INTRODUCTION

Churning Analysis is the process of using data to understand why your customers have stopped using your product or Services[1].

Churn shows customers who stop doing business with a company or a particular service[1].

What most businesses could do is try to understand the reason behind churn and tackle those factors, with responsive strategy[1].

There are various reasons a customer would decide to cancel a particular service or wants to stop doing business with a company, this ranges from poor service quality, poor customer support, change in prices and market competition[1]. There is really no particular reason for this to happen, but a combination of various reasons that has ended up in the customer's dissatisfaction[1].

Given Data: Telco-Customer Data

In the given data we have a total of 7043 Customers and 21 attributes out of which 5174 are Satisfied (not churning) while 1869 are Dissatisfied (churning), this shows that the data is unbalanced.

The focused attribute for this analysis is the Churn.

Keywords: churn, attributes, ca (correspondence analysis)

DEFINITION OF ATTRIBUTES

customerID — Customer ID[4]

Gender — This is whether the customer is a male or a female[4]

SeniorCitizen — This is whether the customer is a senior citizen or not (1, 0) [4]

Partner — This is whether the customer has a partner or not (Yes, No) [4]

Dependents — This is whether the customer has dependents or not (Yes, No) [4]

Tenure Month — Number of months the customer has stayed with the company[4]

PhoneService — This is whether the customer has a phone service or not (Yes, No) [4]

MultipleLines — This is whether the customer has multiple lines (Yes, No, No phone service) [4]

InternetService — Customer's internet service provider (DSL, Fibre optic, No Internet). [4]

OnlineSecurity — This is whether the customer has online security (Yes, No, No internet service) [4]

OnlineBackup — This is whether the customer has online backup (Yes, No, No internet service) [4]

DeviceProtection — This is whether the customer has device protection (Yes, No, No internet service) [4]

TechSupport — This is whether the customer has tech support (Yes, No, No internet service) [4]

StreamingTV — This is whether the customer has streaming TV (Yes, No, No internet service) [4]

StreamingMovies — This whether the customer has streaming movies (Yes, No, No internet service) [4]

Contract — This shows the contract term of the customer (Month-to-month, One year, Two year) [4]

PaperlessBilling — This whether the customer has paperless billing (Yes, No) [4]

Payment Method — The customer's payment method (Electronic check, Mailed check, Bank transfer (automatic), Credit card (automatic)) [4]

Monthly Charges — The amount charged to the customer monthly[4]

Total Charges — The total amount charged to the customer

Churn — This whether the customer churned or not (Yes or No)

HYPOTHETICAL QUESTIONS TO CONSIDER

1. What Services are these customers dissatisfied (churning) about?

2. Why are they dissatisfied (churning) toward these Services?

3. How can the problem be fixed?

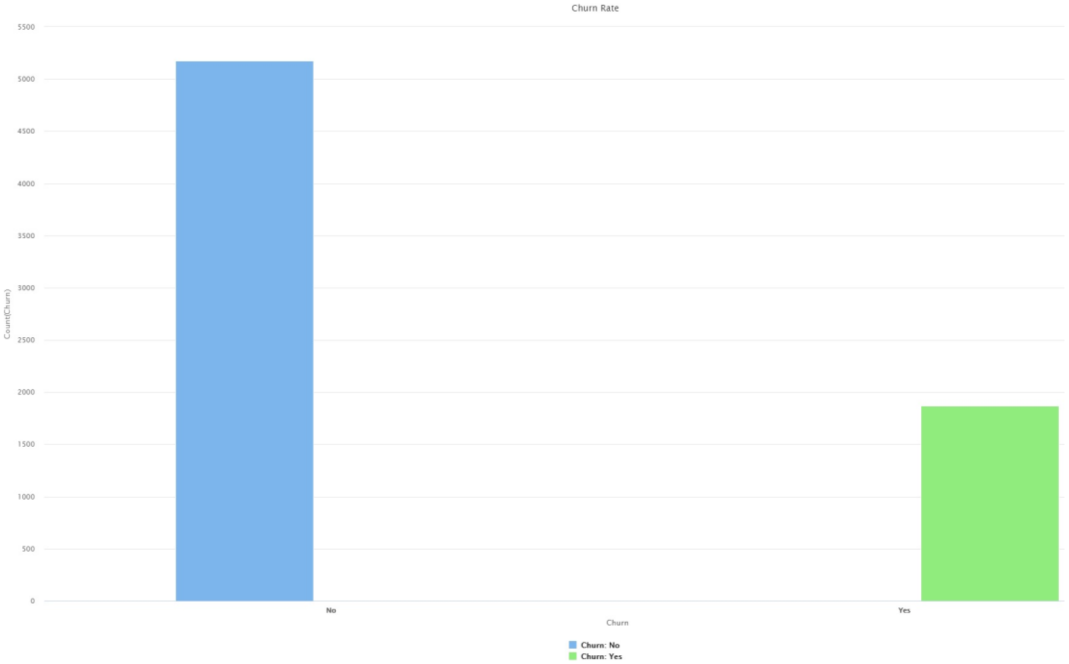
And to help find solution to the above questions, I will go through the following steps;

- ✓ Data Cleansing
- ✓ Data exploratory analysis
- ✓ Outlier Check
- ✓ Statistical hypothesis testing
- ✓ Model Evaluation

DATA CLEANSING

Data cleansing was carried out using the Turbo-Prep tool to remove rows with missing values. A total of 11 records were removed from the data set. We now have a total of 7,043 rows and 21 columns (17 nominal, 4 numerical

DATA EXPLORATORY ANALYSIS



The major purpose of this analysis is to identify the gaps that can yield to customer's churn with the rate at 73.5% (5174) {No} as against 26.5% (1869) {Yes} and also helps to answer the question of what services the customers are churning. I will be concentrating mainly on the churn portion of the data for the data exploratory analysis.

$$\textbf{Churn Rate:}\quad \frac{\text{No of churn for an Attribute}}{\text{Total no of churn in the data set}} \quad \text{X 100}$$

Looking at attributes like, Gender, Dependents, SeniorCitizen, Partner, Contract, PaperBilling, Payment Method, Phone services, Multiple lines, Internet service, Online Security, OnlineBackup, Device Protection, Tech Support, StreamingTV and StreamingMovie.

