ISQA 8720 – Course Project Milestone 2

TEAM MEMBERS

DILANGA ABEYRATHNA
ASHWATHY ASHOKAN
ANOOP MISHRA
FINAL TEXT FOR REVIEW

Abstract

This document serves as the final report of a machine learning project that predicts if a customer account of United Communication will churn or not. If the prediction is that the account will churn, an estimated churn date is also predicted. We merge firmographic data with the company provided dataset to enhance the data quality. Following the CRISP-DM model, data understanding, and data preparation have been done using conventional approaches.

The overall approach consists of two models, a classification model to predict churn while a regression model to predict churn date. In the classification model, use and test multiple models including ensemble models, i.e. models that integrate multiple predictive qualities from different machine learning algorithms to make the prediction. Similarly, the regression model predicts the duration of customers being active until they churn. Models are evaluated based on standard evaluation metrics and the best model or combination of models is recommended. It was seen that the ensemble model with three base models(XGB tree, Random Forest and SVM radial) performed the best in classifying if a customer will churn or not. LASSO regression was seen perform best when it came to predicting the churn date of the customer.

1. Course Project Description and Business Understanding

This project is about building machine learning models from a dataset provided by Unified Communications. Unified Communication provides communication solutions to its customers to bridge the gap between people, teams, clients, suppliers, and partners' situations across the globe. The primary goal of this project is to build a classification model to predict the churn probability of a customer account. Once an account is predicted to churn, the approximate churn time (in terms of the number of years after which the churn would occur) is estimated using a regression model. The main dataset used for modeling gives information about the customers of United Communication and

certain attributes that model their interactions. As a supplementary source, the firmographic dataset is also used to obtain additional details about the customer accounts such as their geographic information, nature of the company, years in existence, revenue, etc. For an accurate prediction of the churn and churn time, it is imperative to model the patterns of customer engagement by leveraging the feature space. Feature space is highly useful to approximate the relationship between historic data with customer churn. A set of variables are also derived from the firmographic dataset to be able to make predictions about the factors affecting customer churn.

One of Intrado's business units, Unified Communication, is about bridging gaps between people, teams, clients, suppliers, and partners. The goal is to make communication easier connecting co-workers across offices or partner firms across the globe. From time to time, due to competitors, change in client requirements, or technical issues customers' churn. The goal of this Machine Learning project is to identify customers who have a high potential to churn. The problem statement here is to classify customer accounts into *Churn* or *No-Churn* based on how likely they are the churn based on historical churn data. A machine learning model can be employed to learn historical customer churn patterns and predict future occurrences. The customer churn can be determined by multiple predictors variables within the primary and supplementary dataset. The ease of implementation and the ability to automatically learn from historical data patterns set the machine learning approach superior to traditional data modeling approaches. Exploration data analysis will be conducted first to study the dataset and identify striking patterns and correlation between the variables. In the light of this information, feature space to train the classification model will be chosen, and various classification models trained to predict if a customer will churn or not. Models will be evaluated, and the best model will be chosen. Next, the churned customers' data would be extracted to build a regression model to estimate a churn time for each of the customers.

The final outcome of this project would be predictions of a customer account's likelihood to churn and an estimated churn year. Along with these, a list of predictor variables significantly affecting the churn decisions is also provided. How the various variables in the dataset affect or do not affect the churn decision, and the relationships between them are also discussed. This should provide valuable insights to United Communications and can take measures to contain churn rates to protect their customer base as it is more economical to invest in retaining existing customers than in gaining new customers.

