39612295

1 Introduction

Predicting and analyzing customer churn is a crucial aspect of most businesses, in order to investigate and improve customer retention and retainment–overall increasing profits and customer satisfaction. Depending on the context, data, and purpose, different analysis techniques and applications can be conducted to dig into different key factors. Understanding which variables contribute to customer dissatisfaction can help enhance different aspects that may go unnoticed–for example, high rates of churn for a specific branch could help resolve branch specific issues, such as location or staff unfriendliness.

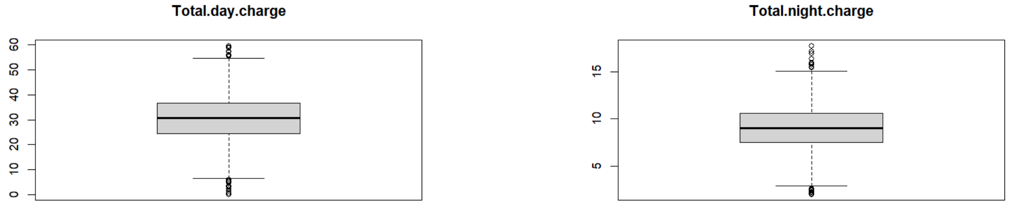
In this task, a telecom company wants to examine which factors might prompt a customer to stop using their services. This aids them in fixing and analyzing their business model, marketing strategies, or even different programs the company might have to offer. The dataset contains 20 variables, which are mixed between categorical and numerical variables, where the "Churn"variables, which is our target variables, is a categorical variable with a binary, True or False outcome. In order to identify the variables which are the most relevant and predictive of the churn, different applications and techniques will be applied to help derive and explain sensible results.

Different exploratory data analysis and visualization techniques will be investigated, including various plots for univariate analysis and visualization and for understanding distributions. The data will be investigated for missing data or the presence of any outliers, and different dimensionality reduction and feature extraction techniques–namely PCA and MDS, or principal component analysis and multidimensional scaling respectively, will be applied in order to investigate feature correlations and patterns, and benefit from them. Among this investigation, the results and findings will be used to construct, assess and moderate different statistical and predictive models–mainly being Logistic Regression, K-Nearest Neighbors (KNN), and Decision Trees. Different performance assessment and modeling selection techniques will be used, as well as different penalization or regularization techniques in attempt to optimize the models and obtain the best outcomes for predicting churn with optimized accuracy and performance.

Certain findings and principals will be highlighted and discussed which are relevant to data regarding predicting customer churn. One phenomena induced is the specificity v.s. sensitivity tradeoff, which is extremely common when dealing with data pertaining to a binary outcome on imbalanced data. The findings will be discussed and assessed in order to provide a holistic view on the customer churn analysis task at hand.

Exploratory Data Analysis

Upon inspecting the dataset, the first step is to investigate whether or not there are any missing values, duplicates or anomalies. It was found that are no missing values or duplicates present within the dataset. By plotting the boxplots for each numerical variable, there is a visual understanding of the distribution given the median and interquartile ranges. Each plot for each variable was investigated, and for each, there are some samples that fall outside of the range. Despite this, given the context of the dataset and the meaning or implication of each variable, for all, the values still fell in a reasonable range for the given variable. Thus, there were no anomalies deemed necessary of either imputation or removal. As a result, there were some outliers found given the IQR boxplot method, but no anomalies that needed to be handled.



(a) Total.day.charge Variable Boxplot

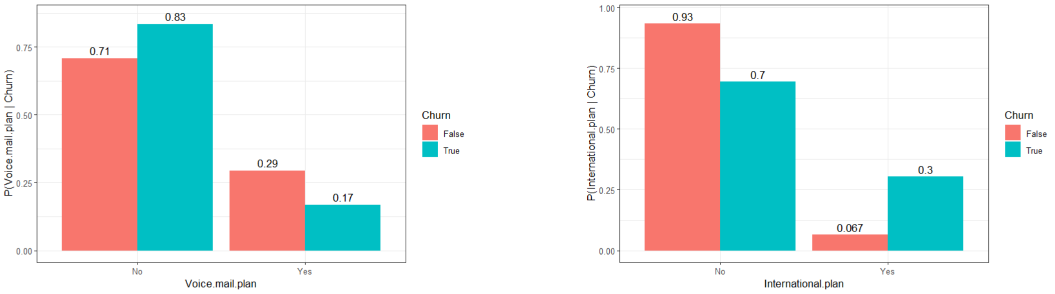
(b) Total.night.charge Variable Boxplot

Figur 1: Boxplots for Numerical Variables To Display No Outlier/Anomaly Behvaior

Given the presence of both categorical and numerical variables, it is necessary to demonstrate the distributions differently in order to visually investigate ad assess them, and their contributions in predicting the target variable. Remembering that the main task of our analysis is to analyze and predict the customer churn, it is pivotal to derive basic insights on the variables at hand. One way to investigate this is to assess the variables related to the customer's churning. First, it is worth calculating the percentage of customers within the data that churned, as opposed to the customers who did not. It was found that approximately 14.6% of customers left the Telecom company, while 85.4% stayed. These percentages indicate that the data is noticeably imbalanced, which already implies a future challenge and consideration for the prediction modelling.