The count charts can provide some meaningful conclusions such as:

- **The Males and Females from our data have equal chances of churning, with the Males at 1 time less likely to churn.**

  In the data given, the number of Males are 3555, Males that are **churning are 930** while 2625 are not churning.

  The number of Females are 3488, Females that are **churning are 939** while 2549 are not churning.

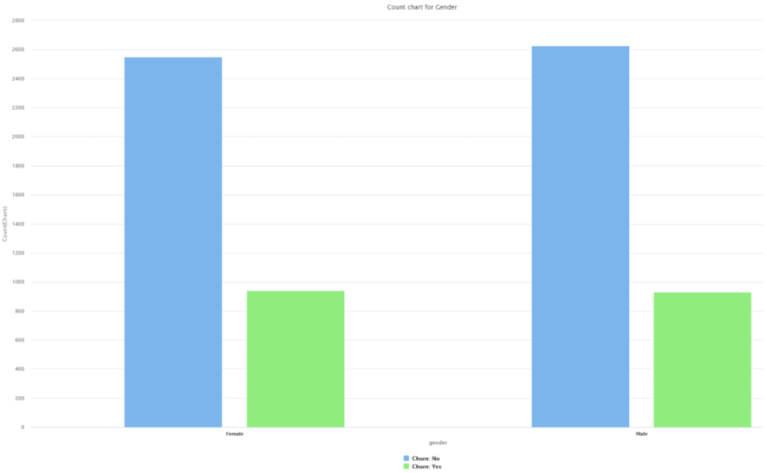
  The total number of churn is 1869.

  Churn rate for Males = no of churn for Males**/**Total no of Churn x 100

  =930/1869*100 = **49.8%**

  Churn rate for Females = no of Churn for Females/Total no of Churn x 100

  = 939/1869*100 = **50.2%**



- **Customers without dependents are five times more likely to churn**[3]

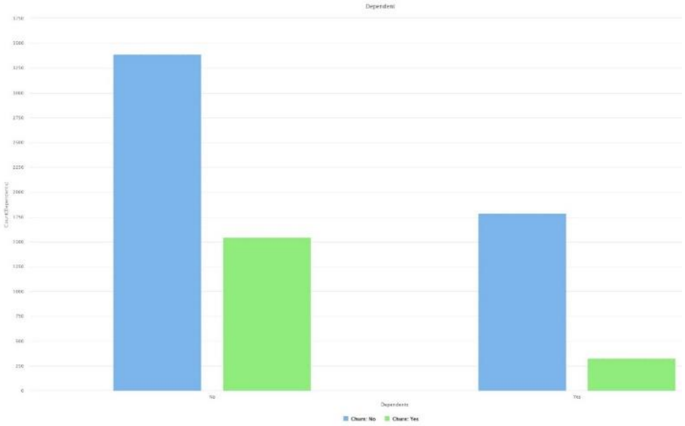
  Total Churn = 1869

  Dependent churn {Yes} = 326

  Non Dependent churn {No} = 1543

  Churn rate for dependents {Yes} = 326/1869*100 = 17.4%

  Churn rate for Non-dependents {No} = 1543/1869*100 = 82.6%



- **Senior citizens {1, Yes} are three times less likely to churn**[3]

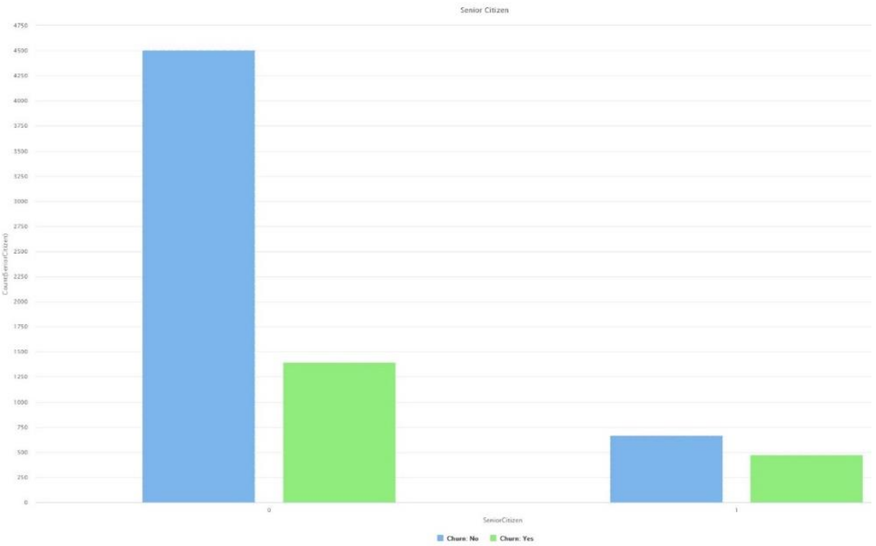
  Senior Citizen {1, Yes} Churn is 476

  Non Senior Citizen churn {0, No} is 1393

  Total churn for both Senior Citizen and Non Senior Citizen is 1869

  Churn rate for SeniorCitizen {1, Yes} = 476/1869*100 = **25.5%**

  Churn rate for Non senior Citizen {0, No} = 1393/1869*100 = **74.53%**



- **Partners are almost two times less likely to churn**[3]

