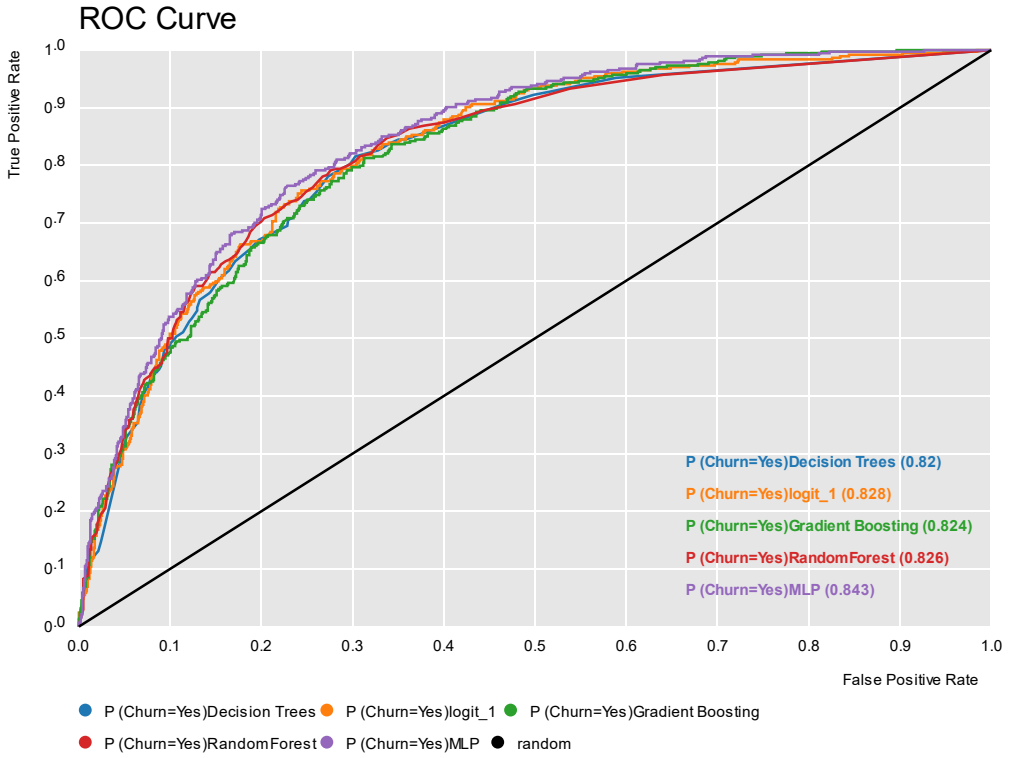
A summary analysis is done on the next page to compare the five models trained on the data