To start the visualizations, the numerical variables are plotted as histograms, and the categorical variables as barcharts. Upon looking at the categorical variables, it appears that the International plan and Churn variables have notably similar graphs; they both have a very similar ranges for the Noänd Falserespectively, as well as for the "Yesänd True"variables.

This prompted the question of how correlated are they two variables? How many customers without an international plan actually stayed with the company? Are the numbers actually similar or do the variables just follow a similar distribution? In order to further investigate this, a conditional probability barchart plot was plotted. It plots the probabilities of a customer churning or not, given if or not they have an international plan.



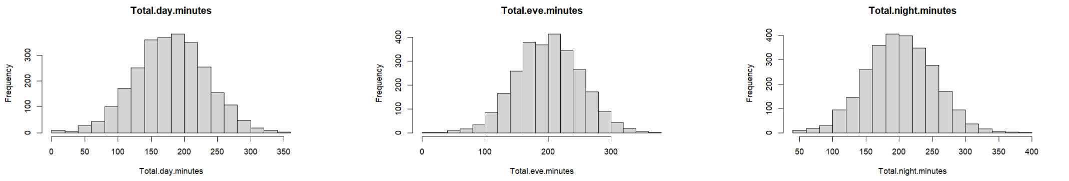
(a) Voice.mail.plan Conditional Probability Bar Chart with Churn Variable

(b) International.plan Conditional Probability Bar Chart with Churn Variable

Figur 2: Bar Charts for Categorical Variables To Display Conditional Probabilities with Target Variable Churn

After plotting this graph, it is inferred that the variables are somehow correlated–93.3% of customers who stayed at the company do not have any international plan, while the 6.7% who ended up leaving, do have one. These numbers are very helpful in predicting the customers staying; however, 70% of customers who left the company did not have an international plan, while 30% who left, did. This indicates that the distribution is somewhat similar–and might help to indicate whether or not a customer may churn or not, and can be a helpful variable in determining it, but moreso for predicting the customers staying at the company. This investigation can be applied similarly to the Voice mail plan variable, which visually looks more different from the churn, but actually follows a closer/higher conditional probability with it. 71% of the customers who do not have a phone plan stayed at the company, while 29% left. 83% of the customers who do have a phone plan stayed at the company, while 17% left. This shows also that this variable might be helpful in predicting the churn, as very few people who have a voice mail plan actually end up churning from the company.

For the numerical variables different histograms were plotted. In general, many of them followed what appear to be normal distributions, making these variables good for statistical analysis and prediction.



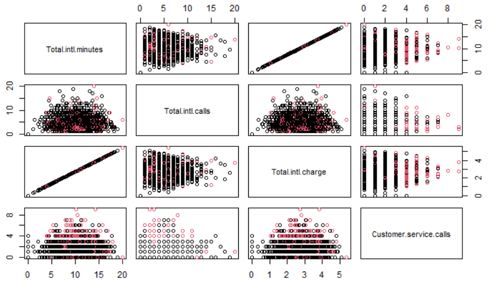
(a) Total.day.minutes Variable Histogram

(b) Total.eve.minutes Variable Histogram

(c) Total.night.minutes Variable Histogram

Figur 3: Different Numerical Variable Histograms to Display Distributions

The scatterplots of the variables were also plotted, and the variables generally do not follow any visible trend. The samples churning and not churning are embedded together and overlap with eachother. However, you can see that certain variables are essentially perfectly linearly correlated. For example, total day minutes and total day charge. Total evening minutes and total evening charge as well. Same for the total night and total international variables for the minutes and charges.



Figur 4: Different Spreads of Variables to Visualize

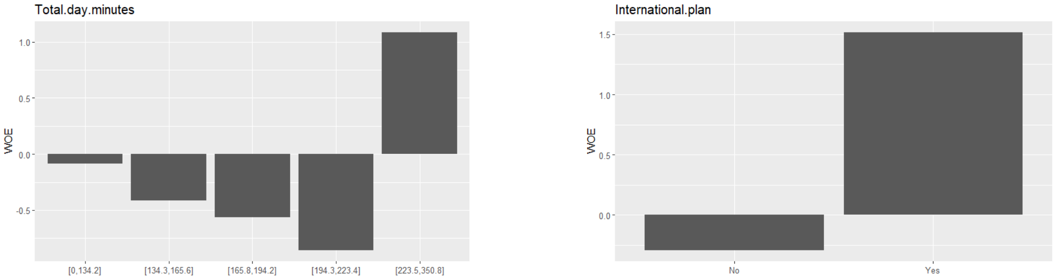
This implies that certain variables are essentially duplicated"in the data, as in having one of the variables is enough because they follow the same distribution.

Next, the information values and weight of evidences were computed, to further investigate the variables and their relationships with the target variable.

First, it is notable that the variables such as total day minutes and total day charge, again, have the same information value, as they are perfectly correlated. This makes sense understanding that the total day charge depends on the total minutes the customer has used, and their is a very direct correlation. This same trend follows for all the variables that have minutes and charges. The Total.day.minutes and Total.day.charges variables also have an information value that is the highest among the variables, 0.52, which is over 0.5, indicating that they are very strong predictors, but they fall just over into being of a level of suspicious predictive power. This could mean that the variable is in fact very predictive of churn, or there could be a data leakage here, meaning it is directly associated with churn, and thus useless to have in the predictor variables. Thus, these variables need further investigation and evaluation to understand whether their is a valid predictive influence or this could just lead to overfitting. Further investigation for this can be conducted as well.