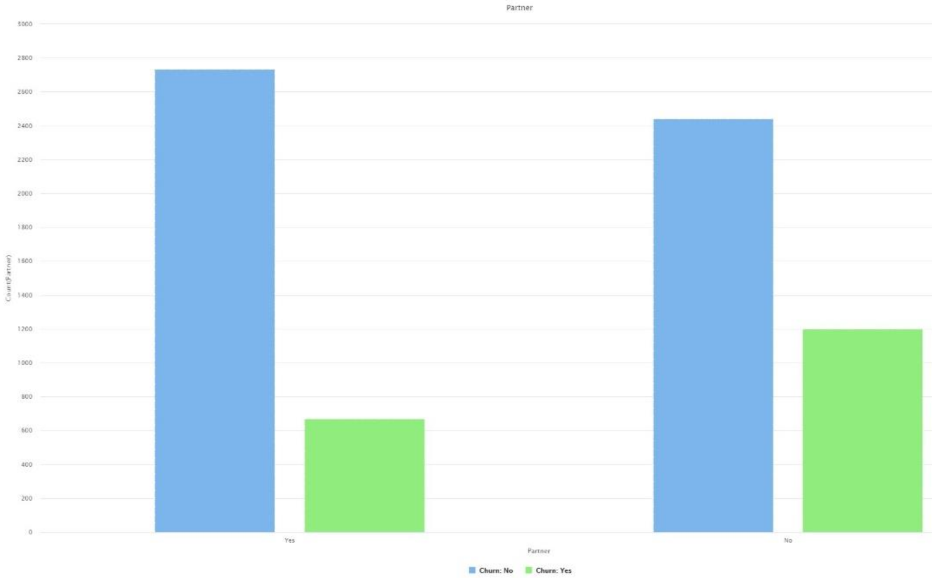
  Partner {Yes} Churn is 669

  Non Partner {No} churn is 1200

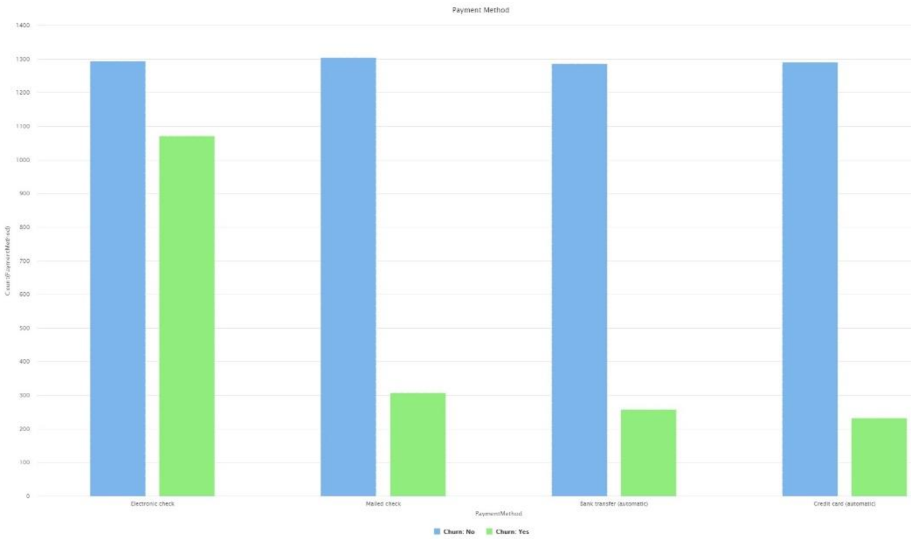
  Total churn is 1869

  Churn rate for Partner {Yes} = 669/1869*100 = **35.8%**

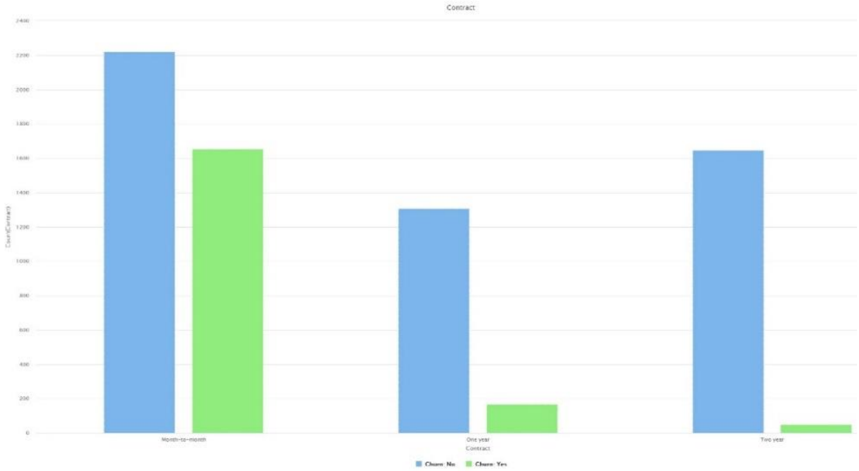
  Churn rate Non Partner {No} = 1200/1869*100 = **64.2%**

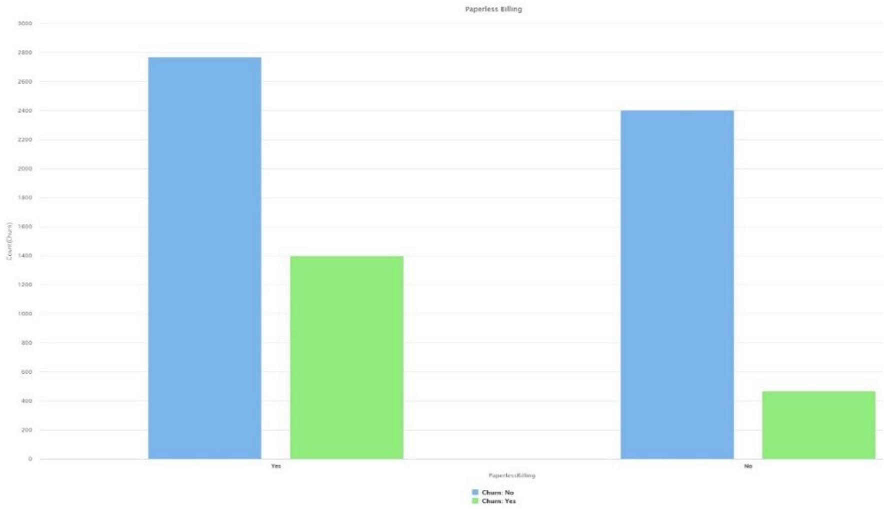


- **Also customers that have Payment Method “*Electronic Check”* are more likely to churn with a churn rate of 57.3%.**[3]

