Group Project - Milestone 2

#### Project Group #5

## Data Description

For this project, we acquired a cross-sectional, individual-level data set that contains 71,047 total observations with 58 features. The target label is a binary outcome, which takes on the value “Yes” if the customer churned and “No” if the customer did not. The data set has been pre-split into train and test, with the test set containing 20,000 observations, implying a 28% test split. Of the 58 features, we have 35 that are quantitative, where 26 are continuous of type float and 9 are discrete of type int. There are 23 features that are categorical of type object. The primary key is an integer and the schema is displayed in Table 7 of the Appendix. The first two pages of this document will detail data issues, data imbalance and future steps. The Appendix is viewer optional, containing figures.

## Data Issues

### Data Missingness

In terms of missing data, there were 14 features and 1295 rows with at least one missing value. The matrix of missingness is displayed in Figure 1 and 2. We also computed the correlation in missingness between features and found many to be perfectly correlated with one another. This means if one column is missing at index i the other column will also be missing at index i. Features that exhibit this behavior include but are not limited to the following pairs: (MonthlyMinutes, MonthlyRevenue), (RoamingCalls, MonthlyRevenue), and (OverageMinutes, MonthlyRevenue). All pairwise correlations are shown in Figure 3 and the distribution of missingness is shown in Figure 4.

For quantitative variables, we use a non-deterministic imputation method called Random Regression (Gelman & Hill, n.d.). First the method fits an OLS model on non-missing data and predicts values for missing observations. The method then adds a Gaussian noise term to the predictions. The noise term has mean 0 and a standard deviation equal to the standard deviation of the model residuals (i.e RMSE). For Random Regression to be viable, all OLS assumptions must hold, most importantly the errors must be normally distributed. The benefits of using non-deterministic imputation is that it yields several sets of data. This allows us to perform sensitivity analysis on our results and allows the imputed data to take on values beyond what is observed in the data. Contrast this with bootstrap sampling, which only generates samples seen in the data.

Categorical data with missing values were left as missing in this Milestone because only one categorical feature (“ServiceArea”) had missing values. In the future we may opt to fill in the missing values with the mode or build a dummy variable for the missing values of “ServiceArea”. It is worth noting that building a classification model to predict missing labels is not tractable for this feature because it has 747 possible labels.

### Data Imbalance

The binary target variable in the train set is imbalanced by a ratio of 0.40 (number of No’s / number of Yes’s). The distribution of imbalance is shown in Figure 5. As a first-pass approach, we use random oversampling to balance the two classes. Random oversampling is the process of drawing random observations from the minority class and adding them in the data set. This effectively duplicates observations at random. After oversampling we arrived at a 1:1 balance as shown in Figure 6.

There are other less crude strategies to ameliorate imbalance such as synthetic minority over-sampling (SMOTE) and class weights. SMOTE has the added benefit of generating samples that are not duplicates of the minority class. However SMOTE is only usable in non-complex feature spaces. If the feature space is too sparse or high dimensional, the method may not accurately represent the data generating process of the minority label (Kim, 2023). Class weights is a simple approach where minority misclassifications are given a higher penalty in the loss function of a model. This forces the model to classify the minority samples better, preventing the model from predicting a constant vector of the majority class. Class weights are usually set using the formula n / (c \* n\_c), where n is the total number of samples, c is the number of classes and n\_c is the number of samples in a given class c (Kamaldeep, 2024).

### Data Scaling

Our scaling strategy of choice is the standardization. We chose this over normalization (min-max scaling) because we foresee the possibility of using techniques such as Lasso or PCA which both require scaling prior to computation. We contemplated mixing scaling methods where features with large ranges and outliers will be min-max scaled while normal features with moderate ranges will be standardized. However this approach is risky because it compromises interpretability and makes PCA or Lasso infeasible because the features are effectively on different scales.

## Summary and Future Steps

In Milestone 2, our group was able to establish a data schema. We identified missing values and applied a Random Regression imputation strategy for quantitative variables and left categorical variables alone, because only one categorical feature had missing values. We chose standardization as our scaling strategy because we foresee using regulation methods in the future. We added a scaling object to a pipeline and intend to build on it for future milestones, which will help simplify complex tasks such as feature engineering or cross validation.

In future steps we will impute categorical variables if necessary, develop new features, construct a baseline model and test it on a preliminary set of model performance metrics.