The weight of evidences show that generally, the lower the variable, the less likely the customer is to churn. Then, after a certain interval, approximately 224 minutes per day, the trend completely flips, and the customers are much more likelier to churn. This makes sense, as the charges would increase as well, which is probably a factor causing churn.

Following this, the international plan variable has an information value of 0.43, meaning it has a strong predictive power, but does not fall in the suspicious predictive power range. Based on the weights of evidence, customers with an international plan have a strong and positive weight of evidence, which means they are much more likely to churn. At the same time, the customers without an international plan displayed very low churn rates. This indicates that the international plan is an important factor, and a significant package or offer the company has which retains customers. This could also imply satisfaction on the customers behalf with the international plan package Deepanshu Bhalla, 2015.



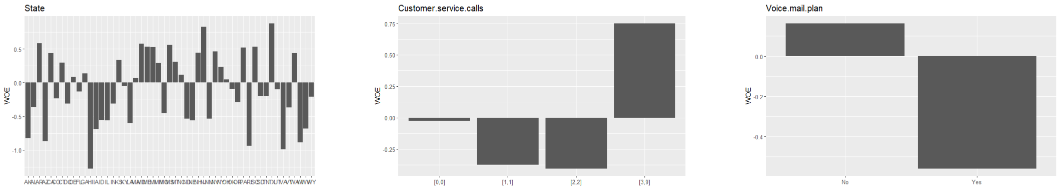
(a) Total.day.minutes Variable WOE Plots

(b) International.plan Variable WOE Plots

Figur 5: Weight of Evidence Plots For Variables With Highest Information Values

Furthermore, the state variable actually obtains an information value of 0.25, and the customer service calls 0.23, which both have a medium predictive power. The state variable has high variations for the different locations. Some have very strong negative weight of evidences, while the others have contrastingly strong positive weight of evidences. These show the different churn patterns. So this means that the churn varies significantly throughout different states, meaning the quality of service or other factors that would retain a customer could need to be improved in specific locations/branches. This is an important derivation for improving and lowering the churn rates. For the customer service calls, the customers churning with 1-3 calls is extremely low, but then drastically increases in magnitude and changes to the positive, displaying that customers that when it reaches 4 calls, the customers churning drastically increases. This provides insights into the customers generally leaving after they make more than 3 calls to customer service. Higher interactions with customer service may highlight different unresolved or frustration issues with the customer, which could further provide business insight into potentially enhancing their customer services and tackling issues discussed with customer service.

Lastly, the voice mail plan has an information value of 0.09, which falls under the weak prediction power, but on the higher side. There are more features that fall under the weak predictive power, including the total international calls, total evening charge, total evening minutes, total international charge, and number of voicemail messages. These features are the last features that fall into a relevant information value level, indicating that the rest can be rendered not useful for prediction. These variables alone have weak predictive power and thus might not be useful to analyze alone, but could be helpful along with other features and patterns to predict the churn. The voice mail plan may be useful to investigate given further or a more detailed investigation is wanted, and follows a general trend of those with a voice mail plan tend to be loyal customers staying with the company.



(a) State Variable WOE Plot (Varied over 50 States)

(b) Customer.service.calls Variable WOE Plots

(c) Voice.mail.plan Variable WOE Plots

Figur 6: Weight of Evidence Plots For Variables With Second Highest Information Values

To further investigate the variables and how they contribute to prediction in terms of importance, Principal Component Analysis will be applied. In attempt to investigate whether or not the metrics and results actually benefit from a feature reduction, the PCA analysis will be performed both on a subset of features and on the full dataset. The subset of features consists of the variables that obtain an information value greater than 0.02, meaning they have any sort of predictive power in the data. Also, total day charges will be removed as it is redundant, following the same distribution as the total day minutes variable, which could create unnecessary confusion. The results will then be evaluated.

The data was first applied as numerical variables in R, because PCA can only work with numerical data, and thus categorical variables need to be one-hot encoded Datatricks, 2025. The target variable "Churn", was also removed from the dataset, and the data was then scaled in order to ensure all variables are on the same range/numerical interval to properly investigate the variables fairly. The results show that the data can be reduced to 13 principal components to capture all the meaningful/useful variance. The first 4 components also explain approximately 50% of the variance Stack Overflow, 2022.

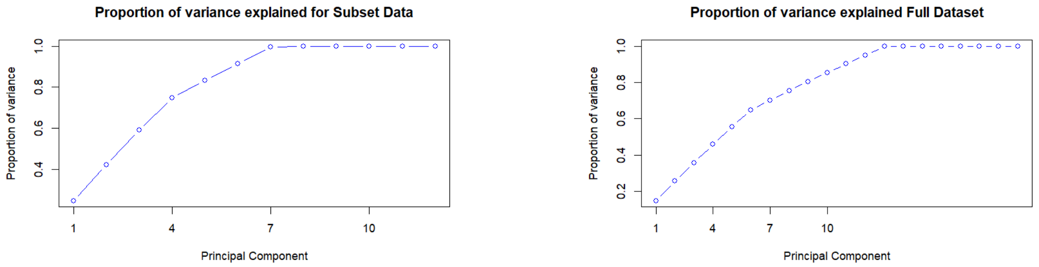
Looking at the first principal component, the voice mail plan and number of voicemail messages have the highest loadings, with voice mail plan "Yesänd number of voicemail messages having positive values and voice mail plan No"having negative values. This indicates that these values are capturing a high amount of variance in the data, which reaffirms what was displayed in the information values and weight of evidences. These variables are very significant in this dataset.