2. Data Understanding

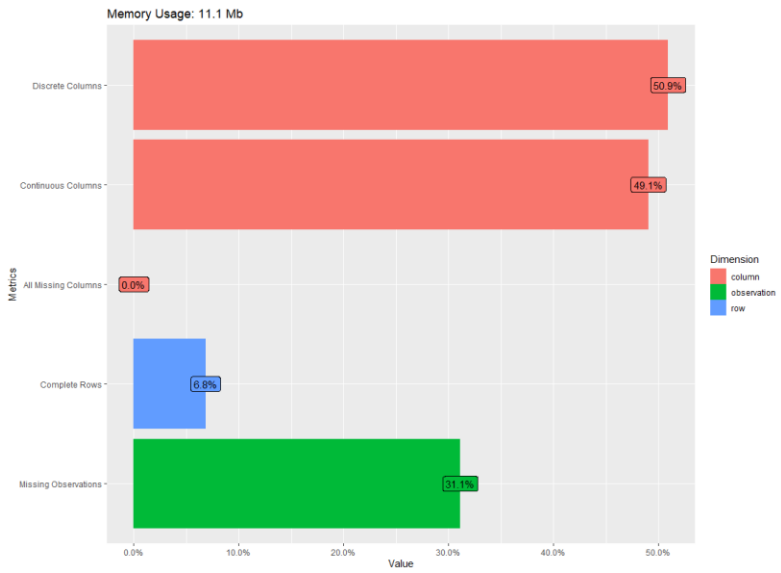


Figure above shows the various types of data from the train set

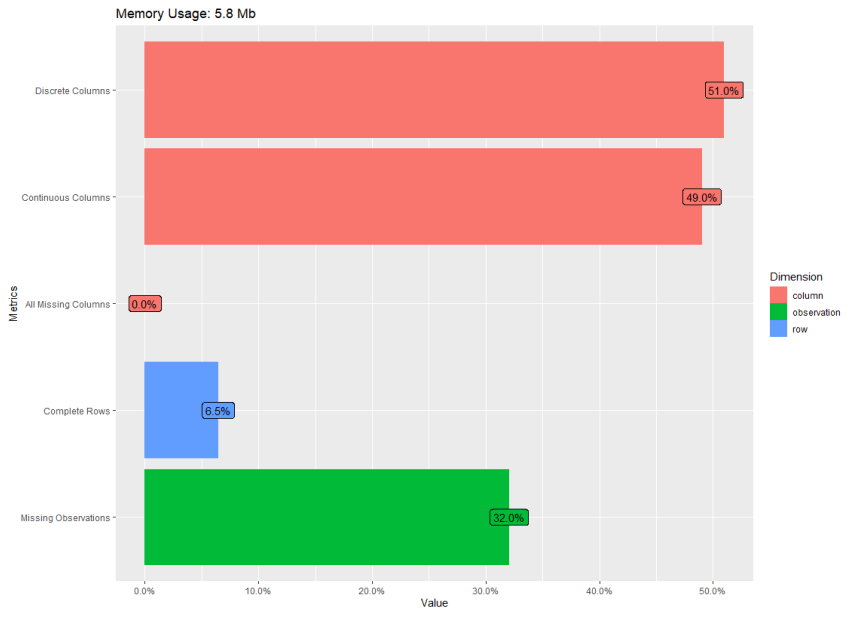


Figure above shows the various types of data from the test set

The primary dataset is provided as training and testing data. They contain information about the customer accounts, i.e., the company that uses service from United Communications. Specific information about the company, its products, transactions, revenue, number of employees etc. are provided. The training set contains information on if the customer has churned and their date of churn. A secondary source of data for each of those companies that form the customer base of United Communications has also been provided. This provides additional information about those companies such as its geographic information, type of company(public/private), years in existence, number of locations, total employee count, type of CEO etc which can provide valuable insights into the factors that affect churn of a United Communication's customer. No external datasets were used other than the provided three datasets (model, test and Firmographic dataset)

Both of the dataset, primary (modeldata_aug2020 and testdata_aug2020) and the supplementary dataset (Firmographic Data_Aug2020) is provided in a structured table format. The analysis will be conducted using primary (modeldata_aug2020 and testdata_aug2020) and the supplementary dataset (Firmographic Data_Aug2020). modeldata_aug2020 and testdata_aug2020 is the primary data source from United Communications and represents the train and test set for the machine learning model.

The train set has a total of 13 variables that quantify the following information:

	Type of information	**Description of the type**
1	Customer churn information	Customer who left and the date of churn
2	Customer account information	Unique customer identifier, date of customership, number of employees within the customer company, number of accounts for the customer company, its location etc.