## Sources

Gelman, A., & Hill, J. (n.d.). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

Kamaldeep. (2024, October 20). *How to Improve Class Imbalance using Class Weights in Machine Learning?* Analytics Vidhya. https://www.analyticsvidhya.com/blog/2020/10/improve-class-imbalance-class-weights/

Kim, M. (2023, December 16). SMOTE: Practical Consideration & Limitations - Minju Kim - Medium. *Medium*. https://medium.com/@minjukim023/smote-practical-consideration-limitations-f0d926b661a8

## 

## 

## 

## 

## 

## 

## Appendix (Optional)

### Figure 1

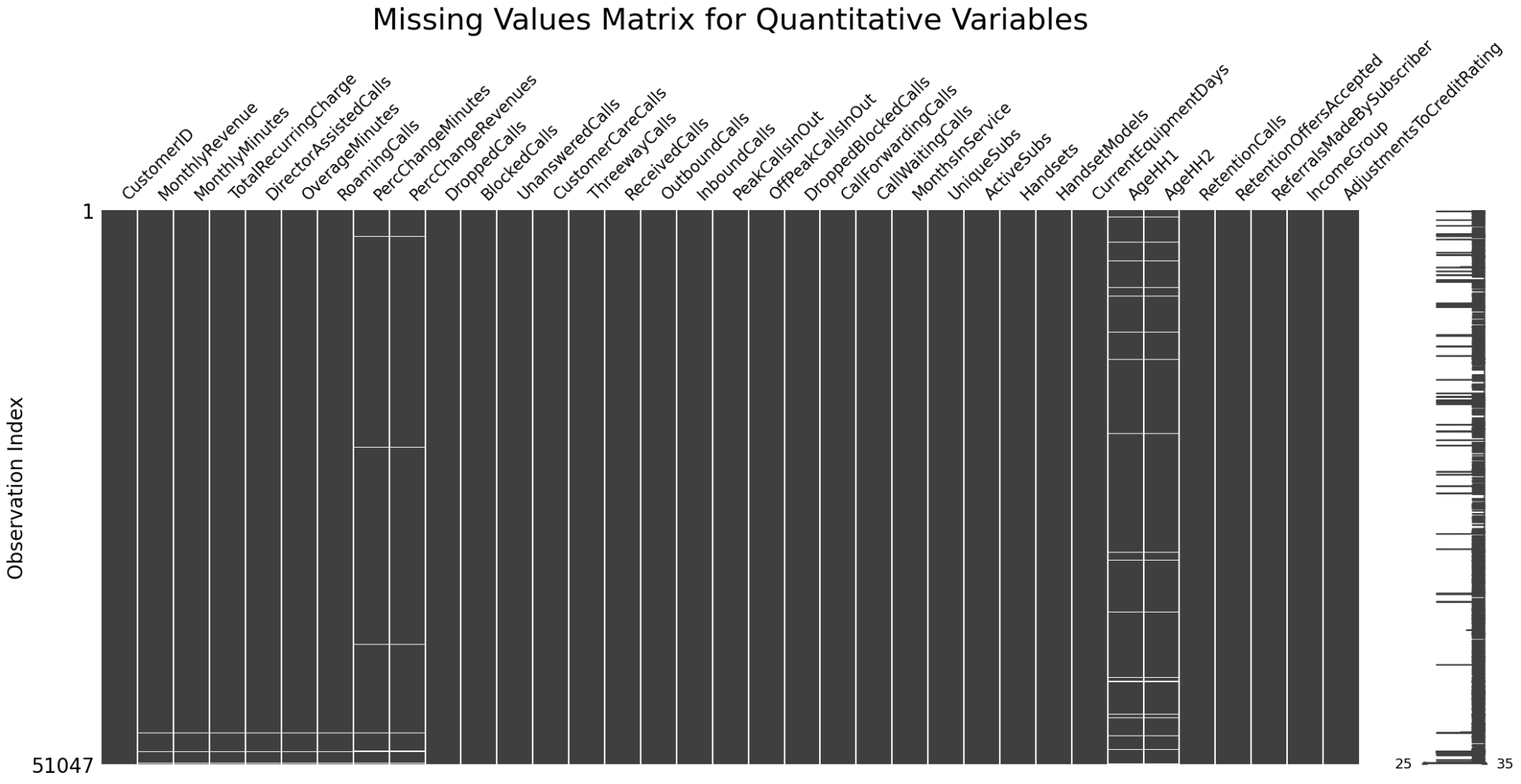
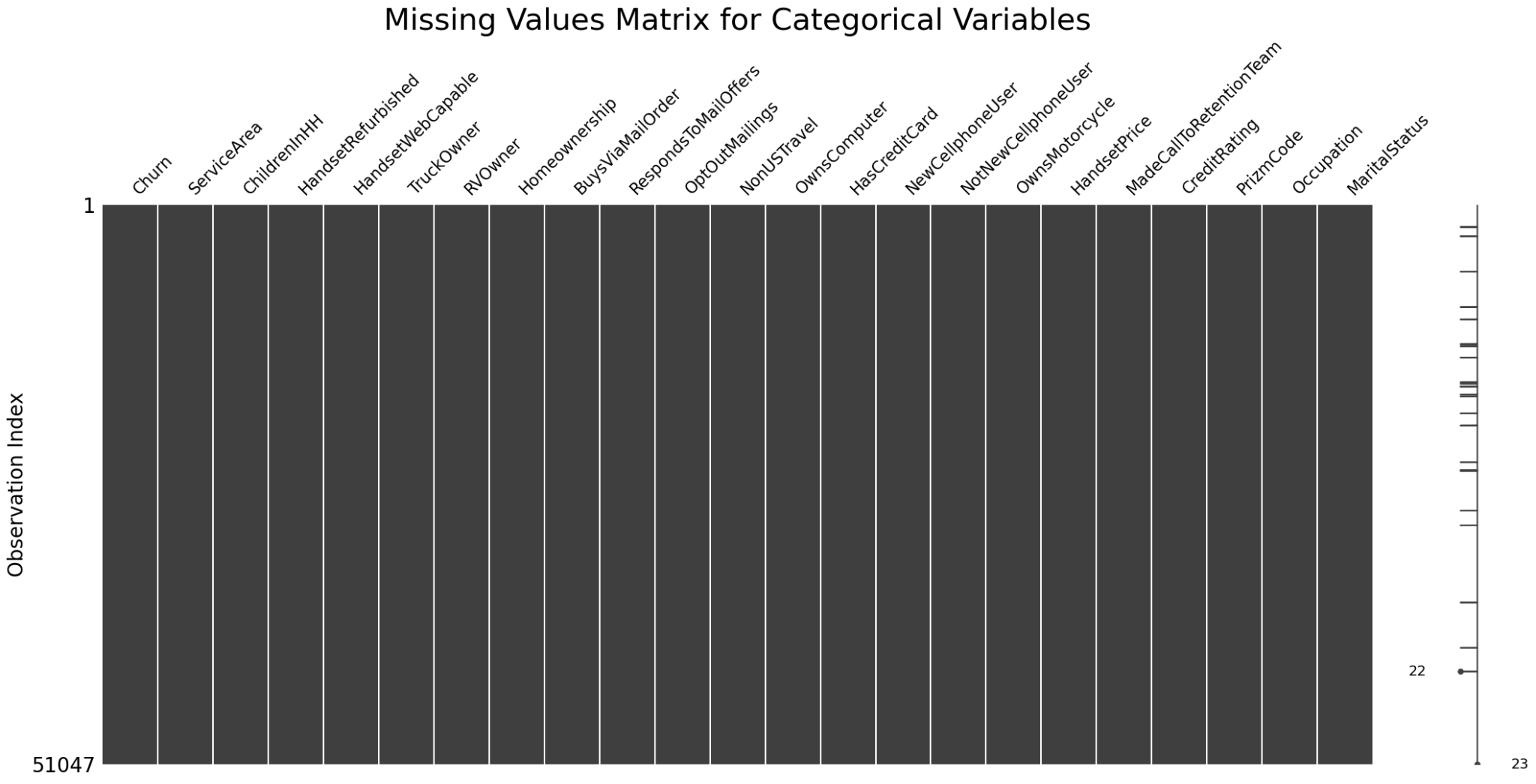