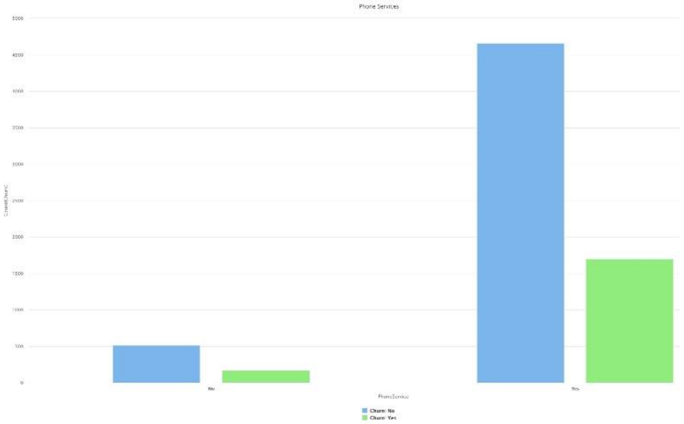


- **Most of the customers that cancel their subscription are on *Month-to-month* Contract with a churn rate of 88.6% and Paperless Billing with a churn rate of 74.9%.**[3]

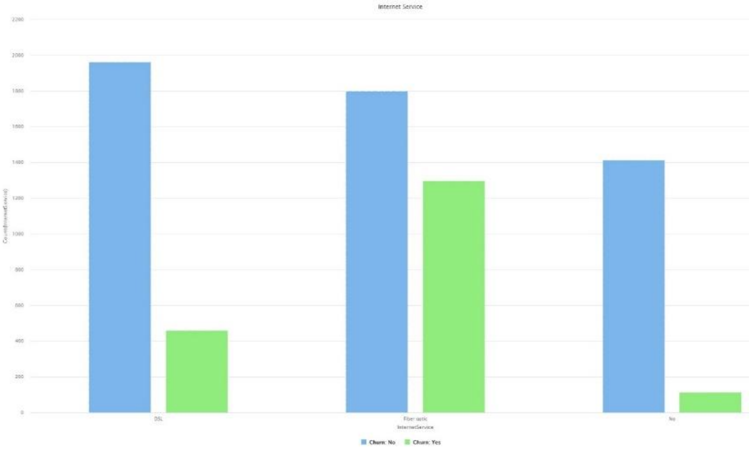




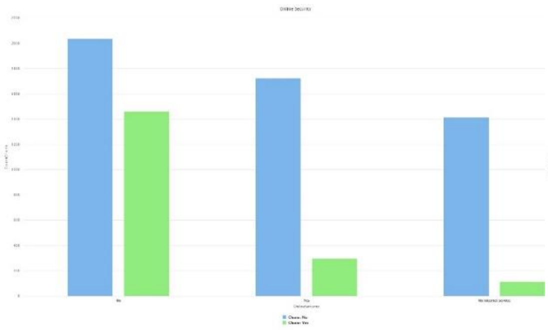
- **Most of the customers that cancel their subscription have Phone Service enabled with the churn rate at 90.9%.**[3]

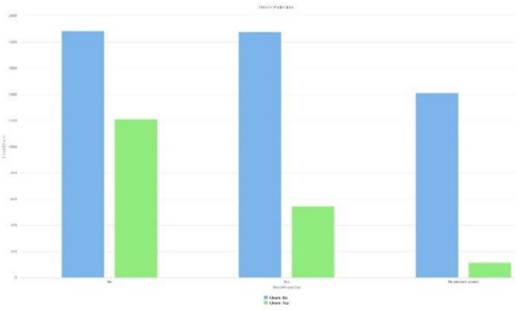


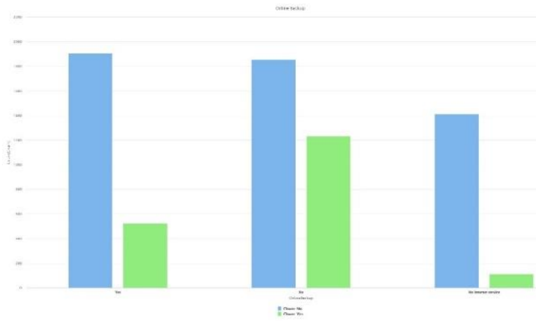
- **Customers that have Internet Service “Fiber-Optic” are more likely to cancel than those who have DSL with their churn rates at 69.4% and 24.6% respectively.** [3]

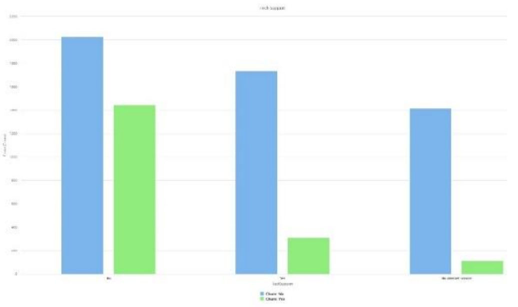


- **Customers that do not have Online Security, Device Protection, Online Backup, and Tech Support services enabled are more likely to leave with their churn rates at 78.2%, 64.8%, 66.0%, and 77.4% respectively**[3]