3	Services availed information	Information are services availed by the customer such as, number of services availed, service usage information, and billing information etc

The test set has a total of 11 variables (as it does not have the labels that is a part of the train set data) that quantify the following information:

	Type of information	**Description of the type**
1	Customer account information	Uniques customer identifier, date of customership, number of employees within the customer company, number of accounts for the customer company, its location etc.
2	Services availed information	Information are services availed by the customer such as, number of services availed, service usage information, and billing information etc

The firmographic data has a total of 41 variables that quantify the following information:

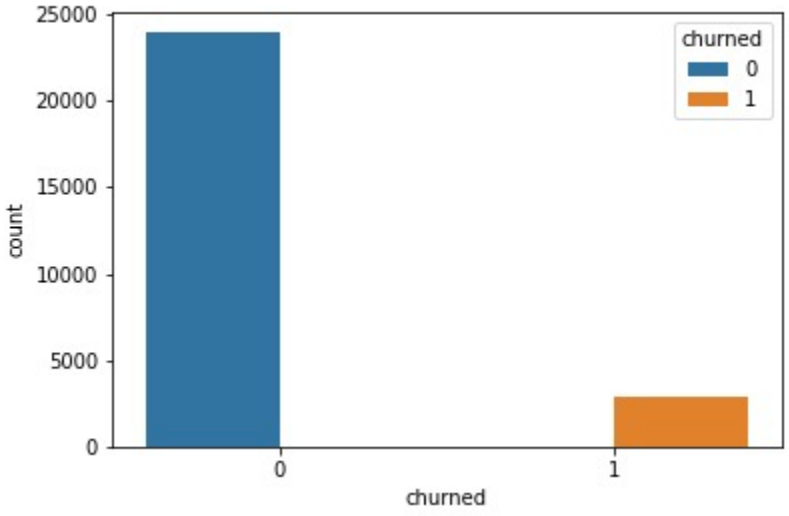
	Type of information	**Description of the type**
1	Basic Company information	Unique company identifier, years in existence, address,location and population information, owner information, employee statistics, revenue information etc.

2	Company characteristics	Information such as industry category, ownership(public/private), import export indicator, business type indication(small/large business), legal status, manufacturing indicator etc
3	Company demographics	Information that can be used to check for bias in the decision model such as location, business type(small/large), minority ownership indicator, CEO_gender indicator and title etc

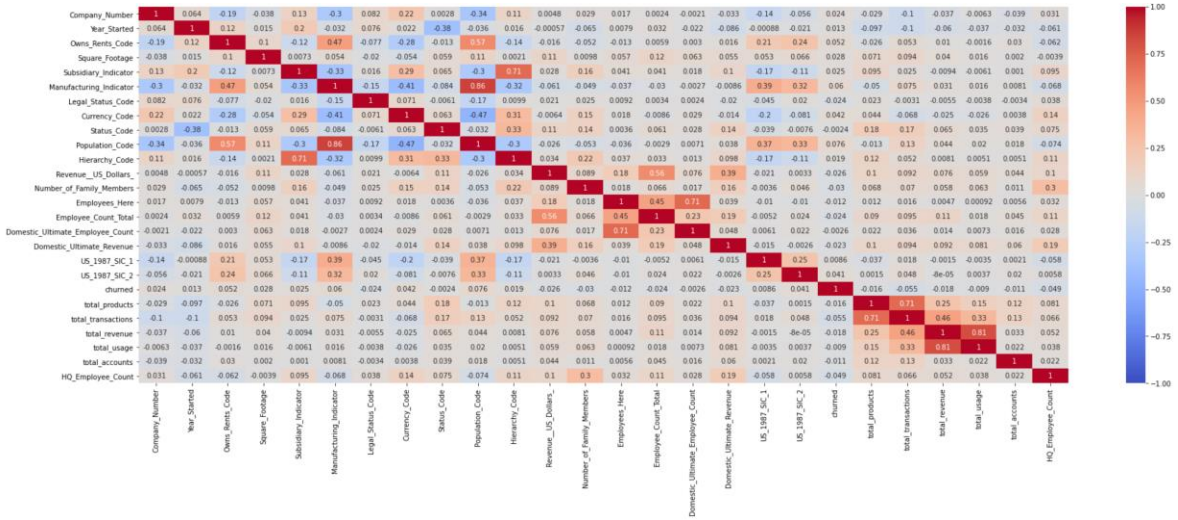
The dataset contains both numeric and categorical variables. Numerical variables consist of different value ranges. The following observations can be made from the results of the Descriptive Statistics on the training dataset.