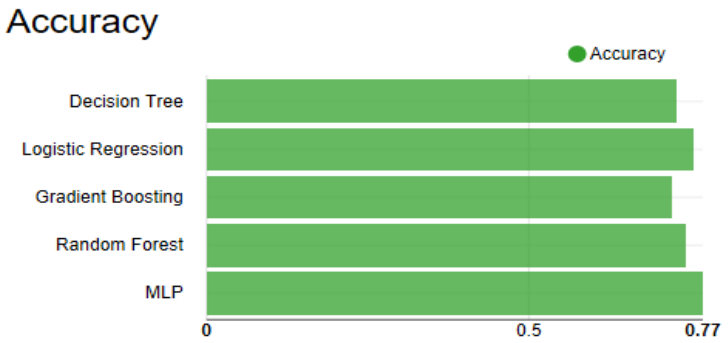


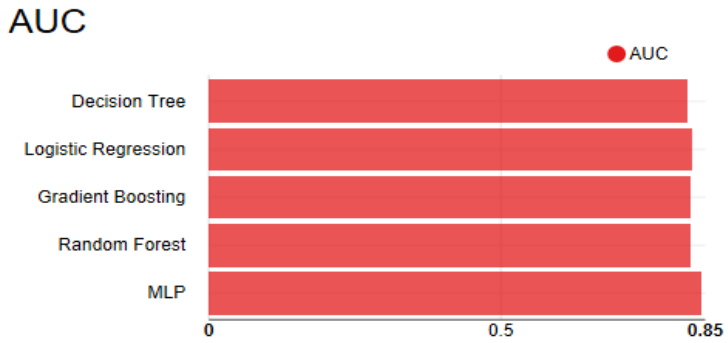


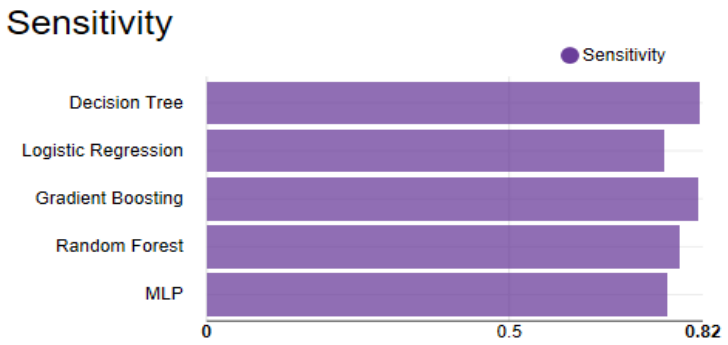
As we can see, MLP dominates on all fronts except sensitivity. However, a 1.5% improvement on AUC over a logistic regression is not substantial enough for us to give away interpretability of the model, especially in the context of customer churn rates where Telco needs to understand the drivers of customer choices in order to improve their services and make better offers, so MLP is not recommended. Overall, the choice seems to lie between a logistic regression and a decision tree.

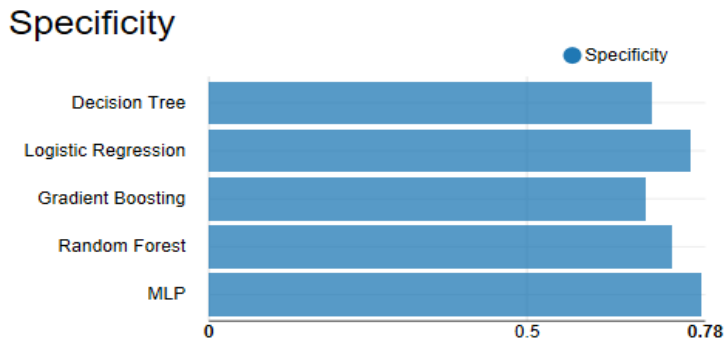
Model Name	Accuracy	Precision	Sensitivity	Specificity	AUC
Decision Tree	0.729	0.494	0.816	0.697	0.82
Logistic Regression	0.756	0.529	0.757	0.756	0.828
Gradient Boosting	0.721	0.485	0.813	0.687	0.824
Random Forest	0.743	0.51	0.783	0.728	0.826
MLP	0.77	0.548	0.762	0.773	0.843