The second component focuses on the international calling behavior. The variables International plan, both Yes and No, again have the highest values in magnitude, and are opposite in sign, and the total international calls and total international minutes variables have the second highest values, indicating that the international package and usage, is in fact highly relevant, capturing a very significant part of the data. This again communicates the importance of the international plan variable and its relevance in this dataset–not even for its correlation with the churn, or its predictive power, but in capturing the variance and highly explaining the data.

The third component captures a relationship between the time of day and the minutes and charges used. The variables total day minutes and total day charge have the strongest contributions, with total night minutes and total night charges being the second greatest contributor here, as well as total international minutes and total international charges. The total day minutes and total day charge hold a lot of the weight here, and they also hold an information variable that falls within the suspiciously predictive range, as seen earlier. Them having a significant contributions further validates that these variables are in fact actually predictive and not suspicious. The total night calls and minutes variables also follow a similar pattern, while the total night charges variables have an opposite sign, being negative. This means that they follow and opposite patterns, in that they have opposite effects in the principal component.

The fourth principal component concerns the evening and international call activities. The relevant variables for this component–ones having the highest magnitude–are the total evening minutes and total evening charge, with positive values, and total international minutes and total international charges negative values. This component resultingly displays the tradeoff or balance now between the evening usages from the international usage. This component thus highlights user behavioral patterns for customers who primarily use the services during evening versus the international services.

This PCA data and the description of the components was applied to the full dataset provided. A further attempt to investigate and understand how the dimensionality can be reduced, and feature extraction can be applied, in order to retain or even achieve better results while investigating the data and maintaining useful data. The results showed that the PCA applied on the subset of features where the information value was greater than 0.02 performed better than on the full dataset. This make sense, as removing irrelevant and the redundant total day charge variable is likely to remove noise, as well as giving better component structure.



(a) Proportion of Variance Explained for Subset Dataset

(b) Proportion of Variance Explained for Full Dataset

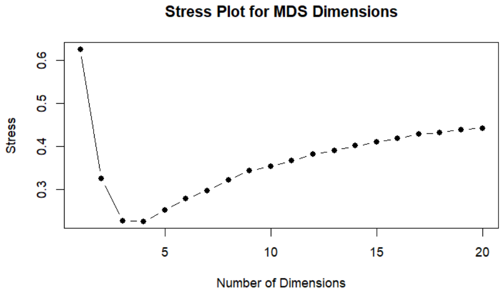
Figur 7: Proportion of Variance Explained for Both Subset and Full Datasets

The PCA on the subset features captured a greater variance, with 99% of the variance being captured in the first 7 components, and almost 60% of the variance explained in the first 3 components. The PCA on the full dataset took approximately 12 components to explain 90% of the variance. This makes sense, as after removing possible noise, redundancy, or irrelevant features, the PCA was actually better able at focusing on the relevant features and patterns. The PCA results for the subset features has similar findings, in terms of which features are relevant for churn predictions, but highlights slightly different feature relationships and different components highlight different features.

The first component has high positive values on the international plan "Yesänd NNo"variables. This follows a similar trend to the first PCA analysis where the International plan is an important variable capturing a significant part of the variance in the dataset. This aligns again with the information values and the weight of evidence plots, displaying that this variance not only has a good predictive power, but also is an important component and feature in explaining most of the variance in the data.

The second component portrays a similar pattern. The international plan variable and the total international charge and total international minutes variables show to be the most important values here. For both components, the first and second, the international plan NNo"has a negative value, while the international plan "Yes"has a positive value. Customers with the voice mail plan tend to follow a certain pattern, while the customers with an international plan are more likely to follow an opposite pattern in terms of behavior. This again reaffirms what the weight of evidences and the information values communicate, and reinforces the finding in the first principal component. This implies that the users who have an international plan are very significant to the company.

The third principal component has strong values on the total evening minutes and total evening charges, as well as on the total international calls and on the total international charge. This emphasizes again the importance of reducing the international plan, and how varied it is in the data, with many people following specific patterns. This again reinforces the importance of this variable in terms of describing or explaining the data. The total evening



Figur 8: Stress Plot for MDS Dimensions

minutes and total evening charges have an even higher value here, which indicates this variable is also important for the dataset, and explaining variance in patterns in the data.

After applying PCA and analyzing its impact, it is worth further investigating using multidimensional scaling (MDS), as MDS, when using the Gower's distance metric, is applicable for datasets containing numerical and categorical variables. Thus, unlike PCA, MDS can take the categorical variables into consideration as well. The results indicate that there are noticeable patterns of customer similarity, with groups or clusters that may be associated to different customer segments. The dissimilarity matrix uses the Gower's distance, as mentioned, and calculates how varied two observations may be, for all the variables.

