Team NC-227027

Team members

- Adarsh S
- Narendran Omprakash

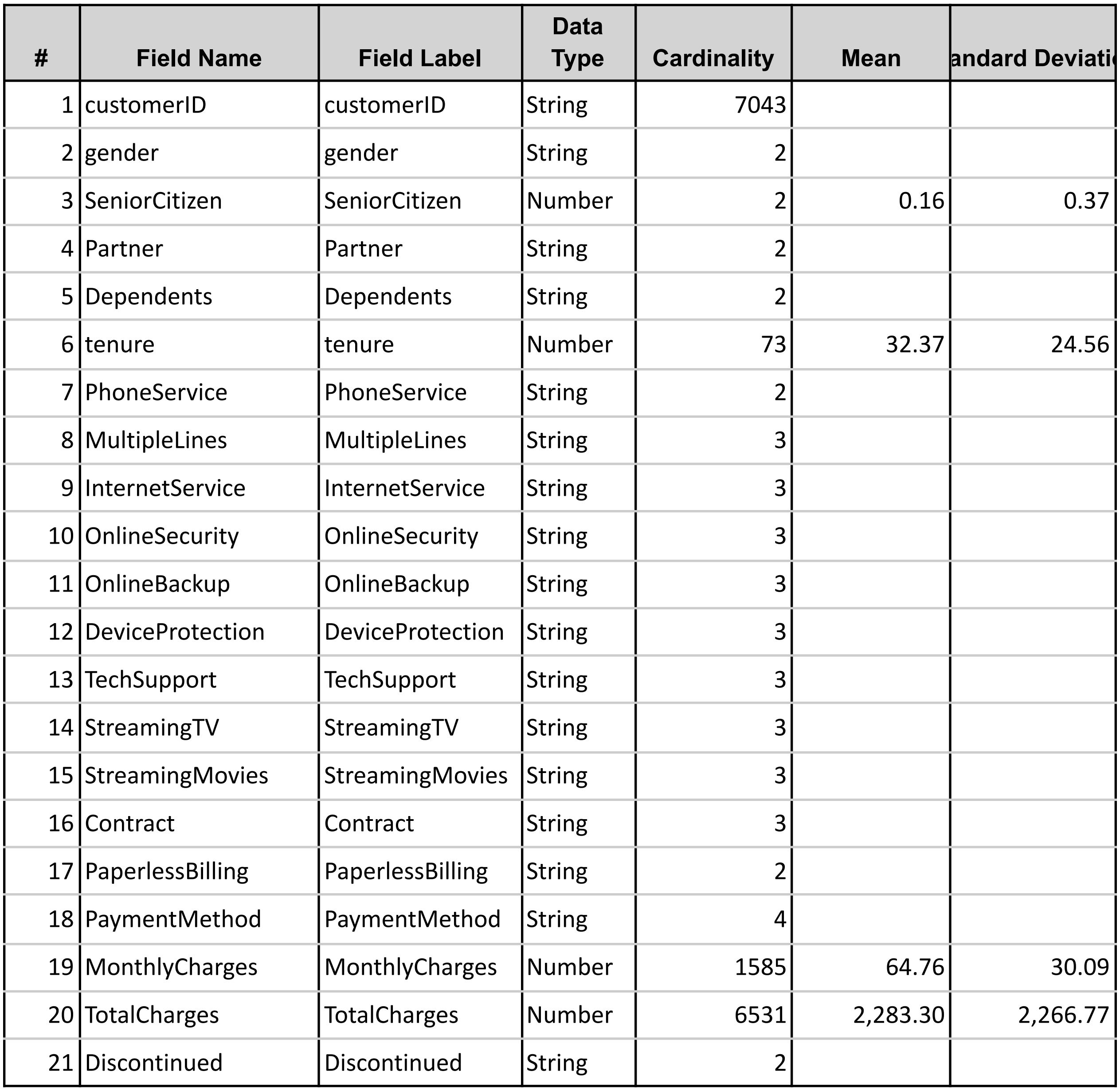
No Code ML

24th January 2022

DATASET OVERVIEW AND IMPORT (Provided in Dataset description)

- We import the dataset into Knowledge Studio using the excel import node
- The shape and size of the data are given below
  - 7043 customer details are provided with their own unique and random customerID
  - Demographical characteristics such as Gender, Senior Citizenship, and presence of partner or dependents are provided.
  - The telecom company provides two basic services: Phone Service and Internet service. Certain other bundled services can also be availed by the customer if they use each of these services. The phone service allows you to get a multiple-line collection.
  - Details of the payment duration, method, and amount for customers have been provided.
  - Finally, the important (output) feature has been provided- whether or not the customer discontinued (stopped using) the service in the past few months.

Size and shape of data



#	Field Name	Field Label	Data Type	Cardinality	Mean	andard Deviati
1	customerID	customerID	String	7043		
2	gender	gender	String	2		
3	SeniorCitizen	SeniorCitizen	Number	2	0.16	0.37
4	Partner	Partner	String	2		
5	Dependents	Dependents	String	2		
6	tenure	tenure	Number	73	32.37	24.56
7	PhoneService	PhoneService	String	2		
8	MultipleLines	MultipleLines	String	3		
9	InternetService	InternetService	String	3		
10	OnlineSecurity	OnlineSecurity	String	3		
11	OnlineBackup	OnlineBackup	String	3		
12	DeviceProtection	DeviceProtection	String	3		
13	TechSupport	TechSupport	String	3		
14	StreamingTV	StreamingTV	String	3		
15	StreamingMovies	StreamingMovies	String	3		
16	Contract	Contract	String	3		
17	PaperlessBilling	PaperlessBilling	String	2		
18	PaymentMethod	PaymentMethod	String	4		
19	MonthlyCharges	MonthlyCharges	Number	1585	64.76	30.09
20	TotalCharges	TotalCharges	Number	6531	2,283.30	2,266.77
21	Discontinued	Discontinued	String	2		

DATA CLEANUP

Missing value treatment

- Missing value substitution is done by the following criteria
  - Discrete Variable - Substitute Mode
  - Continuous Variable - Substitute Mean
- 11 entries in TotalCharges are missing. This is treated by substituting the mean.
- This can be done using the Variable transformations node in Knowledge studio using the following expression

```
CASE WHEN [TotalCharges] IS NULL THEN Avg([TotalCharges]) ELSE
[TotalCharges] END
```

Outliers Treatment

The monthly and total charges columns don't contain any outliers, they lie in the range that was previously specified in the data documentation.

Variable Selection

To highlight potential predictors, we use a variety of different measures such as Entropy, Gini, Information Value, likelihood test statistic ratio, etc. By using the union option, we select the top 2 variables for each measure.

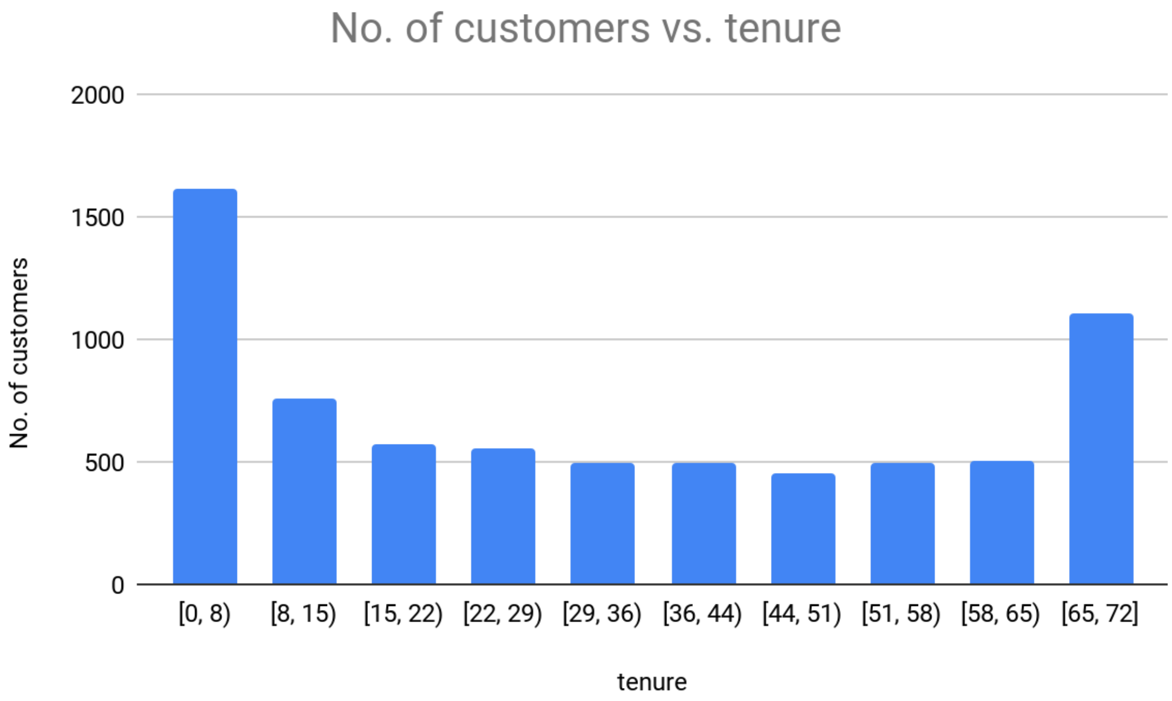
DATA ANALYSIS

Preliminary observations

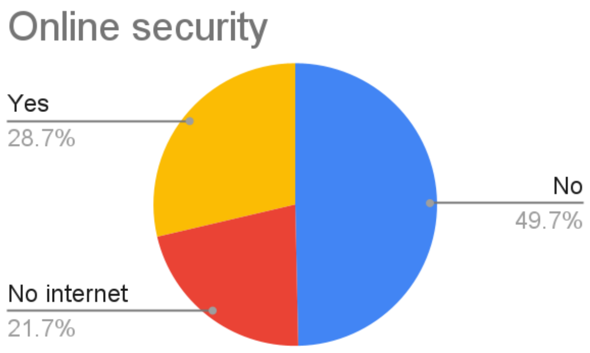
- The dataset has an equal number of male (50.48%) and female (49.52%) records.
- 16.2% of the customers in the dataset are senior citizens.
- Almost half of the customers in the dataset have a partner (48.30%).
- 30% of the customers have dependants.
- The tenure of the customers seems to follow a roughly symmetrical bimodal distribution with peaks at 0-8 months or 65-72 month periods.
- Only about 10% of customers do not avail the phone service.
- 43% of customers avail fiber optic internet service while 34% use DSL and 21% have not opted for internet service.
- 55% of customers are on a month-to-month contract while 20% are on a yearly contract and 24% are on a biannual contract.
- 26.54% of the customers in the dataset have discontinued the service.