Figure 2



### Figure 3

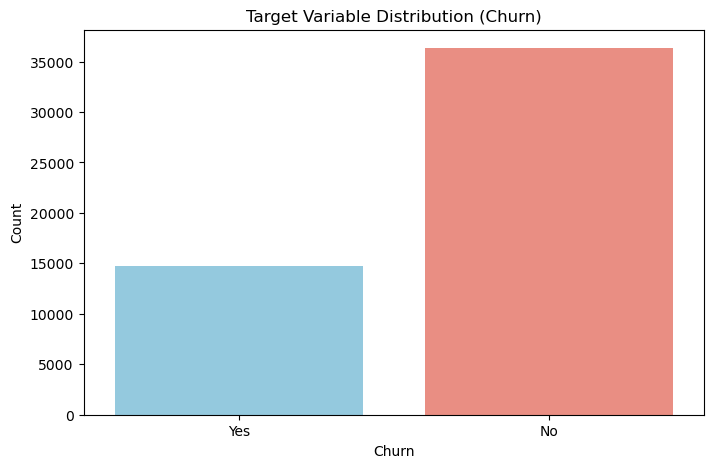
#### 

### Figure 4

### 

### 

### Figure 5



### Figure 6

### 

### Table 7 - Data Schema

****

### Figure 8a and 8b

## 