- Label distribution and other categorical variables in the feature set show data imbalance.
- Churn vs non-churn data shows almost 90% to 10% imbalance.

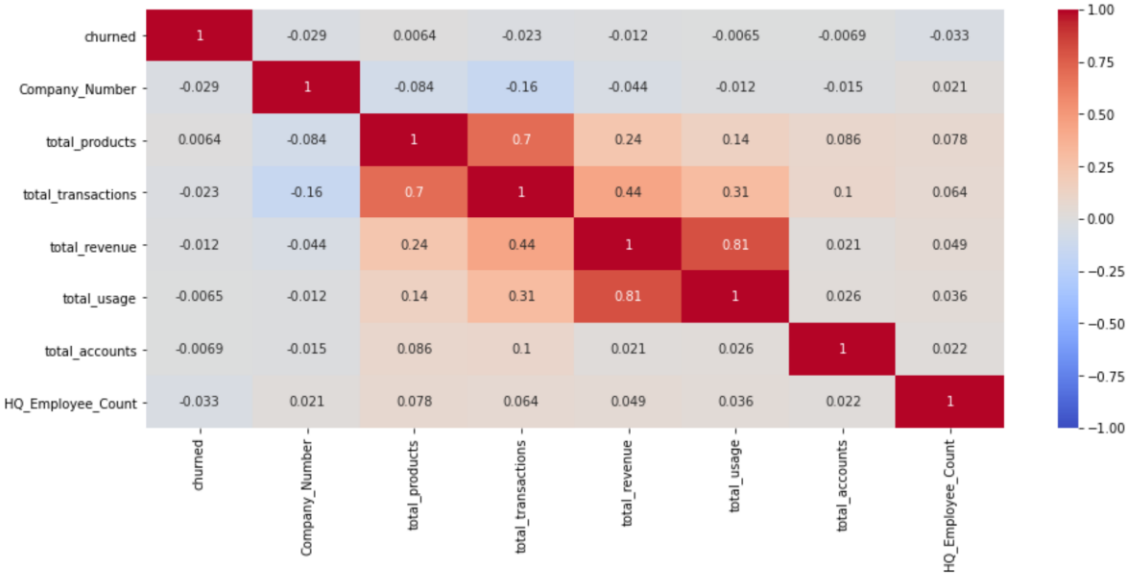
```
        NO       YES
0.8933537 0.1066463
```



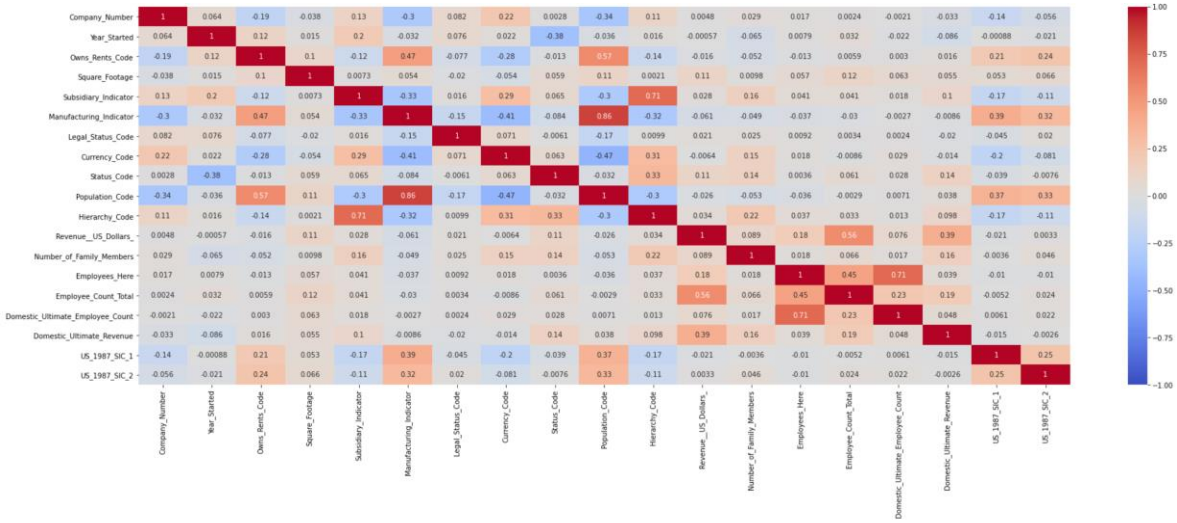
- Initial datasets with their variable correlations are illustrated as follows.