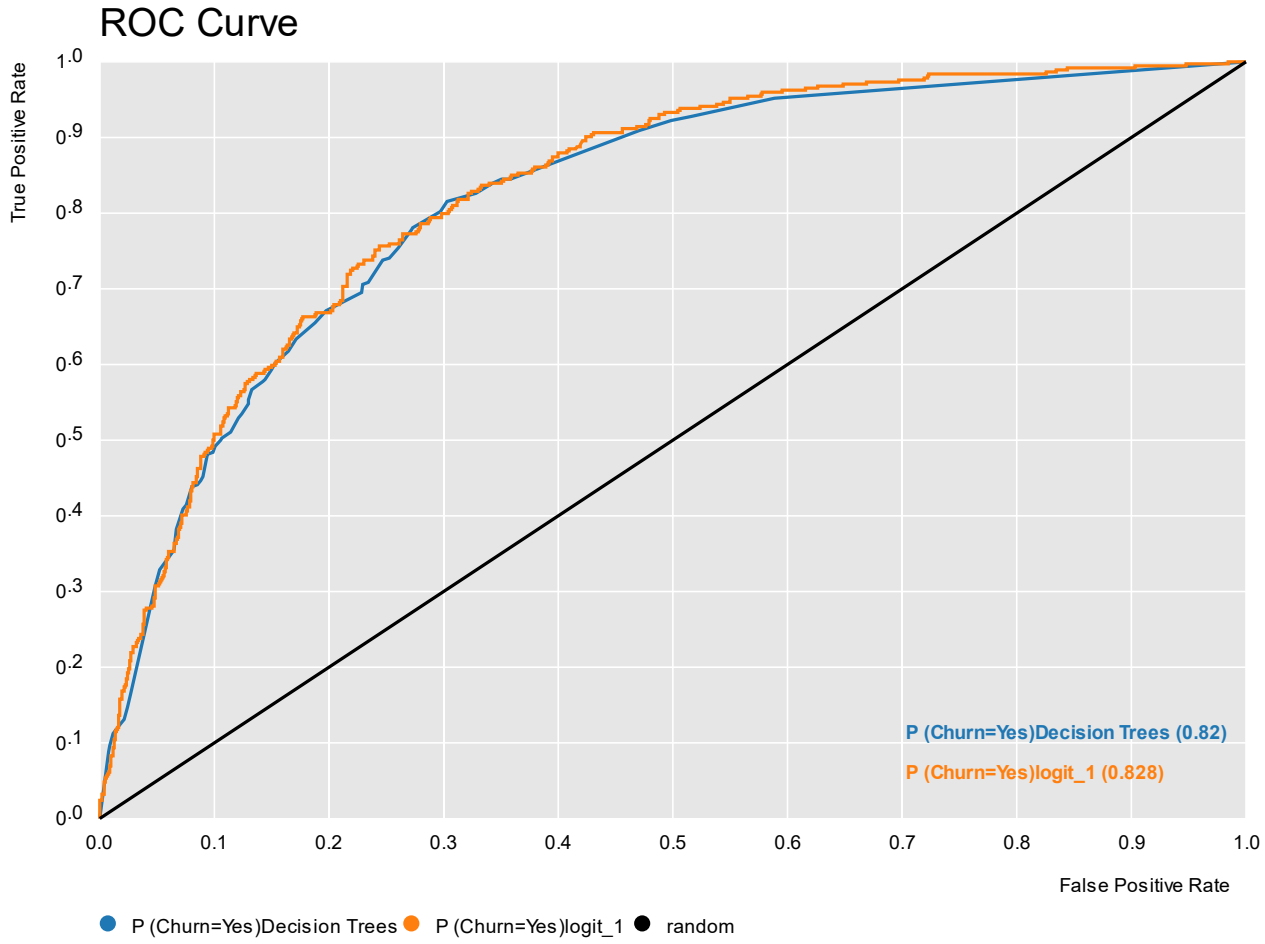
Conclusions

Having Chosen to implement a model with high interpretability, we are left with either a decision tree or logistic regression. For our purposes, as mentioned earlier, given no prior knowledge of specific misclassification costs a false positive and a false negative are equally weighted. Logistic regression can balance both, as well as increase accuracy and AUC. It can also be very easily implemented with the full Telco dataset as it is not computationally expensive to train, especially with reduced features.

So, in conclusion, a logistic regression would be an acceptable way to model Telco's data for prediction and to aid in managerial decision making.

Extra Note:

The 5 tests were done using a train-test split on the original distribution of the dataset without balancing. It was found during the modelling phase that if the data was balanced first this resulted in exaggerated AUC and accuracy scores in all models. It was decided to not balance before splitting the data, as this does not represent a real-world scenario where the test data would also be balanced. It is assumed that the distribution of variables in the original table matches the real world, and thus the test set had to be left unchanged.



Model Name	Accuracy	Precision	Sensitivity	Specificity	AUC
Decision Tree	0.729	0.494	0.816	0.697	0.82
Logistic Regression	0.756	0.529	0.757	0.756	0.828