

**REPLACE WITH COVER PAGE**



# Table of Contents

Data Preparation	1
Definition and Preparation of Class Variables	1
Description of the Final Dataset	2
Modeling and Evaluation	3
Selection and Evaluation of Performance Metrics	3
Method for Dividing Data (Training and Testing Splits)	4
3 Different Classification / Regression Models	4
Analysis of the Models	4
Advantages of Each Model	5
Importance of the Attributes	5
Deployment	6
Measuring the Model's Value	6

# Data Preparation

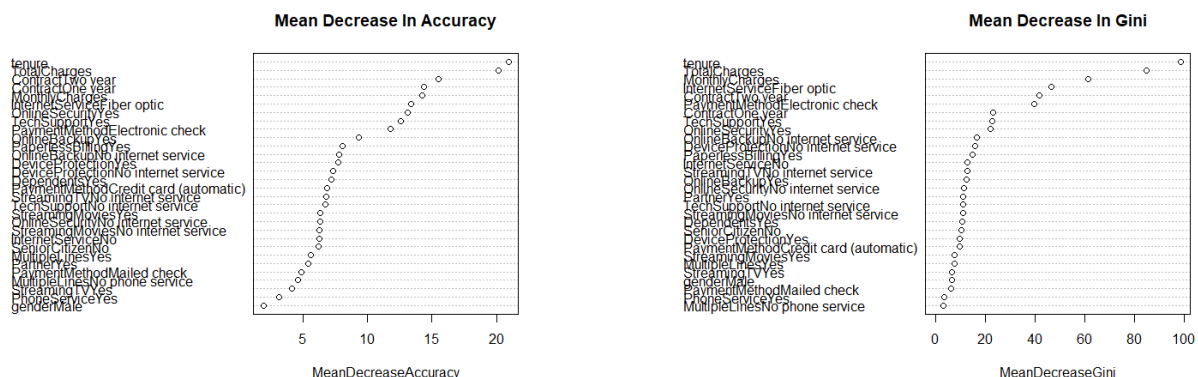
## Definition and Preparation of Class Variables

In the cell phone industry, the CRM tools used to describe customers track myriad data. One of the most significant parts of the process in interactively selecting an appropriate prediction model; is to validate the data points that most significantly impact the models ability to predict the outcome. Overt-fitting is always a concern; the goal is to get the smallest amount of data points with the largest impact on prediction.

Using a variable-importance plot can assist in determining how important a variable will be at classifying the data. Focusing on both the mean decrease in accuracy as well as node purity (Gini), tenure emerged as the most important variable to consider. In fact after TotalCharges and tenure, there is a sharp decline before the next group of predictors emerge.

The Gini coefficient measures the inequality among values of a frequency distribution. A Gini coefficient of zero expresses perfect equality, where all values are the same. Using these tools provides that the cost of service varied with the length of time with the current provider appear to contribute to the homogeneity of threnodies and branches of the resulting Random Forest model. A deduction can be made that a customer's behavior with regard to churn is most similar based on the following attributes and their node purity (note that 2 of the top 5 parameters are moderately correlated: MonthlyCharges and TotalCharges):

- tenure
- TotalCharges
- MonthlyCharges
- InternetService
- PaymentMethod
- Contract
- TechSupport



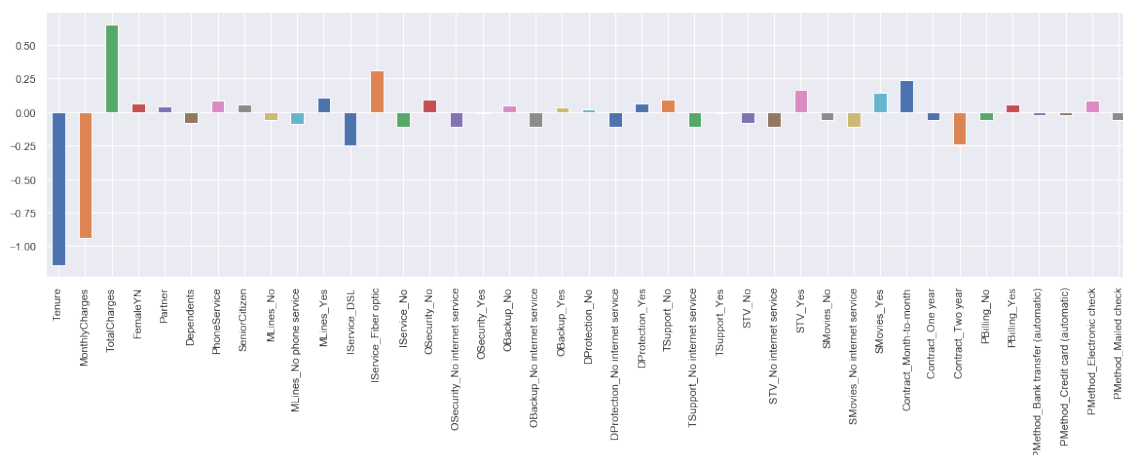
## Description of the Final Dataset

In evaluating the data model, Contract emerged to be problematic in classification; therefore it was decided to separate the Contract variable into "bins". ContractOneYear and ContractTwoYear. The various models supported that most of the churn happens to the clients in the ContractOneYear category. Pairing tenure along with the ContractOneYear was a powerful discovery in tuning the model.