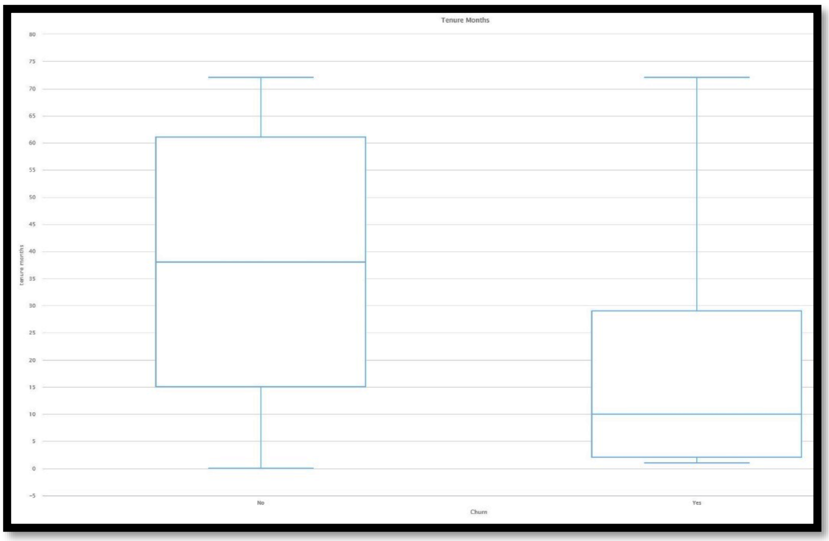




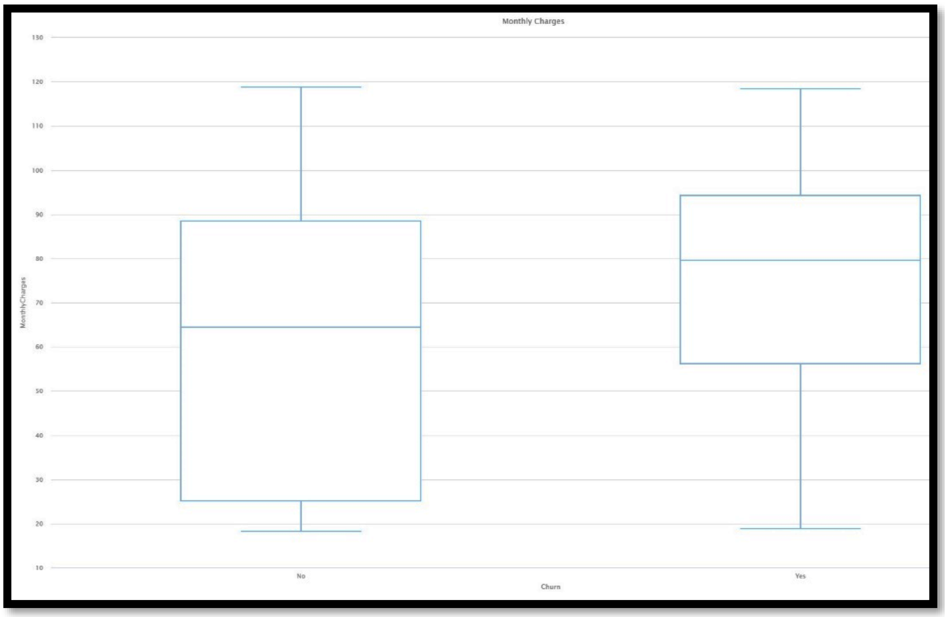




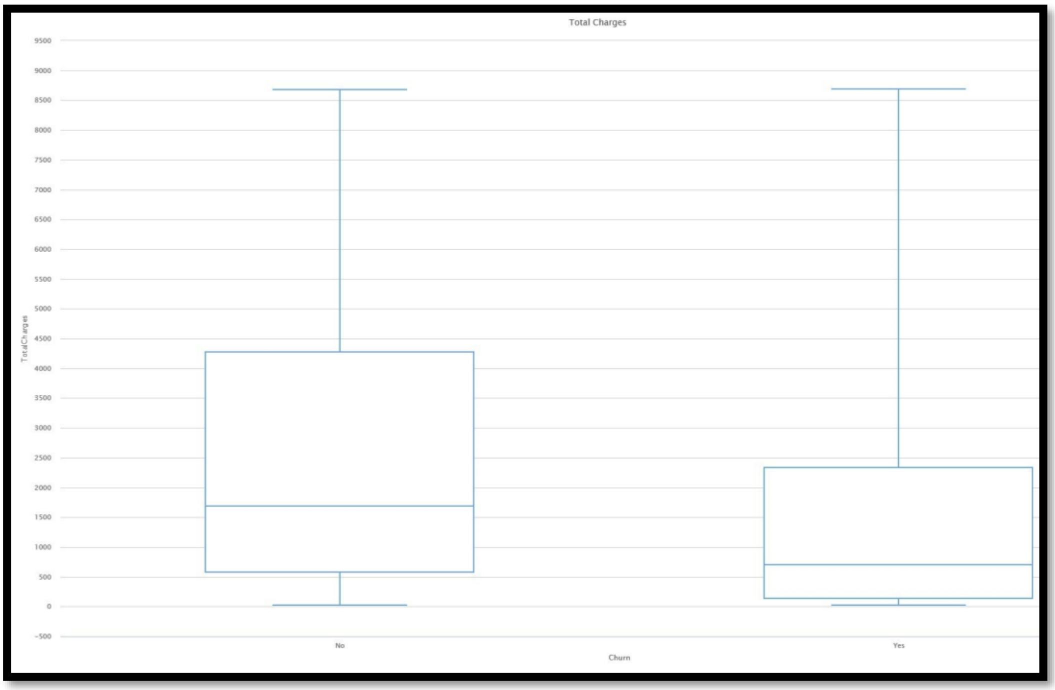
- **Churning customers have much lower tenure with a median of ca. 10 months compared to a median of non-churners of ca. 38 months as shown in the boxplots below.**[4]



- **Churning customers have higher monthly charges with a median of ca. US$80 and much lower interquartile range compared to that of non-churners median of ca. US$65 as shown in the boxplot below.** [4]



- **TotalCharges are the result of tenure and MonthlyCharges, which are more understanding on an individual basis as shown in the boxplots below.** [4]



OUTLIERS CHECK IN NUMERICAL DATA

No outliers in numerical data was found with the Distance-based Outlier Detection and so no adjustments done.

STATISTICAL HYPOTHESIS TESTING

Looking at the gaps between the attributes and connecting them to their potential influence on the customer's churn and at the same time answers the question 'why the customers are churning towards some services. The following hypothesis were made:

- The longer the contract duration the lesser the chance for the customer to churn as he/she is less persistently confronted with the termination and potentially values contracts with reduced effort.
- Customers are willing to cancel simple contracts with fewer additional services faster and more often than complex product packages. They might also be hesitant to cancel a contract when they depend on the additional service.
- Customers with spouses and children might churn less to keep the services running for their family.

- Number of additional services, tenure rate, payment method and contract duration is among the most important determinant of churn.
- More expensive contracts lead to increased churn as the chances to save money by changing providers might be higher.
- Senior citizens tend to churn less due to the long duration of their contracts.