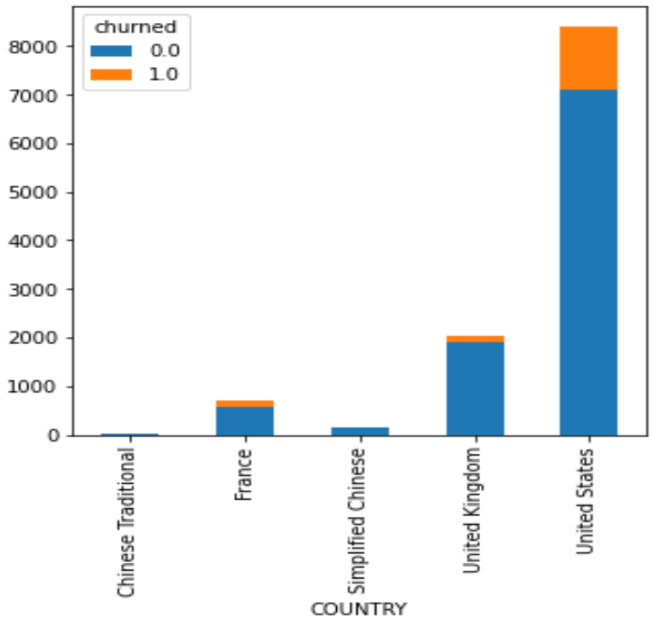
Variable correlations of Firmographic data : As shown in the correlation plot, we can see that some of the Firmographic variables have strong correlations.



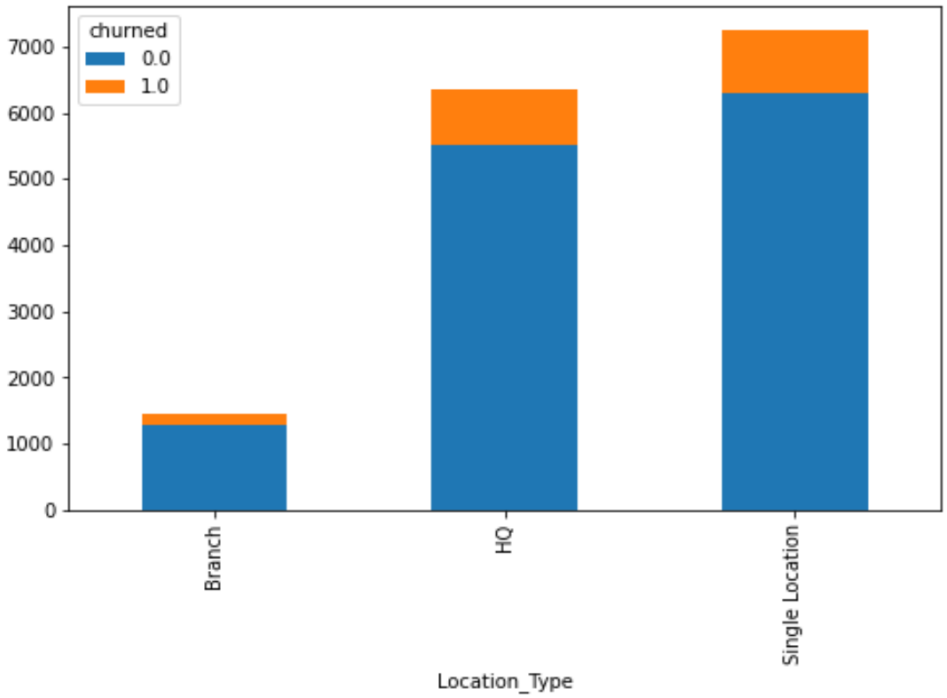
Variable correlations of initial model data : As shown in the correlation plot, we can see that some of the initial model variables have strong correlations.



