REPLACE WITH COVER PAGE

Data Preparation 1

[10 points] Define and prepare your class variables. 1

[5 points] Describe the final dataset 1

Modeling and Evaluation 2

[10 points] Choose and explain your evaluation metrics 2

[10 points] Choose the method you will use for dividing your data into training and testing splits 2

[20 points] Create three different classification/regression models 2

[10 points] Analyze the results using your chosen method of evaluation. 2

[10 points] Discuss the advantages of each model for each classification task, if any. 2

[10 points] Which attributes from your analysis are most important? 2

Deployment 3

[5 points] How useful is your model for interested parties (i.e., the companies or organizations that might want to use it for prediction)? 3

# Data Preparation

## [10 points] Define and prepare your class variables.

Use proper variable representations (int, float, one-hot, etc.). Use pre-processing methods (as needed) for dimensionality reduction, scaling, etc. Remove variables that are not needed/useful for the analysis.

Essentially, the goal is to predict whether a customer will “Churn” based on the selected categories of the model. This will allow the providers to be pro-active in creating promotional programs aimed at the probably of churned customers as well as target the variables that have the most positive impact on customer loyalty (tenure).

**Data Understanding**

The initial pass of data inspection indicates that the data is well formed and provides ample data size for predictive analysis. Given that most people have a device of some kind, our own team can relate to some of the indicators that might lead a customer to change providers. This knowledge is helpful, but also could be harmful if bias is introduced.

**Data Description**

The following fields are contained in the dataset along with a description of the field, values, and potential scale of the variable.

The response variable is Churn. This variable is defined as the customer terminates their contract with the telecommunications provider.

* + Churn: Boolean string showing whether or not the customer 'churned' or terminated services (Yes or No)
  + customerID: Unique alpha-numeric string to anonymously represent an individual customer
  + gender: Categorical String value to represent customer's gender (Male or Female)
  + SeniorCitizen: Boolean int value to show whether the customer is a senior citizen or not (1, 0)
  + Partner: Boolean string value showing whether the customer has a partner or not (Yes, No)
  + Dependents: Boolean string value showing whether the customer has dependents or not (Yes, No)
  + tenure: Numeric value showing number of months the customer has stayed with the company
  + PhoneService: Boolean string value showing whether the customer has a phone service or not (Yes, No)
  + MultipleLines: Categorical string value that shows if the customer has multiple lines or not (Yes, No, No phone service)
  + InternetService: Categorical string value that shows the customer’s internet service provider (DSL, Fiber optic, No)
  + OnlineSecurity: Categorical string value showing whether the customer has online security or not (Yes, No, No internet service)
  + OnlineBackup: Categorical string showing whether the customer has online backup or not (Yes, No, No internet service)
  + DeviceProtection: Categorical string showing whether the customer has device protection or not (Yes, No, No internet service)
  + TechSupport: Categorical string showing whether the customer has tech support or not (Yes, No, No internet service)
  + StreamingTV: Categorical string showing whether the customer has streaming TV or not (Yes, No, No internet service)
  + StreamingMovies: Categorical string showing whether the customer has streaming movies or not (Yes, No, No internet service)
  + Contract: Categorical string that represents the contract term (Month-to-month, One year, Two year)
  + PaperlessBilling: Boolean string showing whether the customer has paperless billing or not (Yes, No)
  + PaymentMethod: Categorical string that shows the customer’s payment method (Electronic check, Mailed check, Bank transfer (automatic), Credit card (automatic))
  + MonthlyCharges: Numeric value showing the amount charged to the customer each month

**Data Quality**

Naturally, with any large dataset, it is prudent to evaluate missing values and determine imputation strategies that allow statistical analysis to remain sound. This dataset had very few missing data points:

* + MonthlyCharges has 11 missing values. Given it is such a small percentage of the total ~7,000 values, it should not affect our analysis. At this point, we are simply excluding these 11 observations from the initial data exploration and visualization exercises.
  + SeniorCitizen is given as numeric, when for all practical purposes it is a categorical factor. (i.e. a customer cannot be 0.77 of a senior citizen; they either are under 65 or they are not). We will simply convert from an integer to a categorical factor.

To allow for visual reference for each categorical variable, Box Plots for each “Churn” category were used to assess distribution in an effort to determine whether this data can be used for predictive analysis.

The Box Plot for “Churn” and “Monthly Charges” (Figure 1) shows a significant difference in the distribution. Validating this with a [Two-Sample t-Test for Equal Means](https://en.wikipedia.org/wiki/Two-sample_hypothesis_testing), will likely show the two distributions differ significantly.

Since the data are just observations, causality cannot be inferred. However, given that the data are a random sampling; generalizations can be made. The plot indicates there is some correlation between “Monthly Charges” and “Churn”, which can be tested with [Logistic Regression](https://en.wikipedia.org/wiki/Logistic_regression) and/or [Linear Discriminate Analysis](https://en.wikipedia.org/wiki/Linear_discriminant_analysis).