The plot highlights the first two dimensions or coordinates from the MDS conducted. The axes correspond to the different directions that best retain the dissimilarities between the customers. There are 4 distinct groups, with red points corresponding to the churned customers, and the blue corresponding to the customers who stayed. On the top left side of the plot, where values of dimension 1 are lower and for dimension 2 are higher, the cluster displays a region where the churned and retained customers are mixed or essentially overlapping. This displays that in this group, the customers share similar patterns irrelevant of the churn factor. The cluster in the bottom right shows similar overlapping, though much less severe and much more distinct. The clusters in the bottom left and top right convey more distinct separations between the churned and retained customers. This shows that dimension 1 generally demonstrates patterns and distinctions related to the churn variable, meaning it obtains characteristics related to the patterns. These patterns could be related to the overall call activity or international plans, as highlighted first in the principal component analysis. The MDS application using the Gower's distance metric did successfully capture meaningful patterns and distinctions in the customer's and their data. However, the overlapping present also portrays that the churn is not exclusively indicative from the different customer patterns, and other factors unaccounted for may further be responsible for customers churning. This makes sense, as customer churn may be related to factors completely unrelated to data–such as personal preference, relocation, or other personal or irrelevant factors. Adding different contextual data of the customers could further help this case.

In the stress plot for the obtained dimensions using MDS, MDS is applied iteratively in a loop in order to calculate the dissimilarity matrix and stress coefficient for each dimension. The lower the stress, the better the fit, and the greater the conformity, which is more desirable. The higher the stress values, the less the MDS was able to reflect similarities in the data. The plot displays a noticeable drop in the first 3 dimensions, which means that most of the significant or useful disimmilarity information is captured within the first few dimensions. After the 7th dimension, the stress essentially plateaus, further emphasizing this point. The stress value is just above 0.2, which is above the ideal value, and is not great, but still acceptable. There is moderate distortion in the data, and to better preserve the data, more than 3 dimensions should be included or considered.

After considering 3 dimensions, the stress level is approximately the same, meaning including more dimensions did not make any beneficial changes. At this level, the idea that there are too many features present and being used was tested by reducing the number of dimensions to 10. This produced similar results in that the stress level was still slightly over 0.2, indicating the dimensionality was not an issue in the stress level. For the sake of investigation, simplicity, and for assessing model performance, the MDS using 2 dimensions with all of the features will be considered.

2 Modeling

After thoroughly investigating the data, various models will be constructed and assessed in order to successfully predict the customer churn. Logistic regression will be used first in order to statistically model the data and evaluate the data and feature relationships. In attempt to investigate different variable combinations and how different dimensionality reduction techniques might affect the model, different models will be run, the first one using the subset of features with information values greater than 0.02, meaning the features with any sort of predictive power, even weak. The total day charge feature will also be removed from the dataset, as this is a redundant feature. The second model will use the obtained dataset after applying PCA and using the first 4 components. The third one will use the dataset acquired after applying MDS. In this way, a comparison can be performed and the best model can be selected and assessed.

Before starting, a logistic regression model containing all the features was constructed and evaluated against one containing the subset of features. This was attempted to ensure and confirm that this subset of features was valid and more suitable for modeling this data. The result summaries display that the AIC is in fact lower in the second model containing the subset of features. The AIC in model 1 is 1779.2, and 1771.3 in model 2, indicating that model 2 is a better model fit and balances this with the model complexity. Thus, a logistic regression model was devised for the three different models each; one applied on the feature subset, one for the PCA data, and one for the MDS data.

The result summaries show that at a 0.05 significance level, the variables State, for only 4 of the State values, as well as International plan, where specifically customers have an international plan, total day minutes, total international calls, and customer services calls are statistically significant. This reaffirms and aligns with the information communicated through the information tables and weight of evidences. These five variables were all deemed relevant and significant in predicting the churn.

For the PCA logistic regression model, the first 7 components were used as the data, as these 7 components were shown in the plot of proportion of variance to explain all the important or meaningful variance. In the summary of this model, all components are statistically significant for predicting the churn in the model. The AIC value here, is 1958.4, which is worse than the previous models. The residual deviance also is higher, being 1742.4. This means that the subset model is better in terms of interpretability, and fits the training data better, as well as for predictive accuracy. Since the objective is to predict the customer churn, this model will not be considered for further analysis. The dataset using the MDS data also provides an even slightly worse performance, clearly highlighting that the logistic regression model for the subset of features performs significantly better, and is overall a more favorable model, and will be considered for further performance assessment.