MODEL EVALUATION

The Data models considered were:

- Naïve Bayes
- Generalized Linear Model
- Logistic Regression
- Fast Large Margin
- Deep Learning
- Decision Tree
- Random Forest
- Gradient Boasted Trees (XGBoost)
- Support Vector Machines

For Best Performance Assessment I chose the Logistics Regression Model, The Following Metrics Were Used:

- **Weights:** Indicates the top features used by the model to generate the predictions[4]
- **Confusion matrix:** Shows a grid of true and false predictions compared to the actual values[4]
- **Accuracy:** This shows the overall accuracy of the model. [4]
- **ROC Curve:** Shows the diagnostic ability of a model by bringing together true positive rate (TPR) and false positive rate (FPR) for different thresholds of class predictions. [4]
- **AUC (for ROC):** Measures the overall separability between classes of the model related to the ROC curve[4]
- **Precision:** Shows the diagnostic ability by comparing false positive rate (FPR) and false negative rate (FNR) for different thresholds of class predictions. It is suitable for

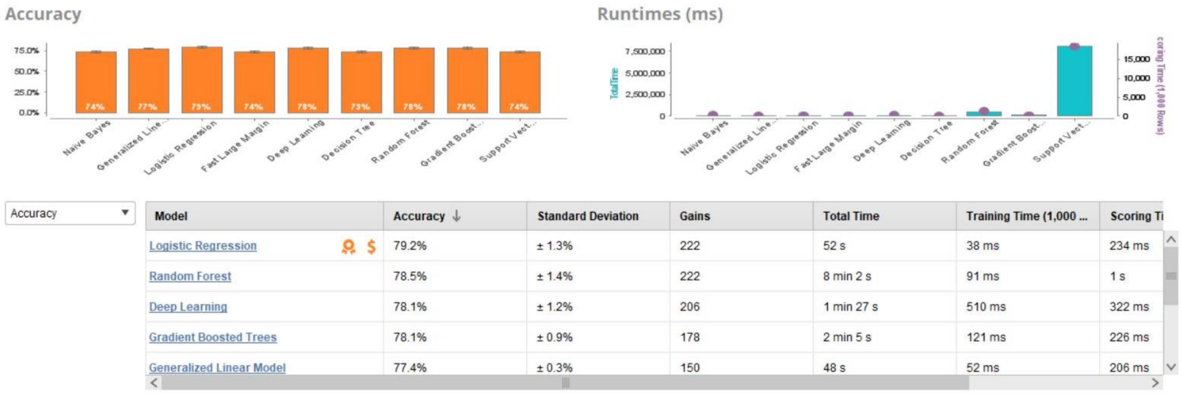
data sets with high class imbalances (negative values overrepresented) as it focuses on precision and recall, which are not dependent on the number of true negatives and thereby excludes the imbalance[4]

- **F Measure:** Builds the harmonic mean of precision and recall and thereby measures the compromise between both. [4]
- **AUC (for PRC):** Measures the overall separability between classes of the model related to the Precision-Recall curve. [4]

SUMMARY

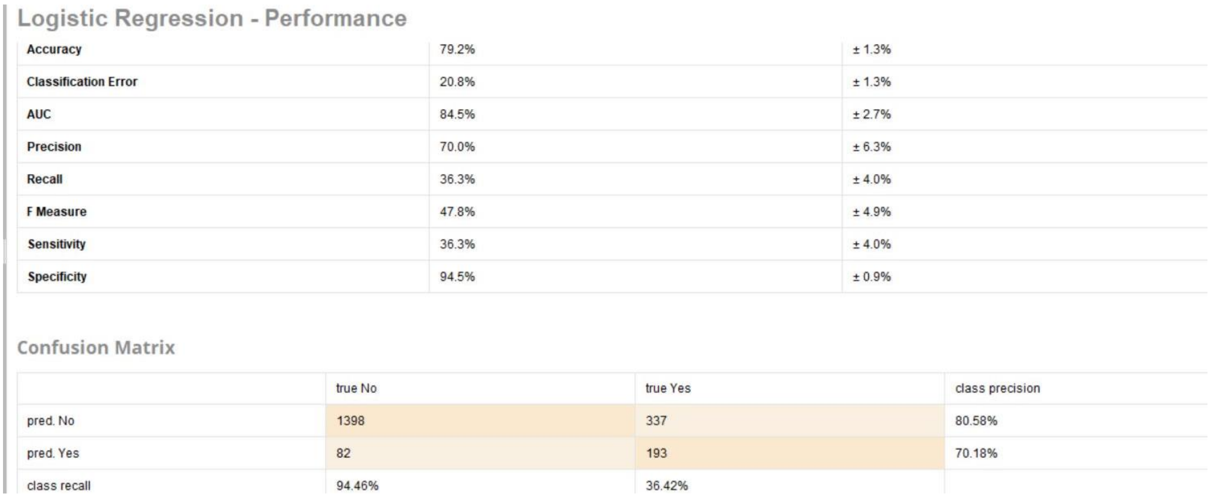
Model Summary

Looking at model results, the best accuracy on the test set is achieved by the Logistic Regression Model with 79.2% and was deployed in the analysis.



Model	Accuracy ↓	Standard Deviation	Gains	Total Time	Training Time (1,000 ...	Scoring Ti
Logistic Regression	79.2%	± 1.3%	222	52 s	38 ms	234 ms
Random Forest	78.5%	± 1.4%	222	8 min 2 s	91 ms	1 s
Deep Learning	78.1%	± 1.2%	206	1 min 27 s	510 ms	322 ms
Gradient Boosted Trees	78.1%	± 0.9%	178	2 min 5 s	121 ms	226 ms
Generalized Linear Model	77.4%	± 0.3%	150	48 s	52 ms	206 ms

Given the high imbalance of the data towards non-churners, it makes sense to compare F1 scores to get the model with the best score on jointly precision and recall. This would also be the neural network with a F1 score of 47.8%.