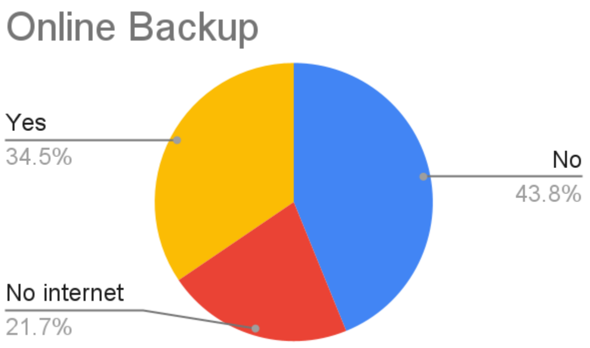
Graphs and charts

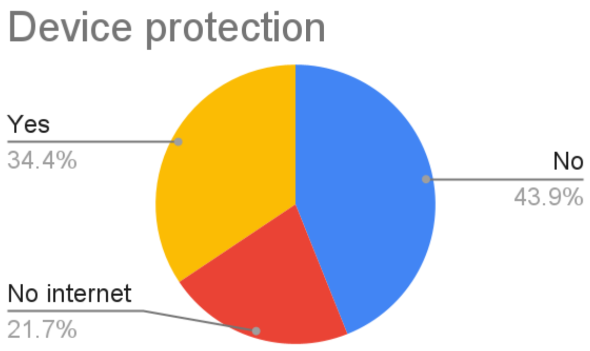
Tenure

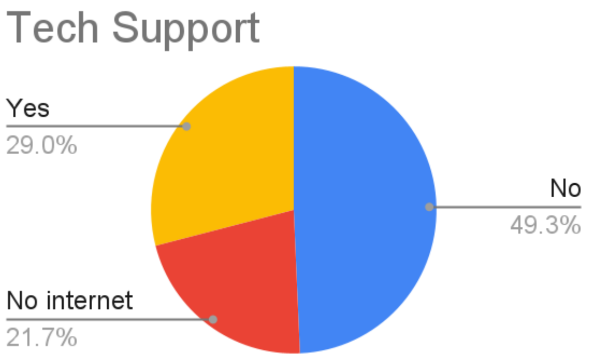


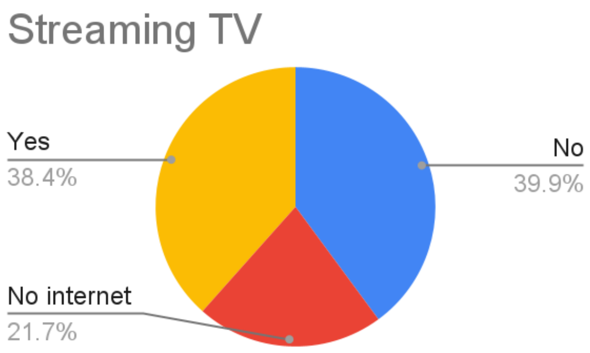
Additional services

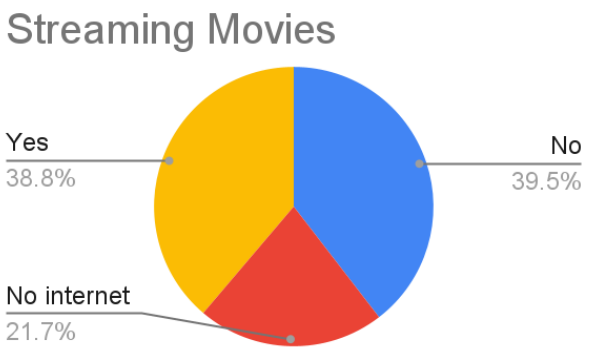




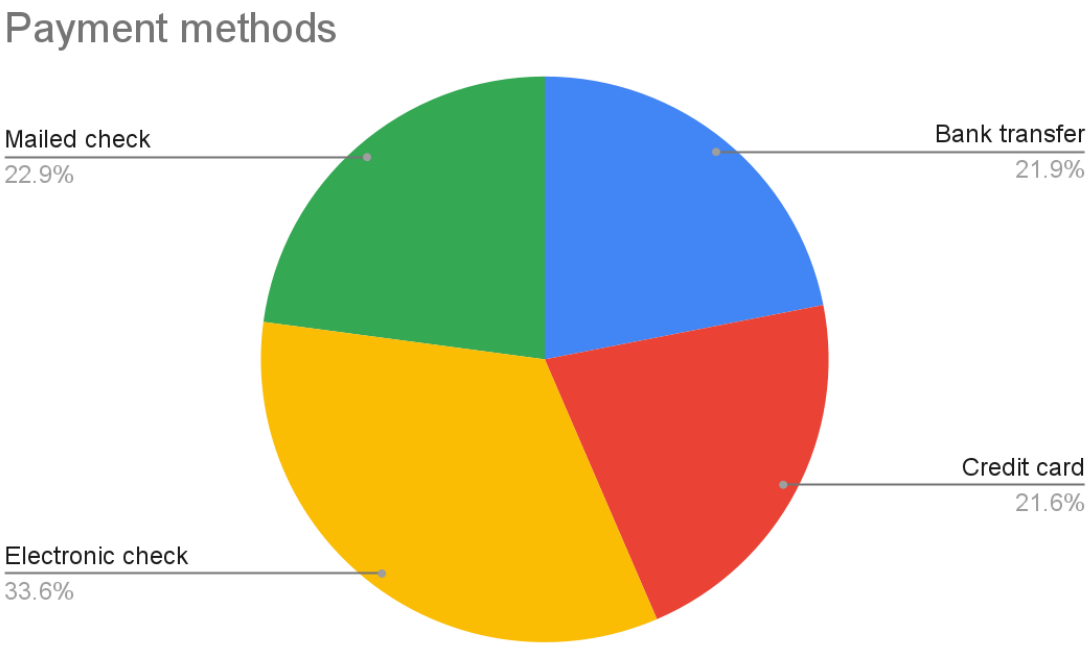




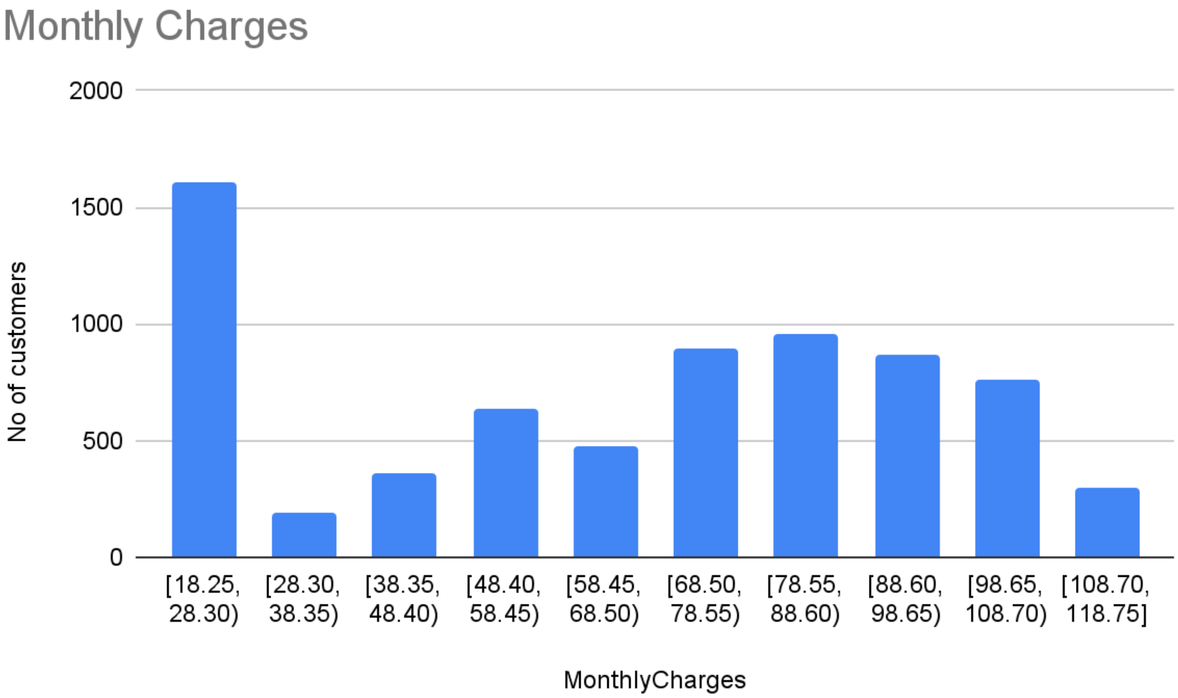




Payment methods

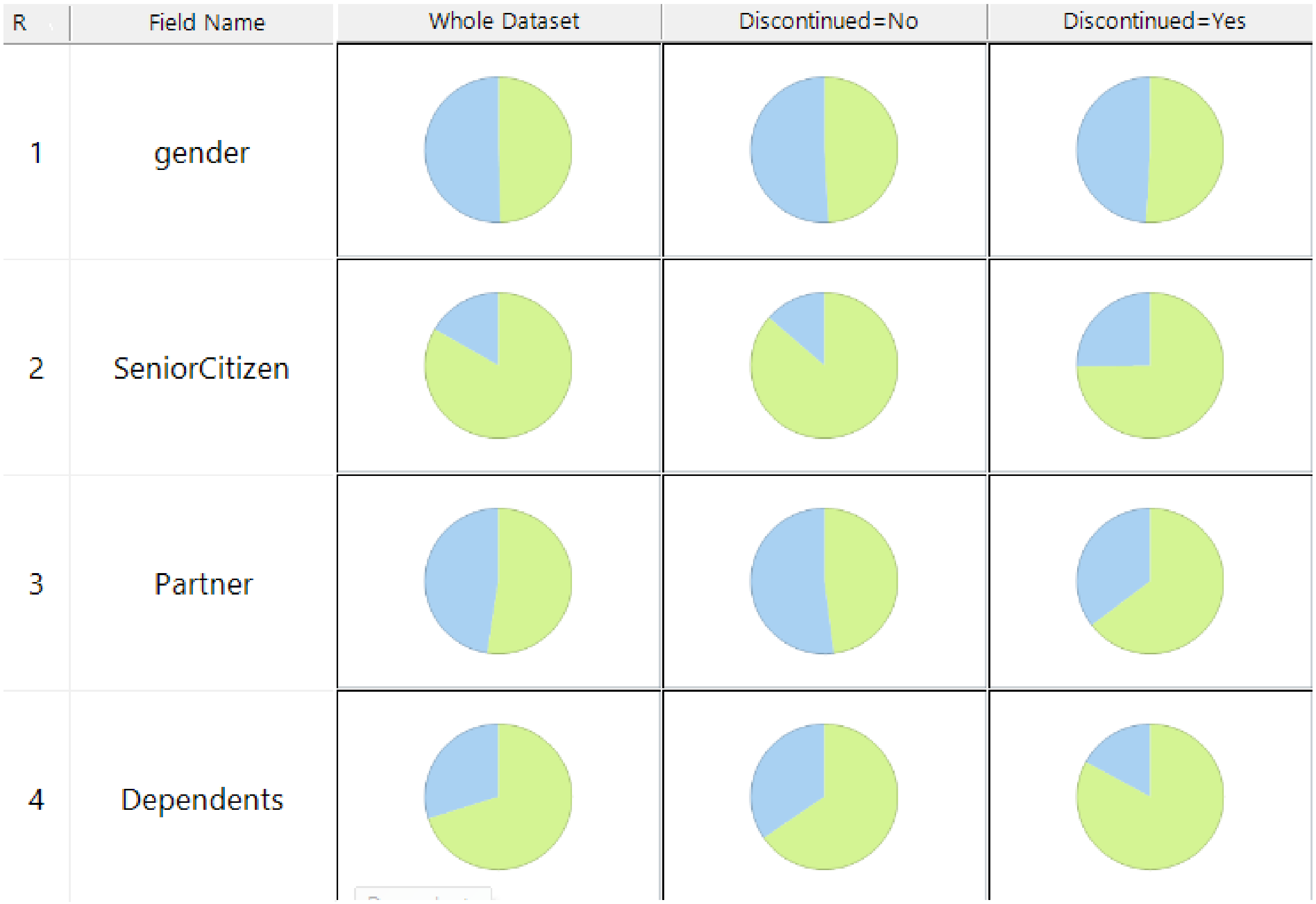


Monthly charges

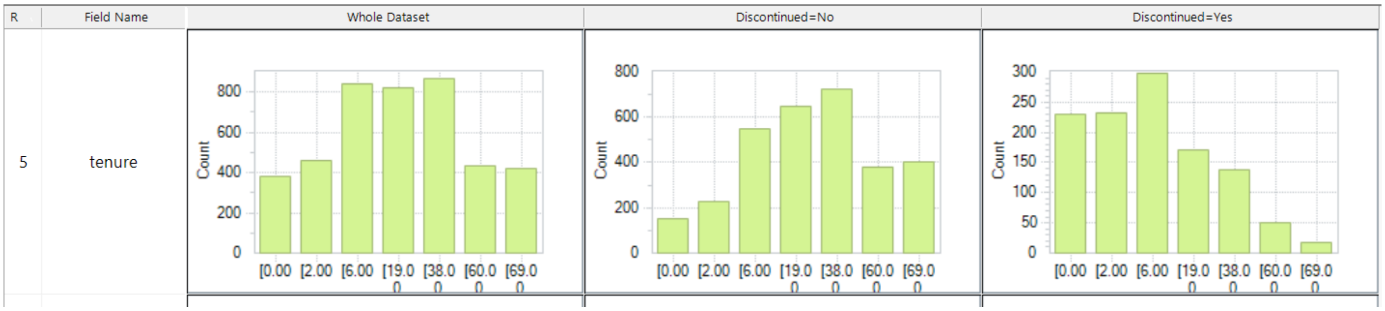


Segment Analysis

We take every value and compare it against the target variable (Discontinued)