Upon investigating “Tenure” against “Churn” (Figure 2), a [Welch’s t-test](https://en.wikipedia.org/wiki/Welch's_t-test) would likely show significant deviations in the two different distributions. This is fairly intuitive since customers who “Churn” usually do so within the first one to two years of their service contract.

Moving more toward the advanced visuals to round out assumptions, the data was split into 2 separate data frames; numerical and categorical. Both will contain the response variable “Churn” which allows for appropriate visualizations based on each of variables considered for analysis.

* + Two classification tasks OR
    1. TASK 1 Churn

2. TASK 2 - Tensure: EXCLUDE

- MonthlyCharges

- TotalCharges

- Churn

## [5 points] Describe the final dataset

* + Two classification tasks OR
    1. TASK 1 Churn

2. TASK 2 - Tensure: EXCLUDE

- MonthlyCharges

- TotalCharges

- Churn

# Modeling and Evaluation

## [10 points] Choose and explain your evaluation metrics

that you will use (i.e., accuracy, precision, recall, F-measure, or any metric we have discussed). Why are the measure(s) appropriate for analyzing the results of your modeling? Give a detailed explanation backing up any assertions.

**For Logistic Regression, SVM, and Random forest we used the accuracy, sensitivity, specificity, and ROC-AUC. In order to prove which model is better at prediction we use the accuracy metric. With ROC-AUC, we measure the curve to see which is more accurate. Sensitivity and Specificity help us understand which model can predict the outcome variable accurately especially with each category of the outcome variable.**

|  |  |  |  |
| --- | --- | --- | --- |
| Metrics | Logistic Regression | Sub-vector-machine | Random Forest |
| Accuracy | 82% | 79% | 80% |
| Sensitivity | 56% | 45% | 47% |
| Specificity | 90% | 91% | 92% |
| ROC-AUC | 81% | 68% | 69% |

## [10 points] Choose the method you will use for dividing your data into training and testing splits

(i.e., are you using Stratified 10-fold cross validation? Why?). Explain why your chosen method is appropriate or use more than one method as appropriate.

The 10 fold cross-validation means we take the **training** data, split it up into 10 equal portions, train the Random Forest on 9 of the 10, and validate on the remaining 1 of 10. It will iterate through all 10 possible combinations, where each of the 10 sections is omitted in a training iteration. This method allows us to get a good model built on nine parts of the data and one part to validate with the new data in order to ensure that the model is not overfit with the training data.

This modeling is based on ensemble models of decision trees called Random Forest, combined with 10-fold cross validation.

The Random Forest means multiple trees will be built, all split in a variety of different ways, and randomly including different predictors, and then based on the majority rules, the "best" tree will be picked.

The combination of these techniques is an "ensemble" or aggregate results of many models

## [20 points] Create three 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.

## [10 points] Analyze the results using your chosen method of evaluation.

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.

## [10 points] Discuss the advantages of each model for each classification task, if any.

If there are not advantages, explain why. Is any model better than another? Is the difference significant with 95% confidence? Use proper statistical comparison methods.

* + **Random forest-less accuracy(comparison of how many you got right in total set), better sensitivity(how accurately do you predict cancer when someone actually has cancer)**
  + **If you mistakenly tell someone they have cancer and they don’t is worse than vice versa that someone is told they are okay yet they have cancer and die. This is worse.**
  + A false positive, false negative is much more severe. Sensitivity matters more in this scenario.

## [10 points] Which attributes from your analysis are most important?

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.

Important feature unique to your model

Random Forest, based on variable importance and higher up in trees.

# Deployment

## [5 points] How useful is your model for interested parties (i.e., the companies or organizations that might want to use it for prediction)?

How would you measure the model's value if it was used by these parties? How would your deploy your model for interested parties? What other data should be collected? How often would the model need to be updated, etc.

We are making our model useful by not leaking in the cost, etc. We are looking at how long people should stay with you based on category, short, medium, long and adjust their marketing plans to that or bundling package to keep them longer. Or how do you put them in production.

Things to consider:

Notes:

1. Two things:
2. Down sampling
3. Down sample nos to match yeses because they were so skewed.
4. Changing search from grid to random
5. Loop so nodesize optimizes at 6. Came up with 6 through loop and found median root squared error and best one you pick.
6. At first score of AUC is 65% however now we have a real curve and AUC goes to 78%.
7. With confusion matris, we care about sensitivity metric. This is the highest sensitivity and being above 80 is great. Sensitivity her is .8365 at 84%.
8. Random forest is better at predicting churn customers of who will leave us.

Predict Churn and Tenure (how long is a customer most likely to stay with us) Which category is more accurate?