Logistic Regression - Performance

Accuracy	79.2%	± 1.3%
Classification Error	20.8%	± 1.3%
AUC	84.5%	± 2.7%
Precision	70.0%	± 6.3%
Recall	36.3%	± 4.0%
F Measure	47.8%	± 4.9%
Sensitivity	36.3%	± 4.0%
Specificity	94.5%	± 0.9%

Confusion Matrix

	true No	true Yes	class precision
pred. No	1398	337	80.58%
pred. Yes	82	193	70.18%
class recall	94.46%	36.42%	

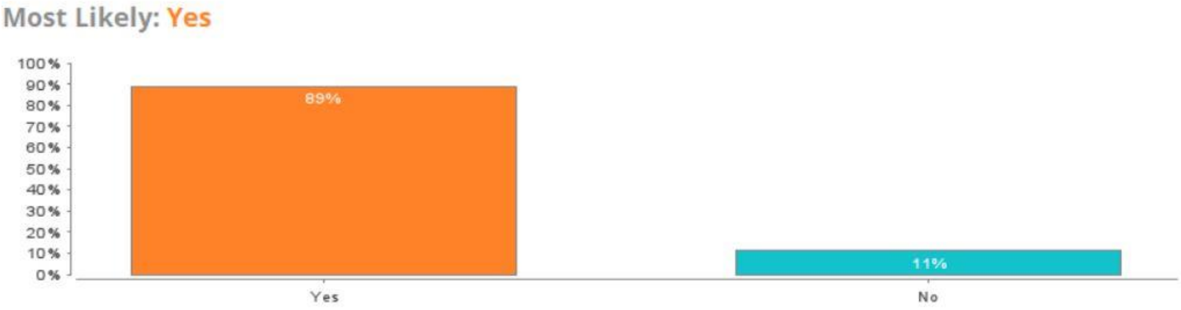
Given the scores of the best performing models, it can be observed that F Measure scores are not much above 50%. Further optimization efforts were carried out to achieve a higher scores and thereby increasing prediction power for more business value.

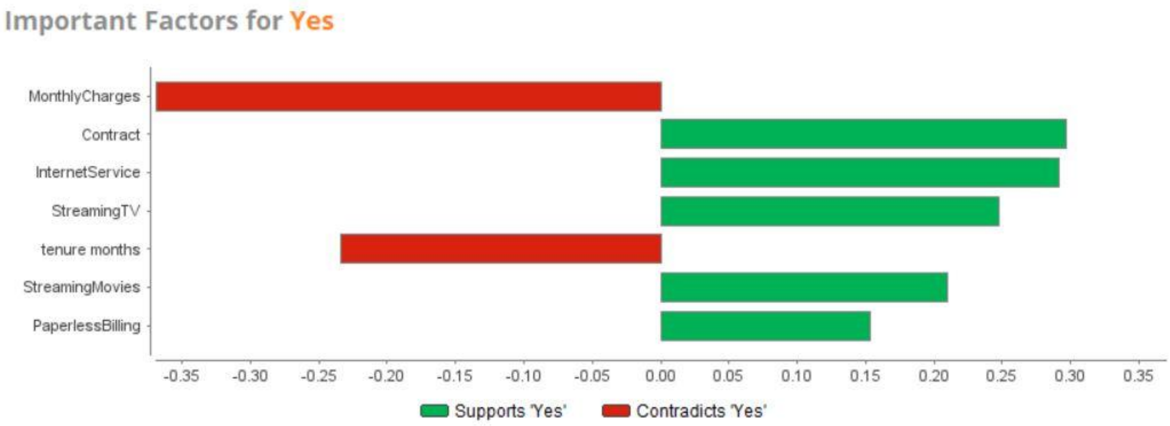
Statistical Hypothesis Check

The most important determinants for Customer churn are;

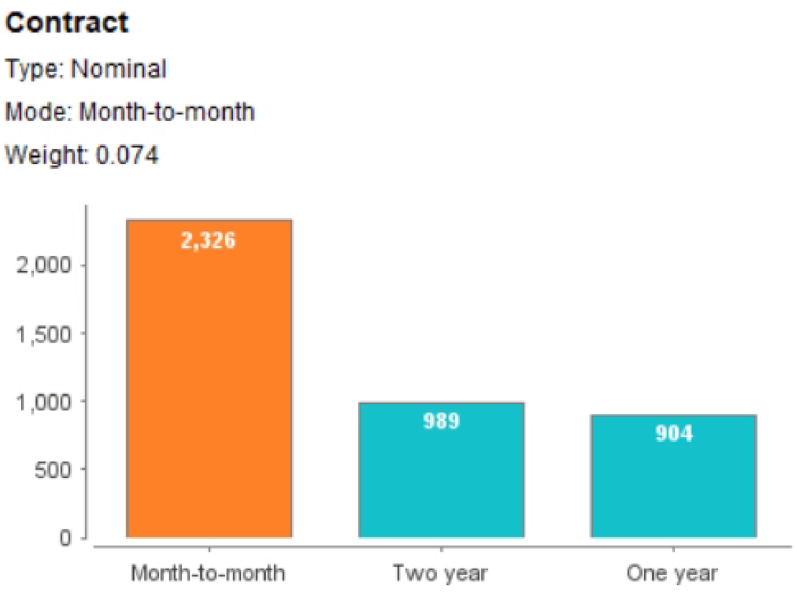
1. Contract Duration
2. Number of Additional Services
3. Dependent
4. Payment Method
5. Senior Citizen

Looking at the evaluation results, specifically the feature weights from the logistic regression, the hypotheses can be directionally supported or refused. [4]





Contract duration: Contract duration month-to-month is the second biggest driver of churn → supported[4]