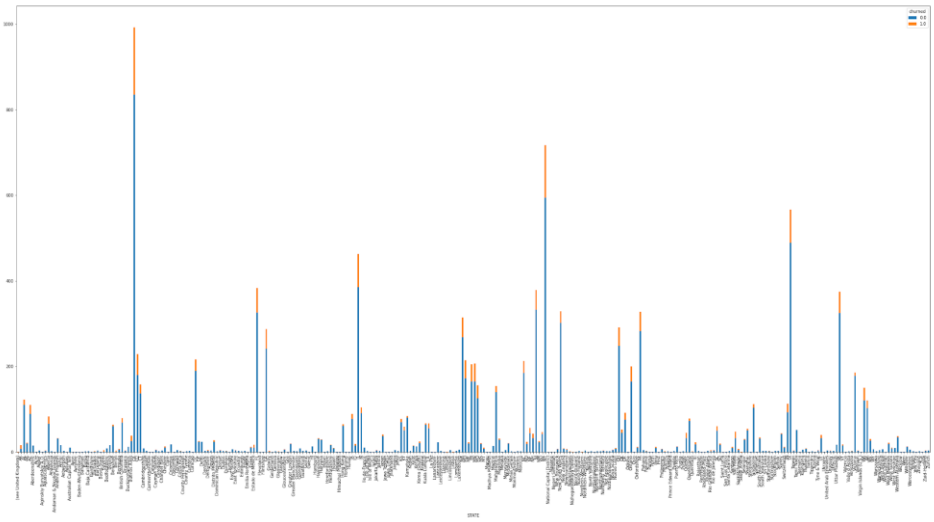
Variable correlations of initial training dataset after left join performed : Here we have
considered the merged (left merged) training data with a new set of combined variables. It is clear that the number of feature variables in the merged dataset shows strong correlations to each other, which showed potential multicollinearity issues.



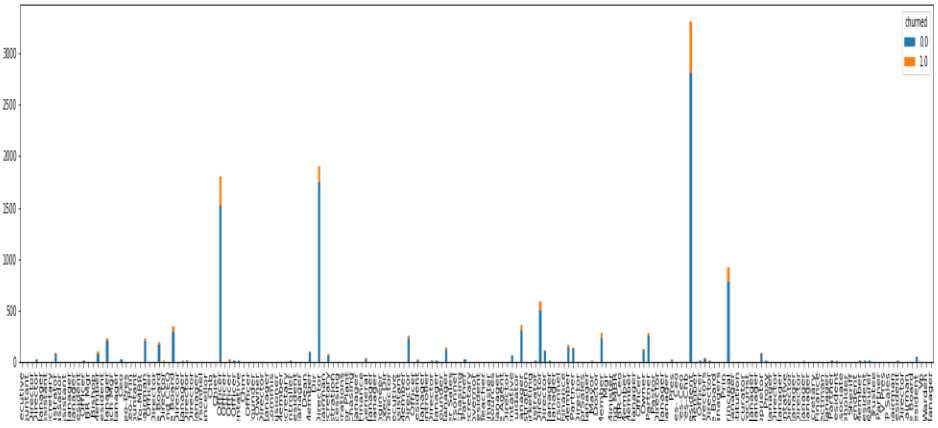
Churned label distribution based on Country



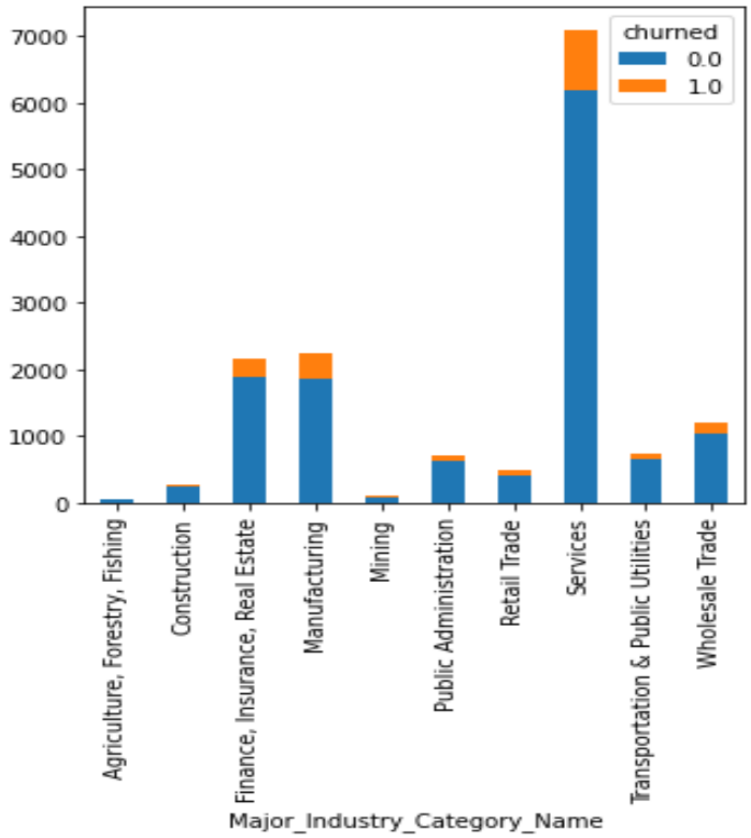
Churned label distribution based on Location_type