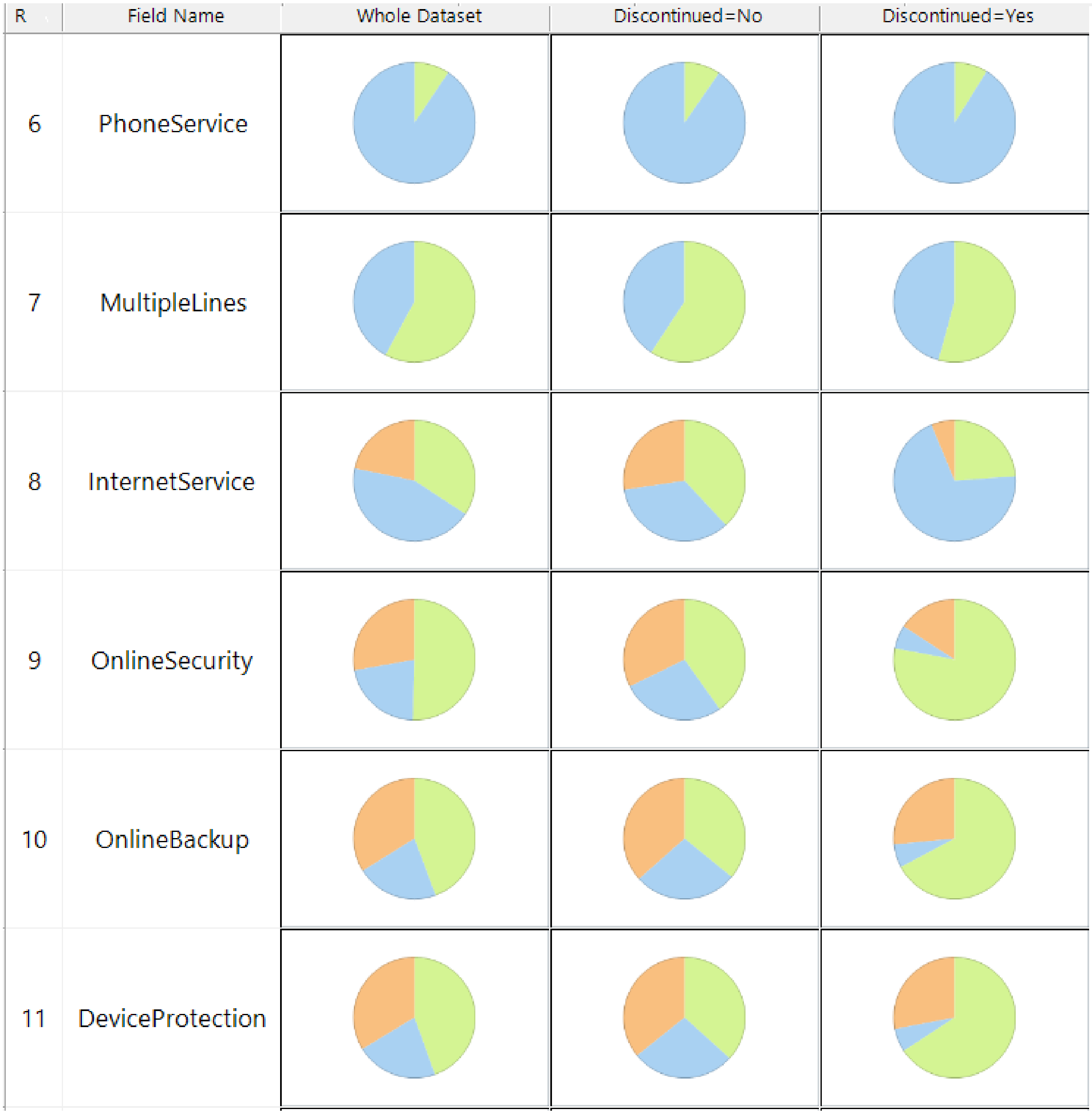
R	Field Name	Whole Dataset	Discontinued=No	Discontinued=Yes
1	gender			
2	SeniorCitizen			
3	Partner			
4	Dependents			



R	Field Name	Whole Dataset	Discontinued=No	Discontinued=Yes
5	tenure			

Observations

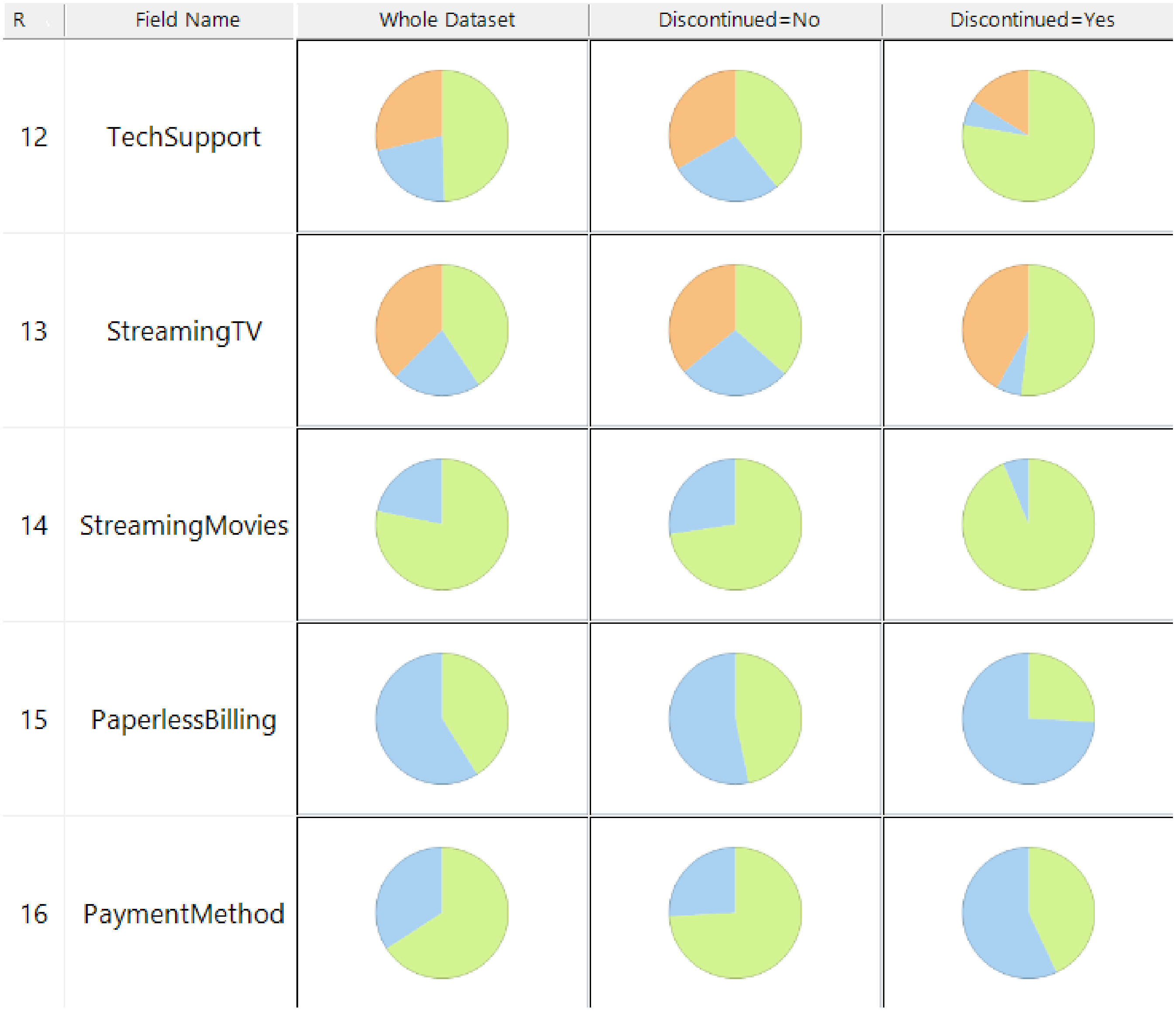
- Gender has no Information value and hence can be eliminated. Partner, Senior citizen and Dependants have an Information value of [0.08, 0.16].
- Customers with less tenure seem to be more probable to discontinue than customers with large tenures.