Contract

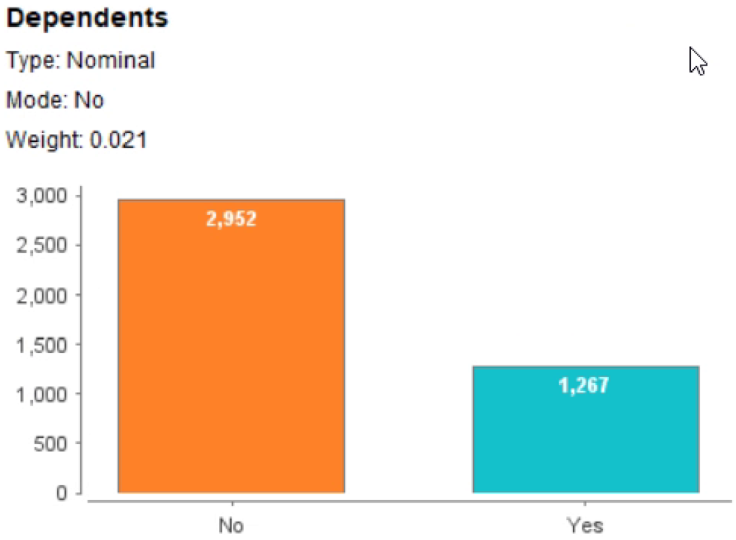
Type: Nominal

Mode: Month-to-month

Weight: 0.074

Number of additional services: This feature does not rank among the top features → refused[4]

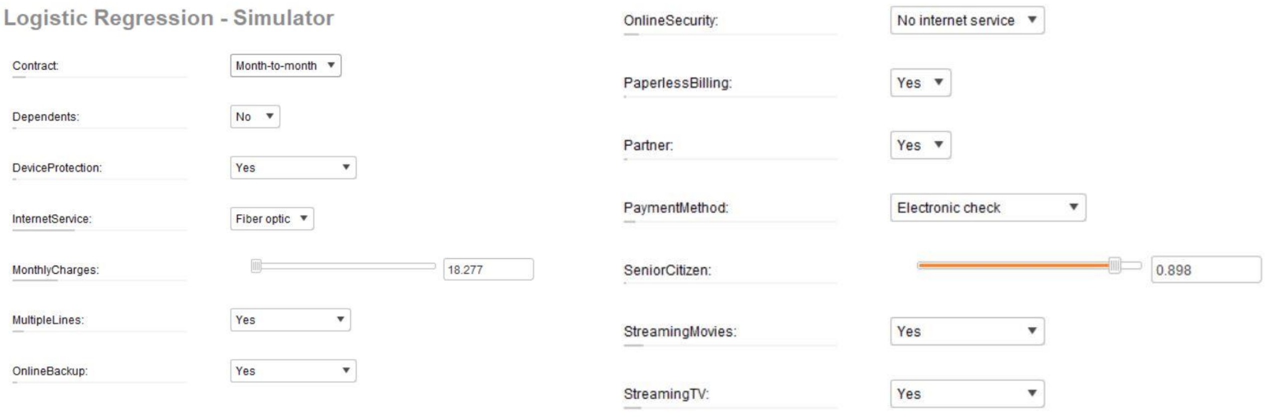
Partners and children: Having children ranks as the fourth feature that drives not churning, but strength is relatively low → partially supported[4]



Tenure: High tenure ranks as the strongest factor for not churning and the strongest feature overall. This is also supported by the boxplot in the Data Exploratory Analysis. → supported[4]

Monthly payment: Total payments, which is the product of tenure and monthly payment ranks as the strongest factor for churn. Indirectly, high monthly payments lead to churn. [4]

Senior citizens: Senior citizens does not have high feature weights. Also the ratio of senior citizens who churn is much higher than that of non-churners → refused[4]



Logistic Regression - Simulator

Contract: Month-to-month

Dependents: No

DeviceProtection: Yes

InternetService: Fiber optic

MonthlyCharges: 18.277

MultipleLines: Yes

OnlineBackup: Yes

OnlineSecurity: No internet service

PaperlessBilling: Yes

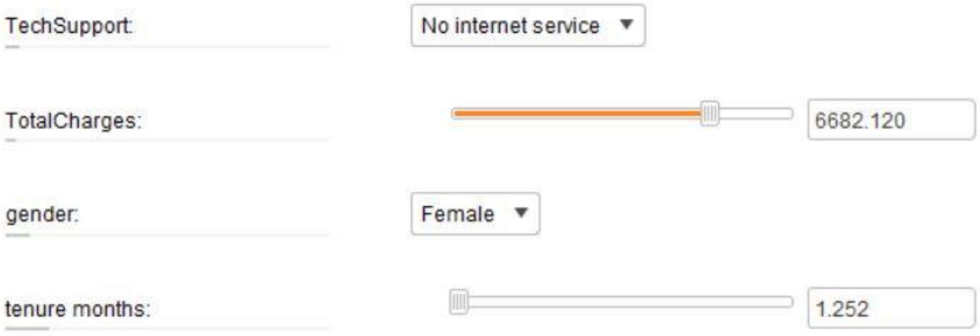
Partner: Yes

PaymentMethod: Electronic check

SeniorCitizen: 0.898

StreamingMovies: Yes

StreamingTV: Yes



TechSupport: No internet service

TotalCharges: 6682.120

gender: Female

tenure months: 1.252

CONCLUSION

Telcos typically have much more data available that could be included in the analysis, like extended customer and transaction data from customer relationship management systems and operational data around network services provided.[4] Also they typically have much larger amounts of churn/non-churn events at their disposal than the ca. 7043 in this case example.

A high accuracy is needed to be able to identify promising customer cases where churn can be avoided as, eventually, the customer returns protected need to outweigh the costs of related retention campaigns.[4]

If the goal is to engage and reach out to the customers to prevent them from churning, it's acceptable to engage with those who are mistakenly tagged as 'not churned,' as it does not cause any negative impact. It could potentially make them even happier with the service. [3]

This is the kind of model that can add value from day one if proper action is taken out of meaningful information it produces. [3]

No algorithm will predict churn with 100% accuracy. That's why it's important to test and understand the strengths and weaknesses of each classifier and get the best out of each.[3]

References

1. https://baremetrics.com/blog/churn-analysis Retrieved on 12th December, 2020
2. https://docs.rapidminer.com/9.0/studio/auto-model/ Retrieved on 12th December 2020
3. https://towardsdatascience.com/customer-churn-in-telecom-segment-5e49356f39e5 Retrieved on 15th December 2020
4. https://towardsdatascience.com/machine-learning-case-study-telco-customer-churn-prediction-bc4be03c9e1d Retrieved on 16th December 2020