Churned label distribution based on STATE



Churned label distribution based on Chief_Executive_officer_Title



Churned label distribution based on Major industry category

Since churned dataset include only 10% of the dataset, we also visualized the churned distribution based on categorical variables as a data understanding activity. SOme key observations based in this are as follows:

1. Churned and not churned distribution is dominated by the US considering country as the variable.
2. Based on Location type Single location and HQ are the dominating levels.
3. Since the US is dominating the country, it was expected that US states will lead the churn distribution. The visualization above shows the same and California is having the highest churn distribution considering states.
4. Interesting observation in Chief executive officer title, where President level is the most dominant level for churn distribution
5. Services industry are having the highest churn distribution in major industry category

However based on these observations for churned distribution, it is concluded that most of the variables are having levels that are strongly dominating the churn distribution. This conclusion is considered in the data preparation activity for feature selection.

total_products with *total_transactions* and *total_revenue* and *total_usage* show high correlation. This is an indicator of multicollinearity between the variables, a key consideration during the feature selection process. Furthermore, from the summary statistics it was clear that there were many missing values across the various columns in the dataset. Summary statistics and visualization on missing values are presented in the Appendix section and data preprocessing section.

Variables with high level correlations were manually removed to avoid multicollinearity. We used a linear regression model with alias(from the VIM R package) to confirm elimination of multicollinearity. The following figure shows the variable alias which represents the collinearity values between feature variables.

	(Intercept)	Business_CodeCANADA	Business_CodeEMEA	Business_CodeOther	Business_CodeUSA	Locatior
BEMFAB__Marketability_unknown9	0	0	0	0	0	
Public_Private_Indicatorunknown10	0	0	0	0	0	
Owns_Rents_Codeownsrentmiss	0	0	0	0	0	
Subsidiary_Indicator2	0	0	0	0	0	
Manufacturing_Indicator2	0	0	0	0	0	
Legal_Status_Code200	0	0	0	0	0	
Status_Code1	0	0	0	0	0	
Status_Code2	1	0	0	0	0	
Status_Code3	0	0	0	0	0	
Population_Code10	0	0	0	0	0	
Site_Statusunknown18	0	0	0	0	0	
Revenue_Rangeunknown19	0	0	0	0	0	
Major_Industry_Category_Nameunknown17	0	0	0	0	0	

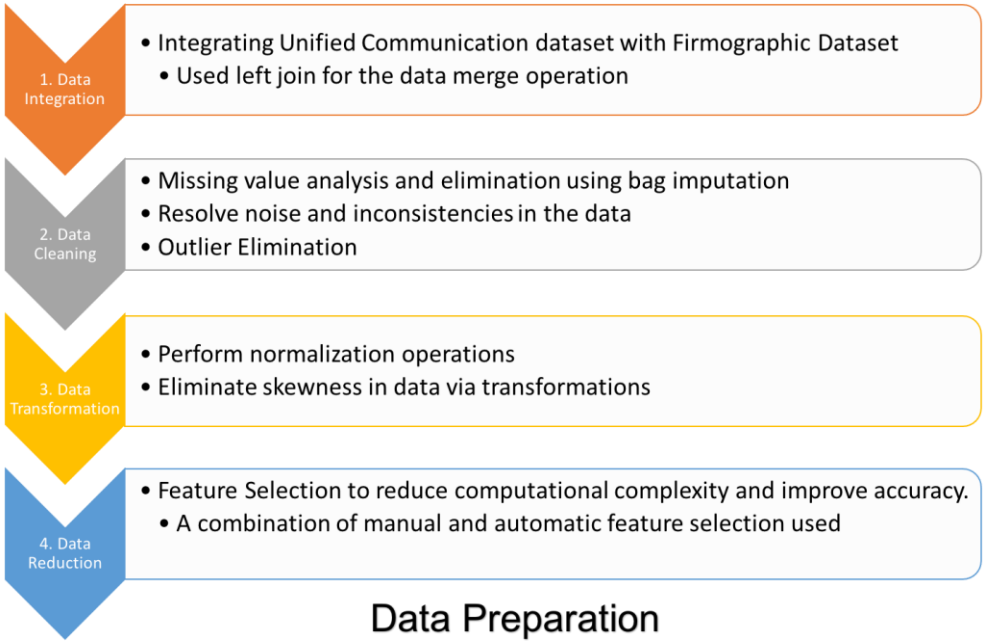
Based on the primary and secondary data provided, a number of new columns have been added to the train dataset such as age of a company, company categorizations based on various revenue levels, total number of products, geographic location and population, employee count etc.