R	Field Name	Whole Dataset	Discontinued=No	Discontinued=Yes
6	PhoneService			
7	MultipleLines			
8	InternetService			
9	OnlineSecurity			
10	OnlineBackup			
11	DeviceProtection			

Observations

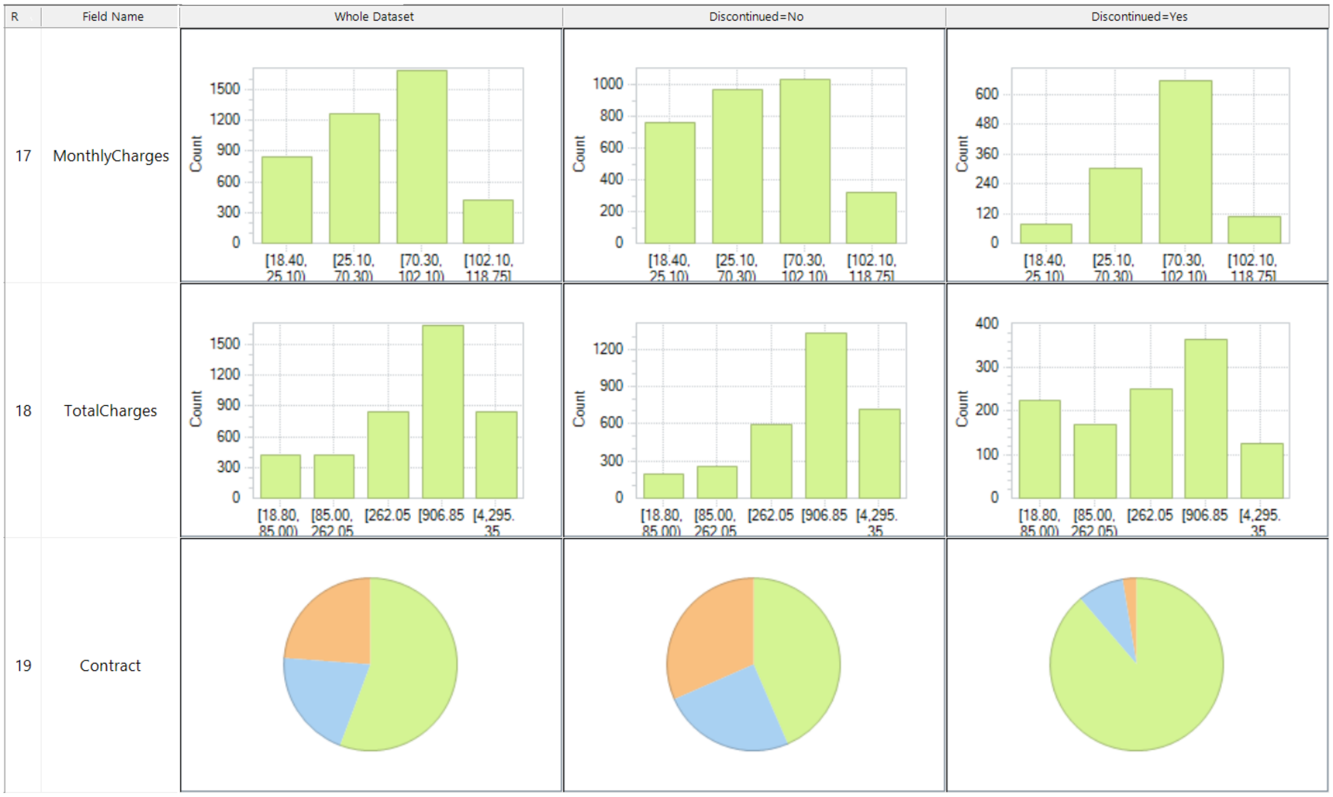
- Phone service, multiple lines have no Information value.
- Customers who opt for fiber optic internet seem more likely to discontinue the service
- Customers who opt for additional services (Online backup, Online security, Device protection, Tech support, Streaming TV, Streaming Movies) are generally less likely to discontinue the service.



R	Field Name	Whole Dataset	Discontinued=No	Discontinued=Yes
12	TechSupport			
13	StreamingTV			
14	StreamingMovies			
15	PaperlessBilling			
16	PaymentMethod			

Observations

- Customers who opt for automatic payments are less likely to discontinue the service. Customers who send electronic checks discontinue the most.



R	Field Name	Whole Dataset	Discontinued=No	Discontinued=Yes
17	MonthlyCharges			
18	TotalCharges			
19	Contract			

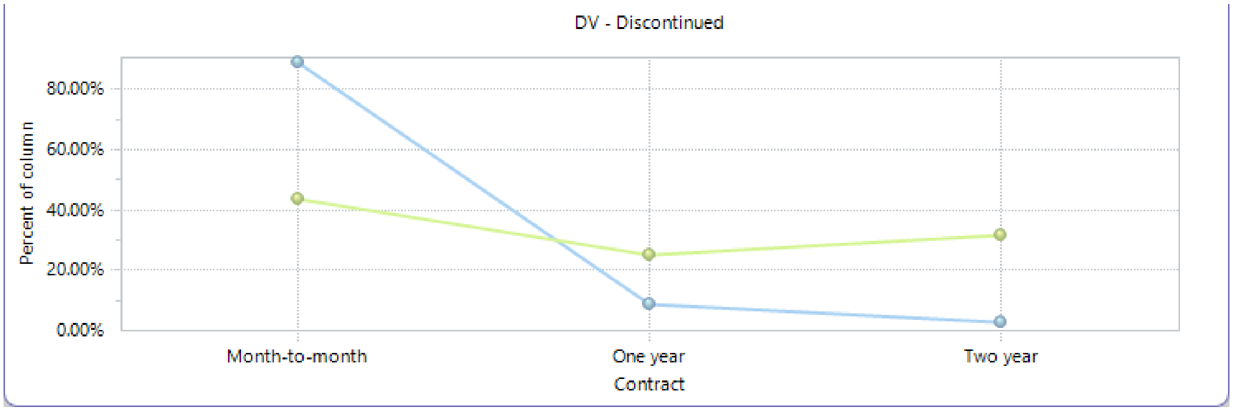
Observations

- Customers who discontinue are mostly on a monthly contract (88%).
- Only 3% of customers with two-year contracts and 8% of customers with annual contracts have discontinued, while 43% of customers on monthly contracts have discontinued.
- Most customers (90%) with low monthly charges have not discontinued.