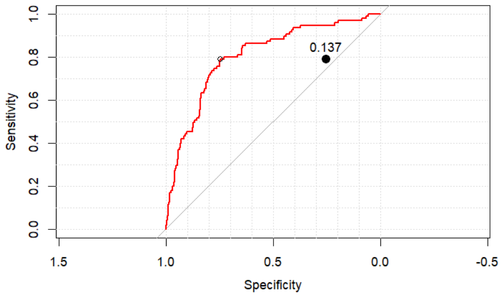
In attempt to further evaluate the model yielded for the subset of variables, the results for the model using the subset of features will be evaluated. For the default threshold of 0.5, the metrics induced were:

	Predicted False	**Predicted True**
Actual False	549	23
Actual True	72	23

Tabell 1: Confusion Matrix for Subset Data using Logistic Regression

Accuracy: 86% Sensitivity: 24% Specificity: 96%

These results indicate that the model performs notably well on predicting the customers who stay at the company, indicated through the high specificity rate mainly, as well as through the high accuracy. However, for predicting the customers who churn–which is the main focus of our analysis–the model performs significantly poor; which is seen in the 24% sensitivity rate, which is the recall value for the customers who do in fact churn. The high specificity value, and the model's ability to correctly predict customers who do not churn, as opposed to performing poorly for predicting the customers who do churn, is mainly due to the great imbalance in the dataset. The imbalance causes the classifier to learn patterns within the data for the customers who stay, much more efficiently and better. This makes the classifier struggle with predicting customers who do churn; as it has not learned as much, nor does it as readily predict the customers who churn due to the lack of presence within the data. Thus, tuning the threshold is necessary here in order to allow the classifier to predict customers who churn more readily and easily.



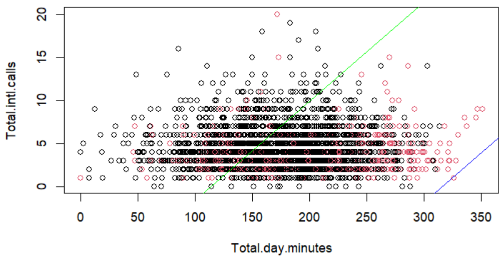
Figur 9: ROC Curve for Logistic Regression with Optimal Threshold Point

After plotting the ROC-curve, the AUC value obtained is 0.86, which is indicates that the model fits the data relatively well, and it performs fairly well at predictions. The optimal threshold was yielded as 0.137, meaning at this threshold, the sensitivity is now improved while the specificity is still at a decent level. This is intuitive, as lowering the threshold essentially tells the classifier that any sample with a predicted probability over 0.137 will be predicted as a 1, or predicted as a customer who churns, which increases the range for predictions for customers who churn, leaving a much greater interval for the customers who churn. This is done to help balance out the effect of the imbalanced data, and allow the classifier to essentially predict more customers who churn, raising the sensitivity, or the accuracy of the recall of churning customers. At a threshold of 0.18, the sensitivity becomes 0.86, which is a significant improvement from 0.24. Also, at this threshold, the specificity is 0.825 and the accuracy is 0.81, which is lower, but also still considerably good. This threshold essentially balances the tradeoff between the sensitivity and specificity, while maintaining optimal metrics–mostly for the sensitivity, as in our case, this is the most important metric for the task.

3 KNN Visualization and Modeling

To further investigate and analyze the data and attempt to explore the predictive abilities of different models, K-Nearest-Neighbors will be used to construct three different models, which will be evaluated. The three models pertain to three different datasets; one consisting of the subset of features chosen from the information values, one consisting of the relevant PCA components derived from applying PCA on the subset dataset, and the last one consisting of the dimensions obtained from the MDS.

First, to understand and confirm the predictive power of different variables in plots using coupled features, decision boundary plots were made in order to accomplish this. Visualizations using KNN can be useful because KNN as a model works differently in nature to logistic regression. KNN can help uncover different patterns locally and feature interactions and patterns that might not be found in statistical models such as logistic regression.



Figur 10: Total.day.minutes V.S. Total.intl.calls Scatterplot with Decision Boundaries at Different Threshold: 0.1 and 0.5

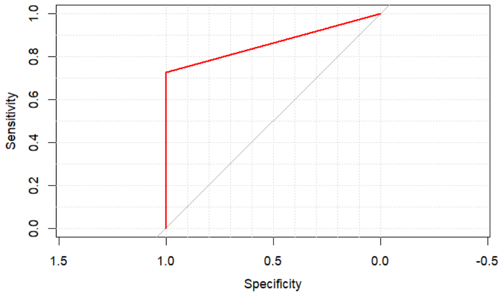
The results from the plots reassure the findings mentioned earlier, such as customers with higher daytime usages in variables such as total day minutes are much more likely to churn. Also, customers with higher customer service calls are much more so associated with churning customers.

The following plot displays two variables already deemed important for predicting the churn, Total day minutes v.s. the total international calls.

In the plot, the black points represent the customers who have not churned, while the red points represent the customers who have churned. The green line represents the decision boundary at a threshold of 0.1, while the blue line portrays the decision boundary at a threshold of 0.5. The plot highlights the difference in the thresholds, and the importance of the threshold in predicting the churn. Considering the green line, anything above this line is classified as a sample that churns. Pushing the boundary allows for the classifier to catch and predict more churn cases, allowing for more churn predictions in the case where the samples do in fact churn. Furthermore, the red dots being more dense where the day minutes variable is higher reaffirms the conclusion that customers with higher amounts of day minutes are much more likely to churn. Also, where the total international calls values are lower, the customers show to churn generally.

The total international calls and total international charge variables against eachother as well as the total international calls plotted against the total international minutes displays that users that use these services frequently and abundantly tend to not churn, showing that this service or plan is generally beneficial for the customer and tends to keep them retained. With this, the total evening minutes against the total evening charge show that there is a perfect linear relationship with these variables, implying that it is not necessary for both variables to be present, and one may be dropped. It is also noticeable that the higher the K value chosen, the less strict and more flexible the boundaries are.