The final dataset for the trees resulted in the following variables:

- tenure
- TotalCharges
- MonthlyCharges
- InternetService
- PaymentMethod
- Contract
- TechSupport

The Sub-vector Machine process had a slightly different take on the variables for prediction and will determine feature importance by coefficient weights. As a general rule, correlation is not causation. The model reflected that bundled services, security services, and data backup features showed positive weights in predicting customers who will remain with the provider more than 12 months and that automatic billing was correlated with longer-term usage.



# Modeling and Evaluation

## Selection and Evaluation of Performance Metrics

As the model continued to unfold, we were able to iterative and incrementally determine the best metrics to evaluate the success of the model. The 4 evaluation metrics chosen were accuracy, sensitivity, specificity, and area under the curve (AUC).

- **Accuracy** is the proportion of true results among the total number of cases examined. While this metric is chosen as a way to evaluate the model, it will likely be less valuable in our business case given that we could predict that a customer will not churn and likely be accurate.
- **Sensitivity** will measure the model's ability to measure the proportion of actual positivities that are correctly identified. This metric will likely be more valuable to the business case; it is more valuable to predict correctly when an asteroid will hit Earth or correctly predict that a patient will have cancer than to simply say a person has cancer and be wrong. The model success in this instance will be that it can predict when a customer will churn.
- **Specificity** will measure the proportion of actual negatives that are correctly identified which is a valuable metric to determine model performance, but not likely as valuable as Sensitivity for the business outcome of the model (which is to help the business know which customers will churn).
- **Area Under the Curve (AUC)** will measure the performance of the classification problem at various thresholds. This metric will also be significant as the model is tuned based on the business problem of classification and the way that the customers have been placed in bins depending on their length of contract. The closer the model can be tuned toward the number 1, the more likely it has a good measure of separability.

In order to prove which model performs better at prediction, the accuracy metric will take precedence. To determine which model is more accurate, the AUC metric and its corresponding ROC Curve will take precedence in the evaluation. Sensitivity and Specificity are important in determining which model is predicting the outcome variable (churn) more accurately within each category.

Metric	Logistic Regression	Sub-vector-machine	Random Forest
Accuracy	0.82	0.79	0.80
Sensitivity	0.56	0.45	0.47
Specificity	0.90	0.91	0.92
ROC-AUC	0.81	0.68	0.69

## **Method for Dividing Data (Training and Testing Splits)**

Iterating through the options, the best method for predicting the business outcome was to use a combination of techniques as an ensemble to the performance of the model and then aggregate the results across the models. The use of decision trees were employed in combination with 10-fold cross-validation.

10-fold Cross-validation was used to evaluate the predictive qualities of the ensemble by partitioning the original data into a training and testing data set. The training data was split amongst 10 equal proportions. The Random Forest was trained on 9 of the 10 and then would validate the results against the remaining 1 of the 10.

The model would iterate through all 10 possible combinations where 1 of the 10 sections would be omitted in the training iteration. This method protected the model from over-fitting. The Random Forest allowed the use of multiple trees to be constructed (all being split in a variety of ways). The model would randomly include various predictors and based on the "majority rules", the best decision tree will be picked.

## **3 Different Classification / Regression Models**

(e.g., random forest, KNN, and SVM). Two modeling techniques must be new (but the third could be SVM or logistic regression). Adjust parameters as appropriate to increase generalization performance using your chosen metric.

## **Analysis of the Models**

Use visualizations of the results to bolster the analysis. Explain any visuals and analyze why they are interesting to someone that might use this model.

## Advantages of Each Model

After analyzing the myriad results of each model and technique, deductions could be made based on the advantages to the business problem (or lack thereof).

- **Random Forest**: This model performed with less accuracy (the comparison of how many churned clients did we get correct). However, it did perform with better sensitivity (how accurately did it predict a client would churn and they actually churned). Based on the business outcome at stake (predicting clients who will churn), sensitivity would be the advantage thus the Random Forest model would be selected.
- **Logistic Regression**: Lorum ipsum dolar
- **Sub-vector Machine**: Lorum ipsum dolar

## Importance of the Attributes

Use proper methods discussed in class to evaluate the importance of different attributes. Discuss the results and hypothesize about why certain attributes are more important than others for a given classification task.

# Deployment

## Measuring the Model's Value

The model selected is useful to the business by using data to predict which clients will churn so retention plans can be deployed to prevent the churn. Focusing the efforts on using Tenure to predict the churn is powerful at preventing the leakage of data.

Marketing research tells the story that it costs 5 times more for the business to onboard and provision new clients than it does to retain the clients. If the model could accurately predict the clients who are poised to change providers, the business could save a percentage of the costs at risk if that client actually changed providers.

Given that only 1/3 of the clients leave the provider due to lower prices, the retention programs can be aimed at what clients truly desire from their providers:

- quality of service
- advancing technologies and media features
- better cell coverage
- better loyalty programs

Shifting some of the on-boarding expense could potentially raise profitability for the business by ensuring the “at-risk” clients did not change providers and better yet, they are happier customers based on the retention program and its elevation of services tuned to what clients truly desire from their providers.

The business outcome would suggest that if the marketing team could find a counter measure to short-term churn such as “first 3 months free”, they could prioritize the lack of churn for those clients. While automatic billing is related to long-term use, marketing team could target these clients with a \$10 off coupon when transitioning to auto-billing. Using this data and the model assist in narrowing down the target retention plans.

Instead of using broad-swathing programs, the marketing team could narrow their focus on the clients identified in this analysis to focus their efforts and impact the bottom line of the business earnings by saving costs of acquiring and onboarding new customers.