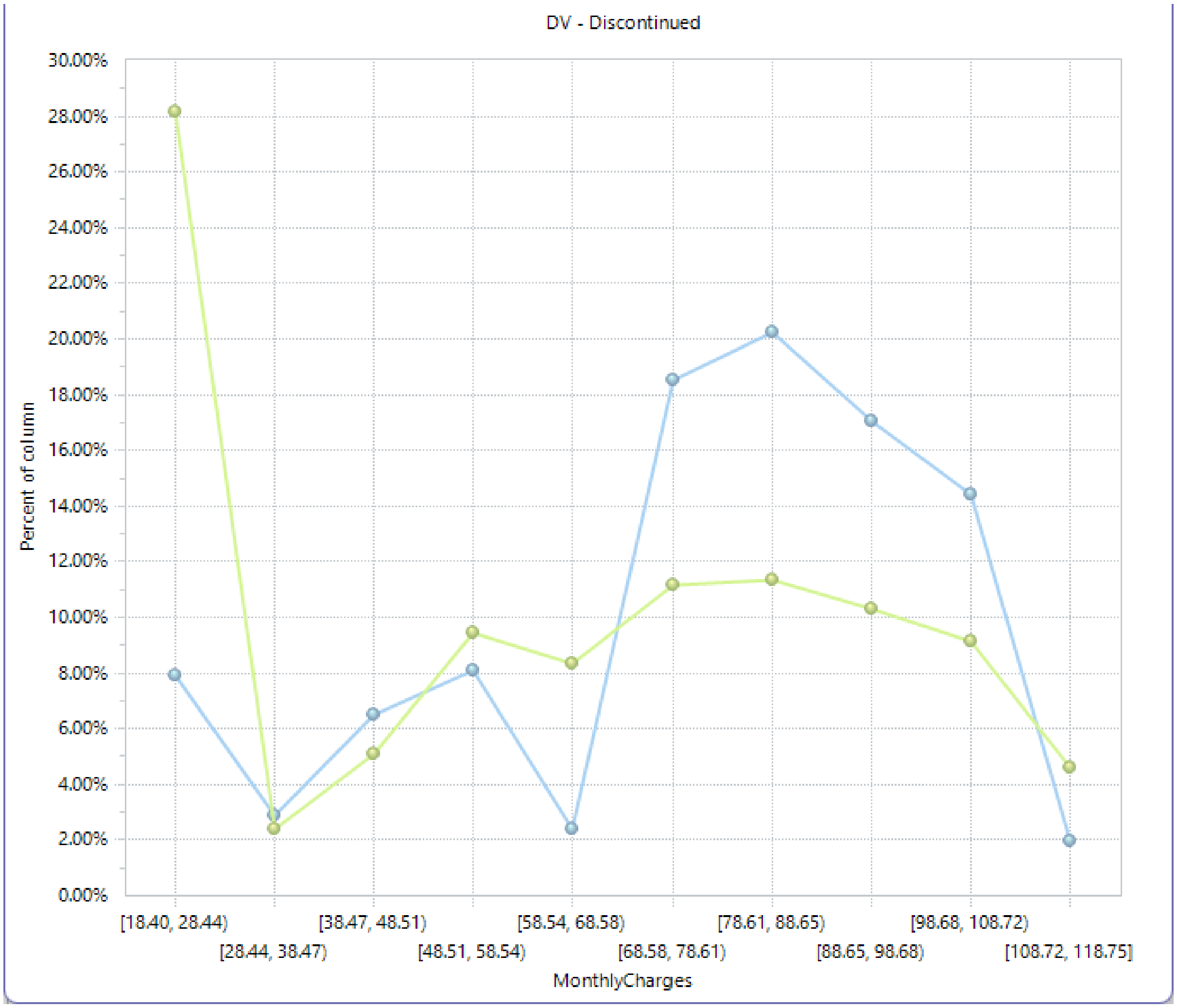
Characteristic Analysis

Characteristic analysis charts

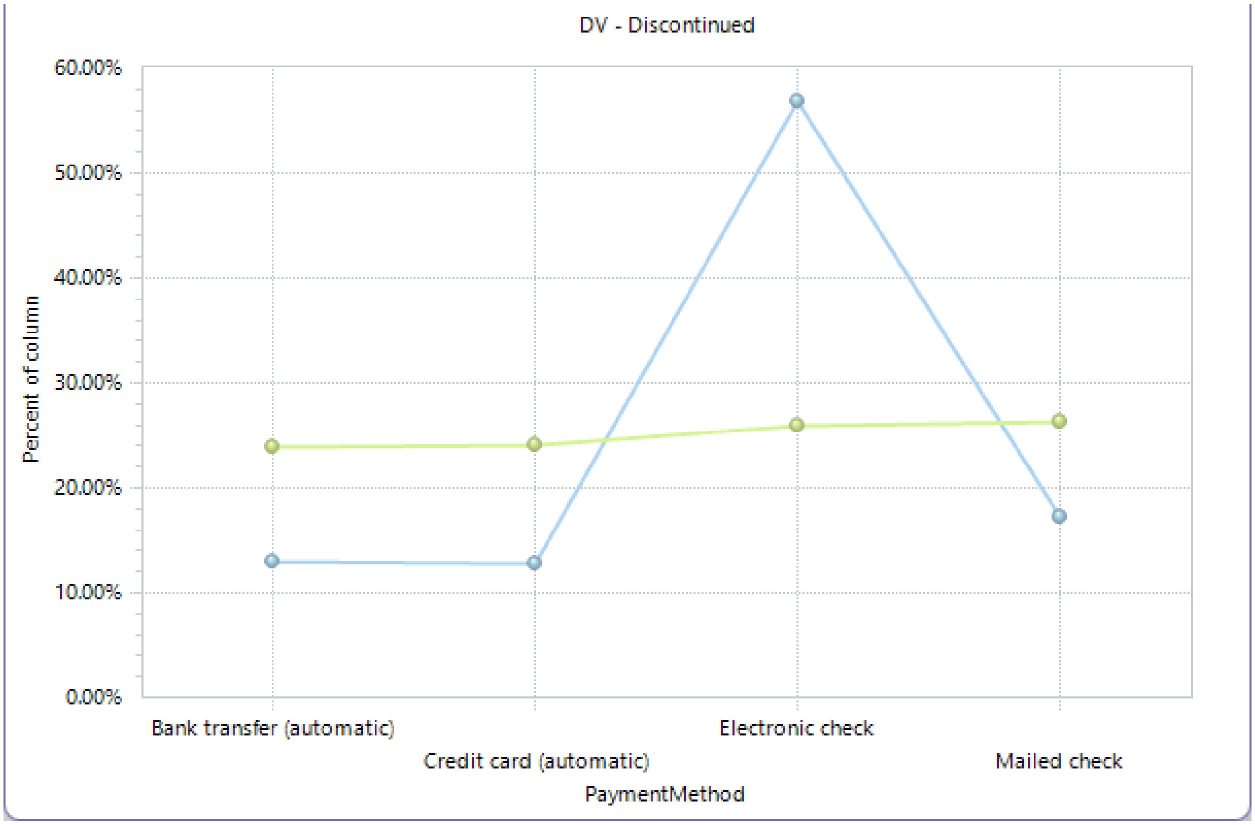
Contract



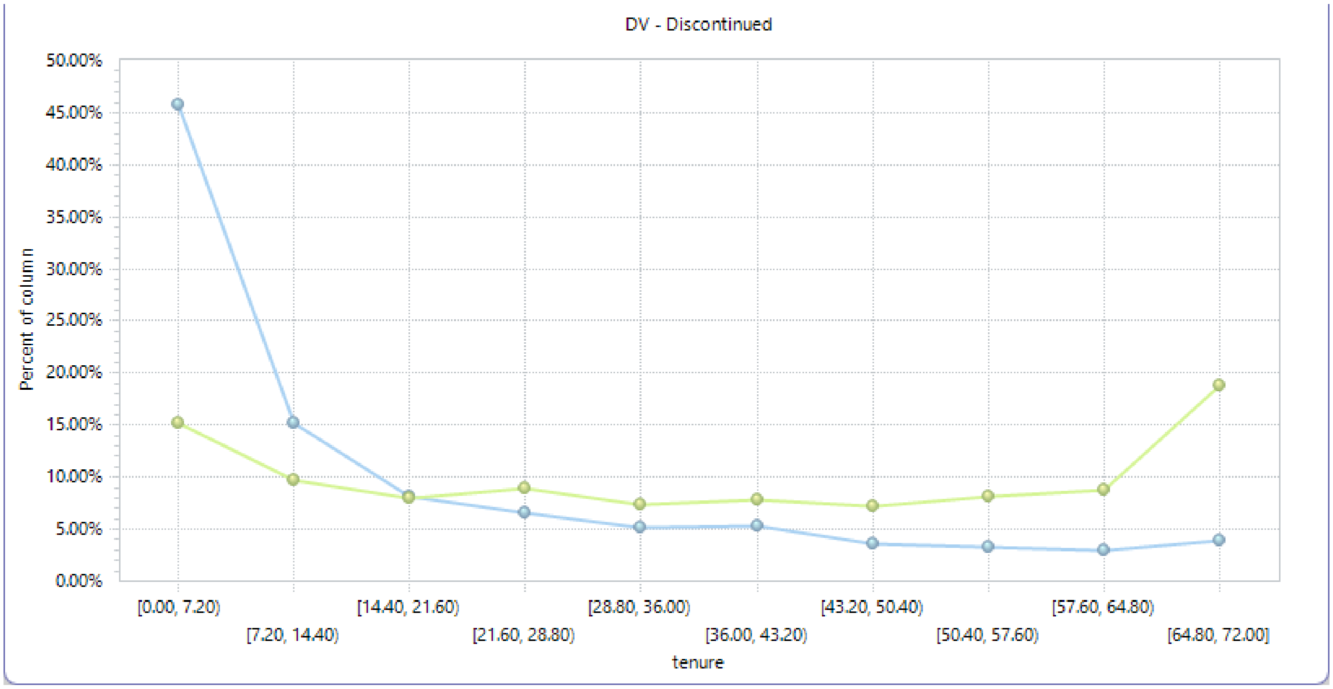
Monthly charges



Payment method



Tenure



DATA PRE-PROCESSING

Target attribute: Discontinued=Yes

Train-test split

The dataset is split into the following partitions:

- Train - 60%
- Test - 30%
- Fit - 10%

This can be done using Knowledge studio's partition node.

Checking statistical characteristics of train and test data

We check for some of the comparable attributes of the train, test and fit data to ensure they have similar characteristics

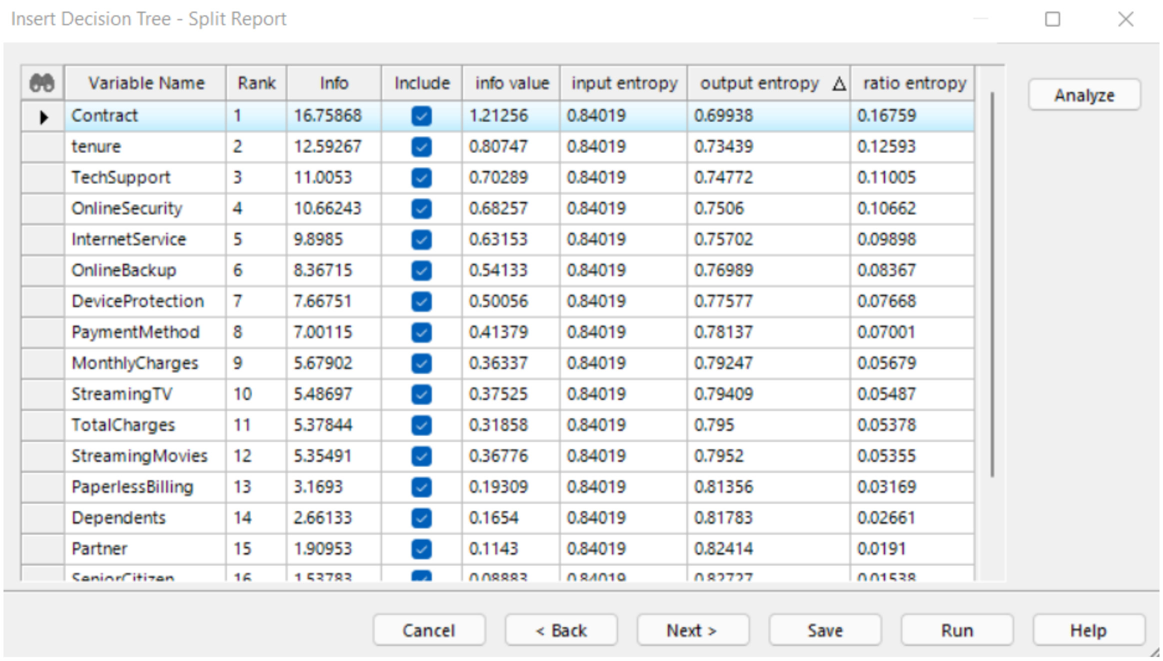
Attribute	Train	Test	Fit
Cardinality	4225	2113	705
Gender	Male: 50.37% Female: 49.63%	Male: 48.84% Female: 51.16%	Male: 56.03% Female: 43.97%
Senior citizen	Yes: 16.73%	Yes: 16.31%	Yes: 15.14%
Phone service	Yes: 90.65%	Yes: 89.65%	Yes: 89.87%
Internet service	DSL: 34.27% Fiber optic: 44.00% No: 21.73%	DSL: 33.19% Fiber optic: 45.67% No: 21.13%	DSL: 34.97% Fiber optic: 43.30% No: 21.72%
Contract	Monthly: 55.72% One year: 20.47% Two year: 23.81%	Monthly: 54.09% One year: 21.39% Two year: 24.51%	Monthly: 53.62% One year: 22.13% Two year: 24.26%
Tenure	Mean: 31.73 Std. deviation: 24.42	Mean: 32.84 Std. deviation: 24.83	Mean: 34.80 Std. deviation: 24.44
Monthly charges	Mean: 64.61 Std. deviation: 29.90	Mean: 64.41 Std. deviation: 30.02	Mean: 66.72 Std. deviation: 31.38
Total charges	Mean: 2218.25 Std. deviation: 2237.5	Mean: 2319.13 Std. deviation: 2278.4	Mean: 2565.79 Std. deviation: 2381.8

MODEL TRAINING AND TESTING

Decision Trees

Decision trees are good at identifying potential predictors that can be used in other models. Their aim is to resolve a dependent variable from a series of independent variables.

Looking at the split report given by the decision tree, we can see that the variable “Contract” plays a vital role in predicting the dependent variable as suggested by its entropy value. The higher the entropy of a particular variable, the lower the information gained from that variable.