The models are all trained with the training dataset, and use the test dataset to make predictions. The data is also scaled, and add the variables are applied as numeric variables. The model on the feature subset produced the following:



Figur 11: ROC Curve for KNN for the Subset Feature Dataset

The ROC curve for the KNN model follows a very steep and straight line until it reaches a sensitivity of about 0.7. After this, the graphs shifts, indicating this is an optimal/changing point for the threshold. This is somehow understandable, because it doesnt give really probabilities, but class predictions.

	Predicted False	**Predicted True**
Actual False	571	26
Actual True	1	69

Tabell 2: Confusion Matrix for Subset Data KNN

Threshold: 0.5 (Default) Accuracy: 96% Sensitivity: 73% Specificity: 99.8%

	Predicted False	**Predicted True**
Actual False	568	4
Actual True	63	32

Tabell 3: Confusion Matrix for PCA Data KNN

Accuracy: 90% Sensitivity: 34% Specificity: 99%

	Predicted False	**Predicted True**
Actual False	572	0
Actual True	94	1

Tabell 4: Confusion Matrix for MDS Data KNN

Accuracy: 86% Sensitivity: 2% Specificity: 100%

From the results of the confusion matrices along with the corresponding metrics, it can be inferred that all three models produce plots with good metrics diplaying the train and test errors, showing that the models generally are decent fits and do not over or underfit the data. Furthermore, it is deduced that the KNN model used on the subset of features performs the best of the three. It has the highest accuracy and sensitivity percentages, and overall is the best at predicting the churners accurately. The sensitivity however does need to be improved. This can be done again by obtaining the optimal threshold, as done similarly for the logistic regression model. This can be done by plotting the ROC curve and using the coords function to get the optimal threshold, in order to effectively balance the sensitivity and specificity tradeoff. The optimal threshold value obtained is approximately 0.1, with a final accuracy of 76% and a sensitivity of 80%. These metrics display higher results than within the logistic regression model, meaning the KNN model is more effective for predicting customers who churn.

	Predicted False	**Predicted True**
Actual False	432	140
Actual True	19	76

Tabell 5: Confusion Matrix for Subset Data KNN with Optimal Threshold)

Threshold: 0.1 (Optimal) Accuracy: 76% Sensitivity: 80% Specificity: 76%

For the model selection and performance evaluation exploration aspect, different logistic regression models were evaluated for analyzing the customer churn prediction. Using the the subset of features again, where the features that have an information value greater than 0.02 are used as predictors, the best subset selection processed was applied on this considering the AIC and BIC criteria and values to balance the model fit and complexity. AIC prioritizes the predictive accuracy while BIC favors model complexity, favoring simpler models. In our case, the AIC criteria is favored as the focus is customer churn, and thus predictive accuracy. This was done using the bestglm() function. The best subset also differs from the stepwise selection process in that the best subset evaluates all possible combinations; thus giving a more well-rounded outcome. The best subset, considering the lowest AIC value, gave an AIC value of 1755.5 and a deviance of 1737.5. This is the best over model for balance and fit. This is even better than the results in the previous section, where the AIC was 1771.3, and the deviance was higher than in this one, being 1737.5. The best subset model considering the BIC value has an AIC of 1756.5, and a deviance of 1740.5, which is still considerably good, but not as good as the first model. The stepwise AIC model chooses the same subset and is identical in its metrics to the first model. The stepwise model for the BIC has an AIC of 1803.6 and a deviance 1740.5, thus same deviance as the best model, but a slightly higher AIC.

The dataset has 11 predictor variables–12 in total with the Churn–after the subset using the information values was considered. The AIC-based best model however, only chose 8 variables. It considered the variables International plan, voice mail plan, total day minutes, total evening minutes, total international calls, total international charge, customer service calls, and number voicemail messages. This means that the state, total evening charge, and total international minutes variables were excluded in the model selection and from the best subset selection, and in the stepwise selection models for all models considered. This indicates that they did not provide beneficial predictive power, and this shows in that the models achieved the optimal metrics and AIC criteria values without them.

The LASSO results again reinforce this. The state, total evening charge and number voicemail messages coefficients were rendered zero. The customer service calls has the highest coefficient, 0.42, with the total day minutes following it with a coefficient of 0.012, and total international calls having a coefficient of -0.096. This indicates that with an increase of customer service calls, the churn increases. This validates and reiterates what was already understood and displayed in the weight of evidences analysis. The case is the some for the total day minutes, an increases leads to an increase in the probability of the customer churning; again already stated. The total international calls has an opposite relationship, as the coefficient given from LASSO is negative, indicating that when this variable increases, the likelihood of the customer churning decreases, which again supports the interpretations made through the weight of evidences and information values; a customer having an international plan and using more minutes means they are less likely to churn.

After applying logistic regression along with performance assessment and model selection techniques, as well as KNN models, it is interesting and beneficial to further model the data using a different model. Tree models are intriguing here because they are able to capture the data and patterns from a different perspective, which is interesting to provide a well-rounded and complete overview of this data. Tree models are great at using gradient descent and capturing deeper patterns in the data, that might not be local, such as in KNN modeling.