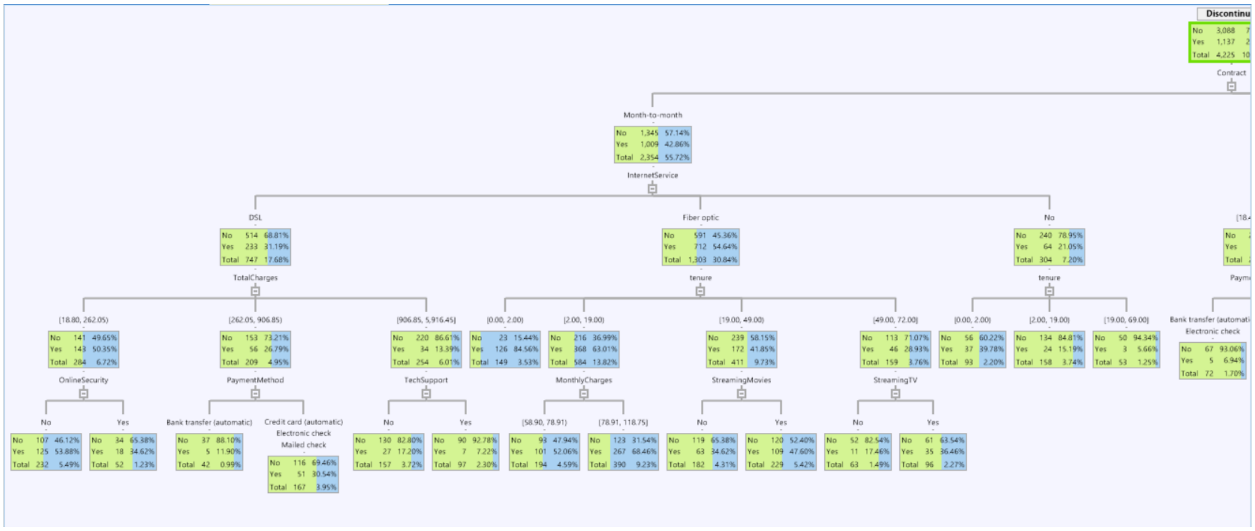


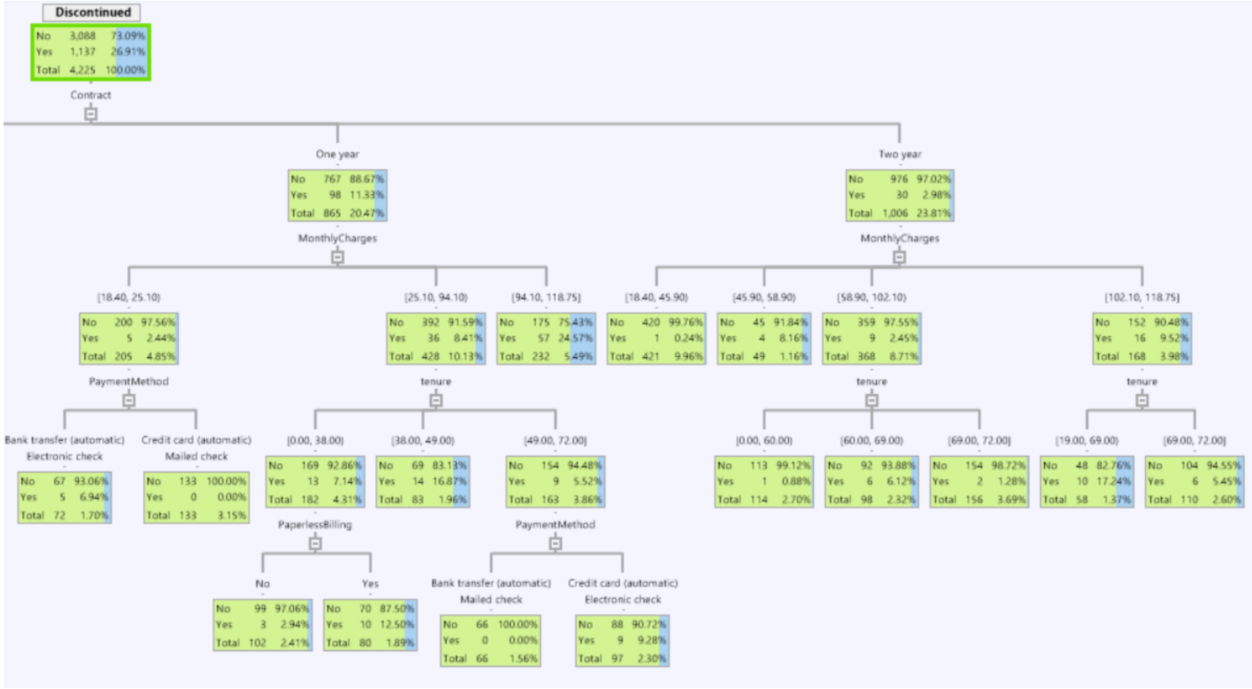
Insert Decision Tree - Split Report

	Variable Name	Rank	Info	Include	info value	input entropy	output entropy △	ratio entropy
▶	Contract	1	16.75868	☑	1.21256	0.84019	0.69938	0.16759
	tenure	2	12.59267	☑	0.80747	0.84019	0.73439	0.12593
	TechSupport	3	11.0053	☑	0.70289	0.84019	0.74772	0.11005
	OnlineSecurity	4	10.66243	☑	0.68257	0.84019	0.7506	0.10662
	InternetService	5	9.8985	☑	0.63153	0.84019	0.75702	0.09898
	OnlineBackup	6	8.36715	☑	0.54133	0.84019	0.76989	0.08367
	DeviceProtection	7	7.66751	☑	0.50056	0.84019	0.77577	0.07668
	PaymentMethod	8	7.00115	☑	0.41379	0.84019	0.78137	0.07001
	MonthlyCharges	9	5.67902	☑	0.36337	0.84019	0.79247	0.05679
	StreamingTV	10	5.48697	☑	0.37525	0.84019	0.79409	0.05487
	TotalCharges	11	5.37844	☑	0.31858	0.84019	0.795	0.05378
	StreamingMovies	12	5.35491	☑	0.36776	0.84019	0.7952	0.05355
	PaperlessBilling	13	3.1693	☑	0.19309	0.84019	0.81356	0.03169
	Dependents	14	2.66133	☑	0.1654	0.84019	0.81783	0.02661
	Partner	15	1.90953	☑	0.1143	0.84019	0.82414	0.0191
	SeniorCitizen	16	1.53783	☑	0.08883	0.84019	0.82727	0.01538

Entropy variance is used as the measure for the decision tree splits at each level. The split search method is cluster. Automatic tree generation is selected for building the tree.

Explaining the decision Tree


Discontinu
No 3,088 7
Yes 1,137 2
Total 4,225 10
Contract
Month-to-month
No 1,345 57.14%
Yes 1,009 42.86%
Total 2,354 55.72%
InternetService
DSL
No 514 68.81%
Yes 233 31.19%
Total 747 17.68%
TotalCharges
Fiber optic
No 591 45.36%
Yes 712 54.64%
Total 1,303 30.84%
tenure
No
No 240 78.95%
Yes 64 21.05%
Total 304 7.20%
tenure
[18.4
Payme
[18.80, 262.05)
No 141 49.65% Yes 143 50.35% Total 284 6.72%
OnlineSecurity
[262.05, 906.85)
No 153 73.21% Yes 56 26.79% Total 209 4.95%
PaymentMethod
[906.85, 5,916.45]
No 220 86.61% Yes 34 13.39% Total 254 6.01%
TechSupport
[0.00, 2.00)
No 23 15.44% Yes 126 84.56% Total 149 3.53%
[2.00, 19.00)
No 216 36.99% Yes 368 63.01% Total 584 13.82%
MonthlyCharges
[19.00, 49.00)
No 239 58.15% Yes 172 41.85% Total 411 9.73%
StreamingMovies
[49.00, 72.00]
No 113 71.07% Yes 46 28.93% Total 159 3.76%
StreamingTV
[0.00, 2.00)
No 56 60.22% Yes 37 39.78% Total 93 2.20%
[2.00, 19.00)
No 134 84.81% Yes 24 15.19% Total 158 3.74%
[19.00, 69.00]
No 50 94.34% Yes 3 5.66% Total 53 1.25%
Bank transfer (automatic) Electronic check
No 67 93.06% Yes 5 6.94% Total 72 1.70%
No
No 107 46.12% Yes 125 53.88% Total 232 5.49%
Yes
No 34 65.38% Yes 18 34.62% Total 52 1.23%
Bank transfer (automatic)
No 37 88.10% Yes 5 11.90% Total 42 0.99%
Credit card (automatic) Electronic check Mailed check
No 116 69.46% Yes 51 30.54% Total 167 3.95%
No
No 130 82.80% Yes 27 17.20% Total 157 3.72%
Yes
No 90 92.78% Yes 7 7.22% Total 97 2.30%
[58.90, 78.91)
No 93 47.94% Yes 101 52.06% Total 194 4.59%
[78.91, 118.75]
No 123 31.54% Yes 267 68.46% Total 390 9.23%
No
No 119 65.38% Yes 63 34.62% Total 182 4.31%
Yes
No 120 52.40% Yes 109 47.60% Total 229 5.42%
No
No 52 82.54% Yes 11 17.46% Total 63 1.49%
Yes
No 61 63.54% Yes 35 36.46% Total 96 2.27%


Discontinued
No 3,088 73.09%
Yes 1,137 26.91%
Total 4,225 100.00%
Contract
One year
No 767 88.67%
Yes 98 11.33%
Total 865 20.47%
MonthlyCharges
Two year
No 976 97.02%
Yes 30 2.98%
Total 1,006 23.81%
MonthlyCharges
[18.40, 25.10)
No 200 97.56% Yes 5 2.44% Total 205 4.85%
PaymentMethod
[25.10, 94.10)
No 392 91.59% Yes 36 8.41% Total 428 10.13%
tenure
[94.10, 118.75]
No 175 75.43% Yes 57 24.57% Total 232 5.49%
[18.40, 45.90)
No 420 99.76% Yes 1 0.24% Total 421 9.96%
[45.90, 58.90)
No 45 91.84% Yes 4 8.16% Total 49 1.16%
[58.90, 102.10)
No 359 97.55% Yes 9 2.45% Total 368 8.71%
tenure
[102.10, 118.75]
No 152 90.48% Yes 16 9.52% Total 168 3.98%
tenure
Bank transfer (automatic) Electronic check
No 67 93.06% Yes 5 6.94% Total 72 1.70%
Credit card (automatic) Mailed check
No 133 100.00% Yes 0 0.00% Total 133 3.15%
[0.00, 38.00)
No 169 92.86% Yes 13 7.14% Total 182 4.31%
PaperlessBilling
[38.00, 49.00)
No 69 83.13% Yes 14 16.87% Total 83 1.96%
[49.00, 72.00]
No 154 94.48% Yes 9 5.52% Total 163 3.86%
PaymentMethod
[0.00, 60.00)
No 113 99.12% Yes 1 0.88% Total 114 2.70%
[60.00, 69.00)
No 92 93.88% Yes 6 6.12% Total 98 2.32%
[69.00, 72.00]
No 154 98.72% Yes 2 1.28% Total 156 3.69%
[19.00, 69.00)
No 48 82.76% Yes 10 17.24% Total 58 1.37%
[69.00, 72.00]
No 104 94.55% Yes 6 5.45% Total 110 2.60%
No
No 99 97.06% Yes 3 2.94% Total 102 2.41%
Yes
No 70 87.50% Yes 10 12.50% Total 80 1.89%
Bank transfer (automatic) Mailed check
No 66 100.00% Yes 0 0.00% Total 66 1.56%
Credit card (automatic) Electronic check
No 88 90.72% Yes 9 9.28% Total 97 2.30%

The first split is done using the Contract variable. It can be observed that 97% of the two-year contracts and 89% of the one-year contracts are not discontinued. But, 43% of the month-month contracts are discontinued.

In the 43% of the contracts that are discontinued, the majority of the users having no internet connection seem to be satisfied with the service having just a 21% rate of discontinuation. Users using fiber optics, on the other hand, have a higher rate of discontinuation with 54%. Within the first two months, 45% of the users who used fiber optics discontinued. Users using DSL, have a comparatively lower rate of discontinuation at 31% but the ones who used it the least with very low total charges discontinued their service.

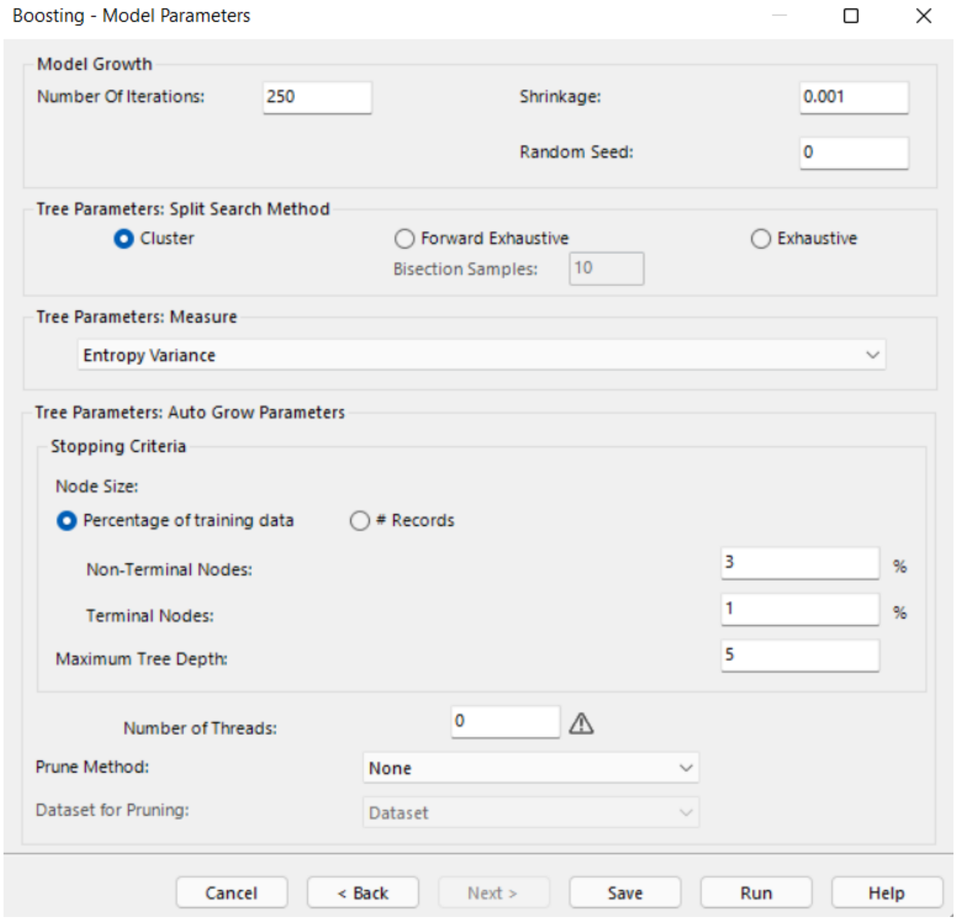
Bagging, Boosting and Random Forests - Ensemble methods

Boosting works by recursively training a particular tree and correcting the errors made by that tree. When correcting the errors made by the initial tree, we create another tree that can classify the records correctly.

Bagging iteratively takes a subset of the dataset and creates a large number of trees. During prediction, they take an average of the predictions made by each of the trees.

Random forest is similar to bagging but it also takes a subset of variables for each of the trees.

For boosting, similar parameters to whatever was given for the decision tree algorithm is given. That is, the split is taken into consideration by Entropy variance and the split search method is cluster. The number of iterations for the model to improve upon its previous tree's failures is set to 250.



Boosting - Model Parameters

Model Growth

Number Of Iterations: 250

Shrinkage: 0.001

Random Seed: 0

Tree Parameters: Split Search Method

● Cluster ○ Forward Exhaustive ○ Exhaustive

Bisection Samples: 10

Tree Parameters: Measure

Entropy Variance

Tree Parameters: Auto Grow Parameters

Stopping Criteria

Node Size:

● Percentage of training data ○ # Records

Non-Terminal Nodes: 3 %

Terminal Nodes: 1 %

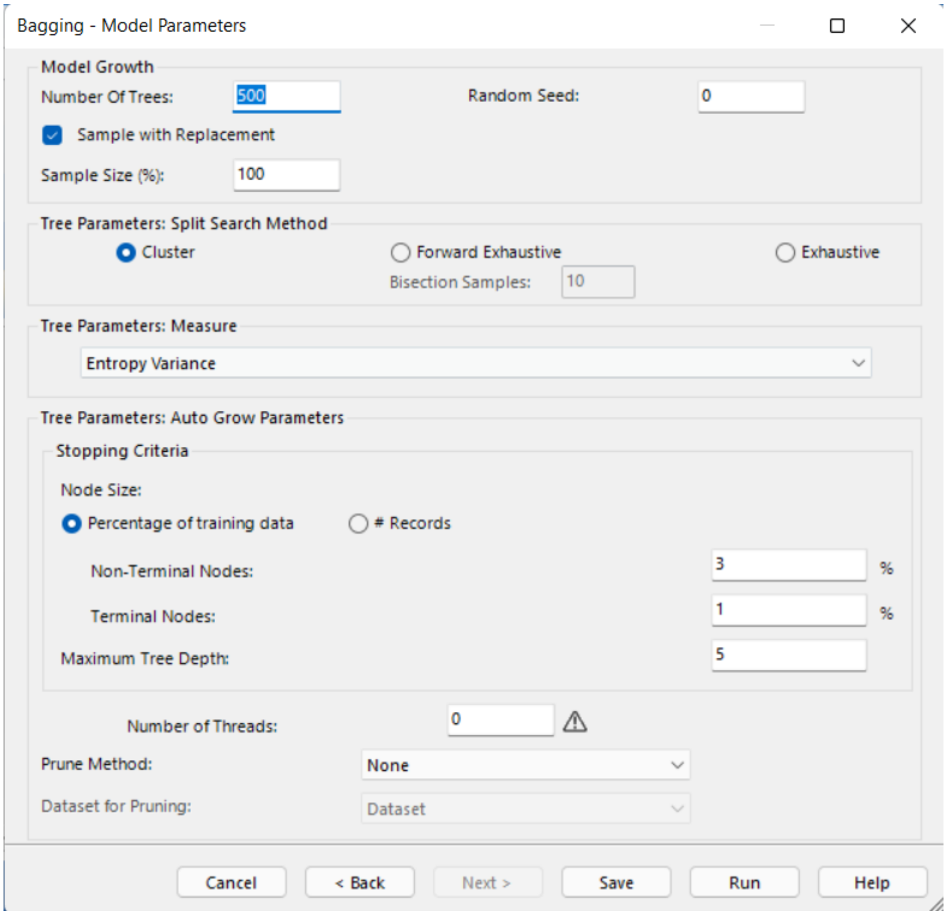
Maximum Tree Depth: 5

Number of Threads: 0

Prune Method: None

Dataset for Pruning: Dataset

Similar properties were set for the bagging model, on top of that, sampling with replacement is checked. Sampling with replacement allows us to reduce the variance without increasing the bias.



Bagging - Model Parameters

Model Growth

Number Of Trees: 500

Random Seed: 0

☑ Sample with Replacement

Sample Size (%): 100

Tree Parameters: Split Search Method

● Cluster ○ Forward Exhaustive ○ Exhaustive

Bisection Samples: 10

Tree Parameters: Measure

Entropy Variance

Tree Parameters: Auto Grow Parameters

Stopping Criteria

Node Size:

● Percentage of training data ○ # Records

Non-Terminal Nodes: 3 %

Terminal Nodes: 1 %

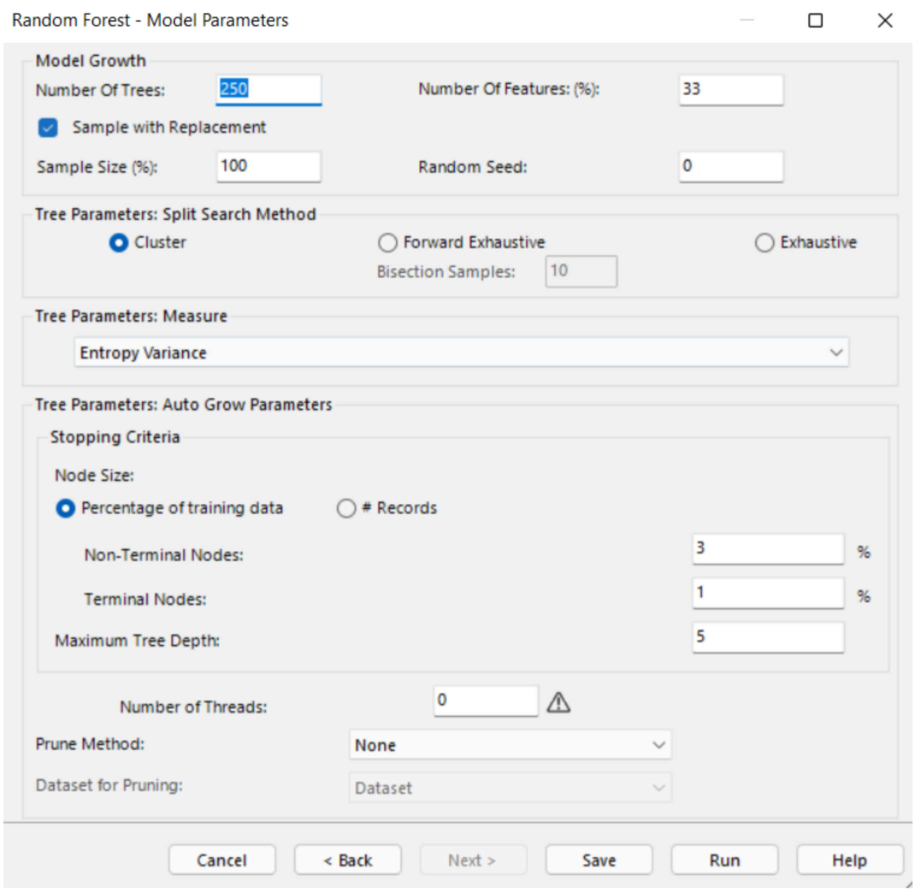
Maximum Tree Depth: 5

Number of Threads: 0

Prune Method: None

Dataset for Pruning: Dataset

As previously mentioned, in random forests, along with the training data, the features are sampled as well. The percentage of features that are sampled is given as 33% for our random forest model.