The results of the decision tree model on the subset feature were analyzed both for the full tree and for the regularized pruned tree.

	Predicted False	**Predicted True**
Actual False	554	18
Actual True	25	70

Tabell 6: Confusion Matrix for Full Tree Decision Tree Model)

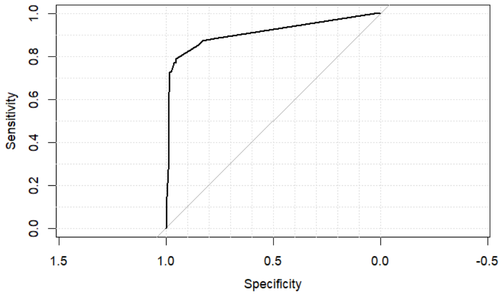
AUC Value: 0.948 Accuracy: 94% Sensitivity: 74% Specificity: 97%

	Predicted False	**Predicted True**
Actual False	556	16
Actual True	25	70

Tabell 7: Confusion Matrix for Pruned Tree Decision Tree Model

AUC Value: 0.905 Accuracy: 94% Sensitivity: 73% Specificity: 98%

The results show that the pruned and the full tree model are both very excellent models for modeling this dataset and for the corresponding metrics derived. The not pruned, basically the full tree model performs the best on this dataset in terms of accuracy and sensitivity, being the recall for the churners. This model also fits the data well, and is a good classifier for the data, as seen through the ROC curve Statology, 2021. The variable importances are first the total day minutes, for predicting the churn. This is very interesting, and emphasizes evidence seen in the weight of evidences Stack Overflow, 2019.



Figur 12: ROC Curve for Full Tree Model for Subset Data

The curve has an AUC value of 0.91, which is the best model, AUC values, and metrics obtained yet. This is also without modifying the threshold value.

This model was also investigated on the PCA dataset for the subset of features. It did not perform as well for this model, and obtained the metrics

This model was attempted for the MDS dataset, but was infeasible and unable to provide an outcome. Overall, the MDS data for this dataset, does not provide any useful insight or aid the prediction models in any noticeable or significant way.

4 Discussion

Overall, this investigation provided some beneficial and useful insights into the dataset and for predicting the customer churn. The values for the variables computed in the weight of evidence align with the PCA findings, as well as the significant variables from the logistic regression analysis. This is interesting, because the information values and weight of evidences, provide insights into the dataset as a whole and in general for prediction, while the logistic regression model is specifically for the ability to predict the churn. The variables further align with the feature importance variables computed from the decision tree model. The most important variables for the prediction being the total day minutes, state, and customer service calls further align with the weight of evidences, and the total day minutes variable is very important for predicting the churn. This means that this variable is the most telling of whether or not a customer is going to churn. This can provide some very significant business insights, especially in analyzing factors affecting customer churn. The total day minutes variable being the most indicative of the customer churning or not, could be for multiple reasons, and it is worth further investigation and testing if the company wishes to decrease the rate of churn. This could also imply that the customers are unhappy with the rates, especially in high amounts, to which the company could focus on providing or devising a specific plan for customers who use great amount of minutes during the day.

Furthermore, the State variable proving highly relevant, again, both in the weight of evidences and in the feature importance for the tree model, as well as in the p-values of the logistic regression, indicate that for a relatively good amount of states, the state itself does affect the churn behavior. This is worth investigating further as well, to analyze different factors relative to the different states of why this could be the case. This could be due to different factors, such as poorer service quality in specific states, maybe unfriendly customer service as well. This would make sense, as the customer service calls variable comes in third importance right after the state, as it was seen in the weight of evidences that customers who make more than 3 customer service calls are the greatest group of churners, and are thus, much more likely to churn. This could probably be attributed to poor, unhelpful, or unfriendly customer service, and this is worth further investigation if the company wants to decrease the churn rates.

The international plan and the total international calls having also a significant importance, conveys that the international plan is probably a key aspect for the customers not churning. Again, 83% of customers with the international plan remain at the company, highlighting an important service that the company provides.

Finally, the best model for accurately and effectively predicting the churn is the decision tree. It is able to achieve the highest accuracy and recall, specifically for the customers who churn, which perfectly caters to the task under investigation.

Referenser

Datatricks. (2025). *One Hot Encoding in R – Three Simple Methods* [Accessed April 14, 2025]. https://datatricks.co.uk/one-hot-encoding-in-r-three-simple-methods

Deepanshu Bhalla. (2015). *Weight of Evidence (WOE) and Information Value (IV) Explained* [Accessed April 14, 2025]. https://www.listendata.com/2015/03/weight-of-evidence-woe-and-information.html

Stack Overflow. (2019). *How do I plot the variable importance of my trained rpart decision tree model?* [Accessed April 14, 2025]. https://stackoverflow.com/questions/56304698/how-do-i-plot-the-variable-importance-of-my-trained-rpart-decision-tree-model

Stack Overflow. (2022). *PCA: x must be numeric in R* [Accessed April 14, 2025]. https://stackoverflow.com/questions/73093813/pca-x-must-be-numeric-in-r

Statology. (2021). *Classification and Regression Trees in R* [Accessed April 14, 2025]. https://www.statology.org/classification-and-regression-trees-in-r/