Random Forest - Model Parameters

Model Growth

Number Of Trees: 250

Number Of Features: (%): 33

☑ Sample with Replacement

Sample Size (%): 100

Random Seed: 0

Tree Parameters: Split Search Method

● Cluster ○ Forward Exhaustive ○ Exhaustive

Bisection Samples: 10

Tree Parameters: Measure

Entropy Variance

Tree Parameters: Auto Grow Parameters

Stopping Criteria

Node Size:

● Percentage of training data ○ # Records

Non-Terminal Nodes: 3 %

Terminal Nodes: 1 %

Maximum Tree Depth: 5

Number of Threads: 0

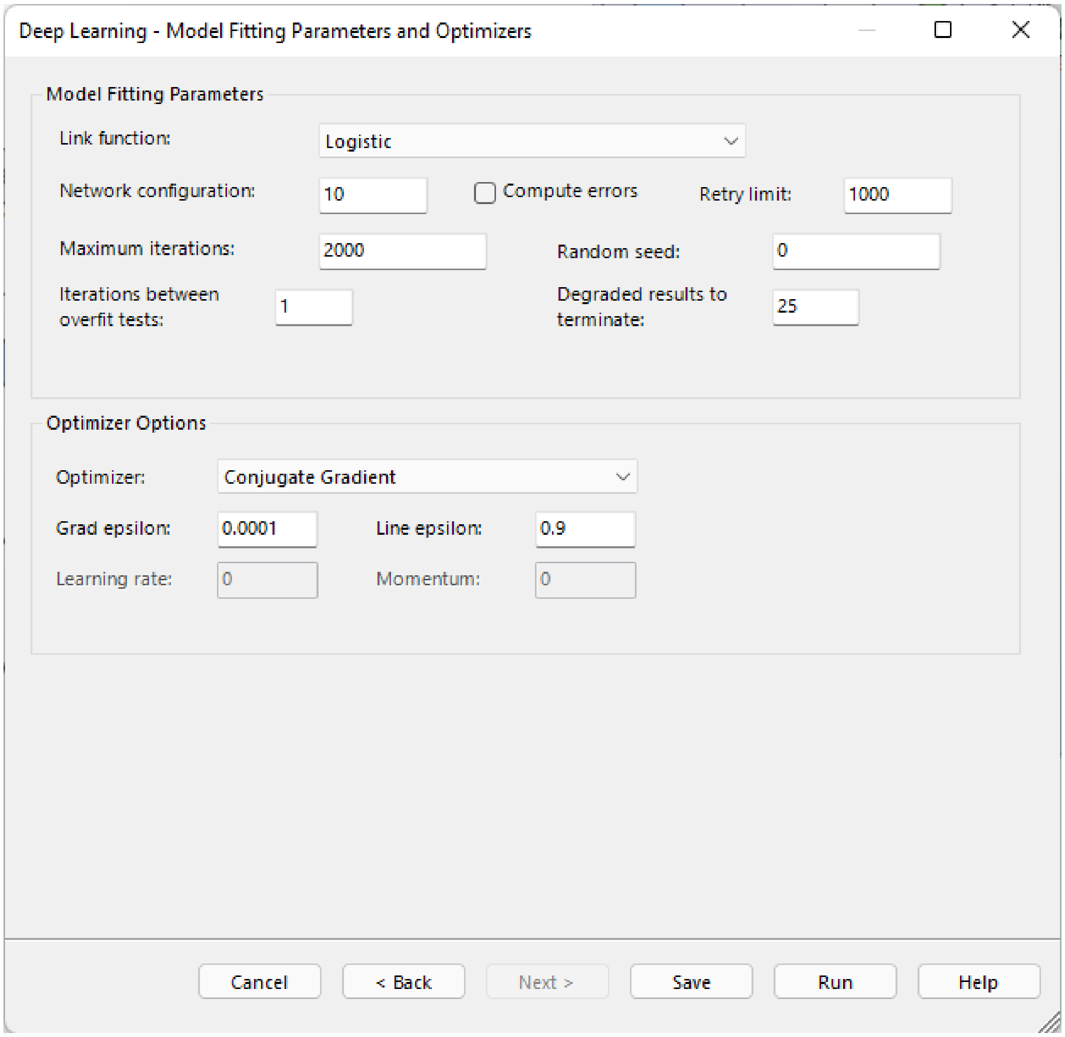
Prune Method: None

Dataset for Pruning: Dataset

Deep Learning

While using the deep learning node in Knowledge studio, we use the test dataset - Dataset_fit, to guard against the model overfitting on the training dataset. Overfitting happens when the model trains too much on the training data to a point where it memorizes it which leads to poor generalization. A model that has overfitted on a training dataset will perform poorly on the testing dataset.

We select all the variables to train the deep learning model. A singular value threshold is set by default, which helps in finding collinearity and excluding variables that are collinear.



Deep Learning - Model Fitting Parameters and Optimizers

Model Fitting Parameters

Link function: Logistic

Network configuration: 10

□ Compute errors

Retry limit: 1000

Maximum iterations: 2000

Random seed: 0

Iterations between overfit tests: 1

Degraded results to terminate: 25

Optimizer Options

Optimizer: Conjugate Gradient

Grad epsilon: 0.0001

Line epsilon: 0.9

Learning rate: 0

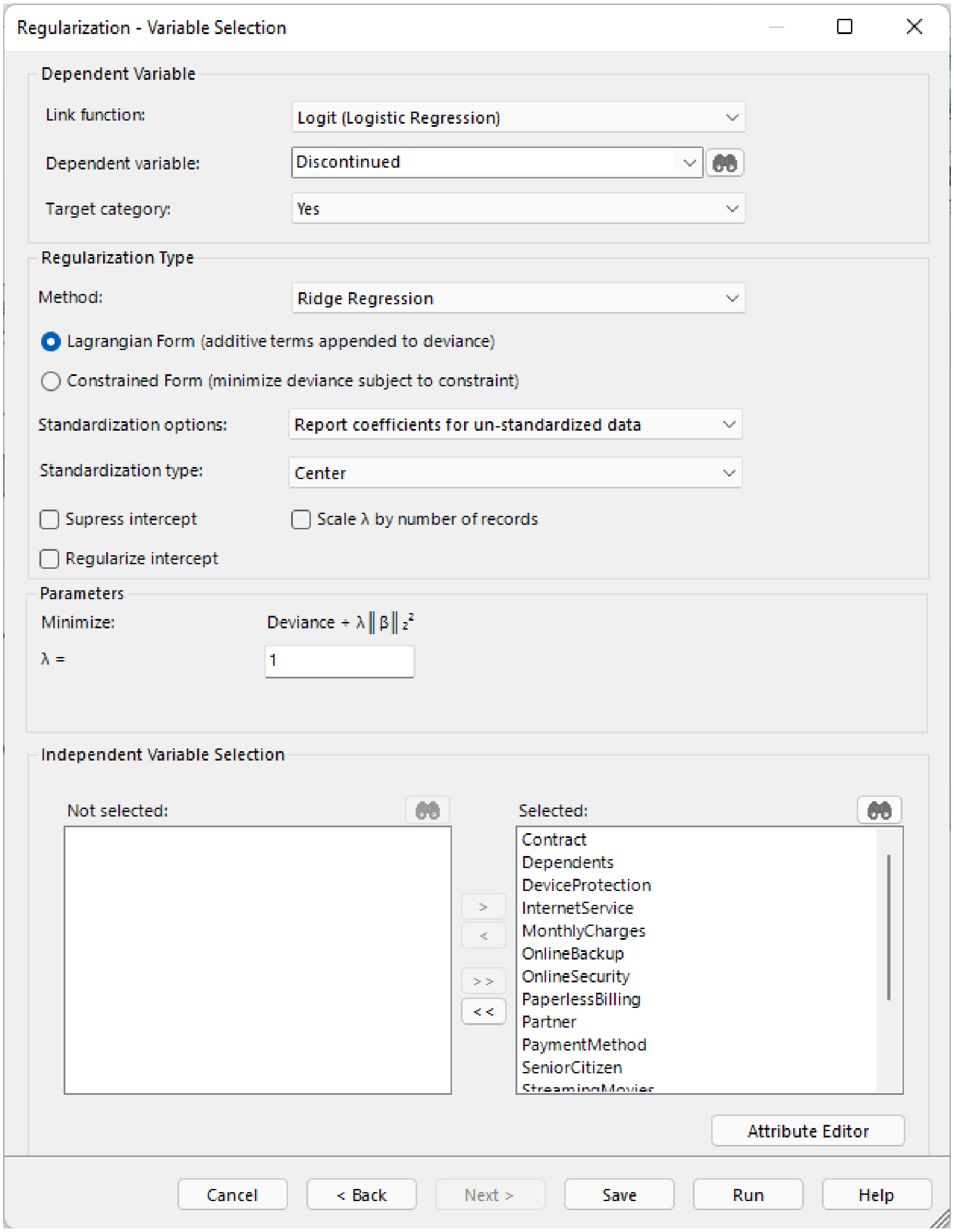
Momentum: 0

The link function is set to logistic which helps in non-linear transformation. The optimizer is set to conjugate gradient, I have no idea what it is.

Regularized Regression

To overcome overfitting, a regularization parameter is added to a model's loss function which penalizes weights and drives them towards zero. L1 regularization can even drive weights towards zero.

We are going to use a ridge regression model, ridge regression uses L2 regularization which squares the weights and adds them to the loss function thereby penalizing them.



Regularization - Variable Selection

Dependent Variable

Link function: Logit (Logistic Regression)

Dependent variable: Discontinued

Target category: Yes

Regularization Type

Method: Ridge Regression

● Lagrangian Form (additive terms appended to deviance)

○ Constrained Form (minimize deviance subject to constraint)

Standardization options: Report coefficients for un-standardized data

Standardization type: Center

□ Supress intercept □ Scale λ by number of records

□ Regularize intercept

Parameters

Minimize: $\text{Deviance} + \lambda \|\beta\|_2^2$

λ = 1

Independent Variable Selection

Not selected:

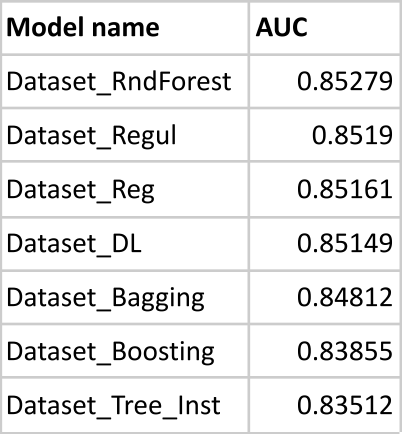
Selected:

Contract
Dependents
DeviceProtection
InternetService
MonthlyCharges
OnlineBackup
OnlineSecurity
PaperlessBilling
Partner
PaymentMethod
SeniorCitizen
StreamingMovies

Train and Test accuracies

Model Selection

For selecting the best model from the previously explained set of models in which we trained our dataset, we use the Model Selector node in Knowledge Studio. We choose the evaluation metric as AUC (Area under the ROC curve). AUC is commonly used to determine the performance of classification algorithms.



Model name	AUC
Dataset_RndForest	0.85279
Dataset_Regul	0.8519
Dataset_Reg	0.85161
Dataset_DL	0.85149
Dataset_Bagging	0.84812
Dataset_Boosting	0.83855
Dataset_Tree_Inst	0.83512

From the table above, we can see that the Random Forest model outperformed the other models but the difference in model performance is pretty low.

CONCLUSION AND IMPROVISATION

We have implemented all different models in Altair Knowledge Studio to predict the customers who are likely to discontinue the Telecom company's services. We have achieved a model with 85% accuracy. Knowledge Studio's AutoML features came in very handy and helped us implement all the models without any hassle.

To provide better results, the company can do one or more of the following

- Provide more data points to research
- Survey customers on why they are satisfied or dissatisfied

From the study, we can conclude that

- Providing additional services help in retaining customers
- A longer term contract up front might with incentives can help retain customers