String type columns representing date/time are converted to years/month/days and time(in seconds, minutes or hours) numeric columns. Also, we have derived new data variables such as company age (in years) and converted the churned date into the churned year. This enhanced the data representation for smooth calculation in model development. To compute the class variable for the regression model, we introduced a

new class named churned period, by subtracting churned year from company creation year.

3. Data Preparation

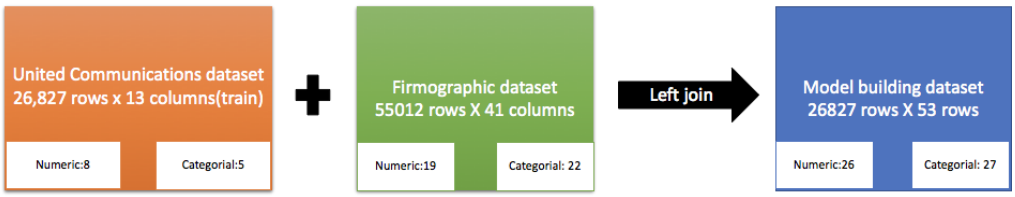
Data preparation is the process of treating and transforming the raw data to get it ready to be fed into machine learning models. It involves cleaning up the data for errors, inconsistencies, handling missing values and outliers, normalizing the scale of observations, transforming data to reduce or eliminate skewness etc. Final step to get the data ready for the model is reduce the dimensionality of the data via feature selection. The data preparation process we performed for this project is illustrated diagrammatically below, followed by a detailed discussion of the details of each step.



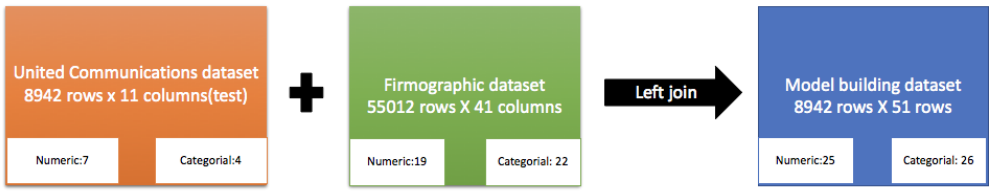
Data Preparation

1. <u>Data Integration</u>

First, we take the metadata description given from the dataset (Data dictionary.csv) to understand what each column represents and their nature of value distribution. First we consider merging train set(modeldata_aug2020) and test set(testdata_aug2020) with the provided secondary dataset (Firmographic Data_Aug2020) to leverage the additional company information in the dataset.



Data Integration – train data



Data Integration – test data

When merging the datasets we consider both inner merge as well as left merge. The use of left merge resulted in 10,000 data instances, though with a lot of missing values across various columns. Such data is known to perform poorly with the machine learning models. Hence, the inner join of Firmographic Data_Aug2020 with training and testing set is the preferred way of merging the datasets. However, later we decide to perform left join with bag imputation to enhance the data quantity while keeping the missing values imputed carefully. We eliminated the variables which had missing values more than 45%.

From here on, the train set refers to the merged training set and the test set refers to the merged test set.

2. Data Cleaning

Data cleaning involved 3 main steps -

a. Missing value analysis and elimination

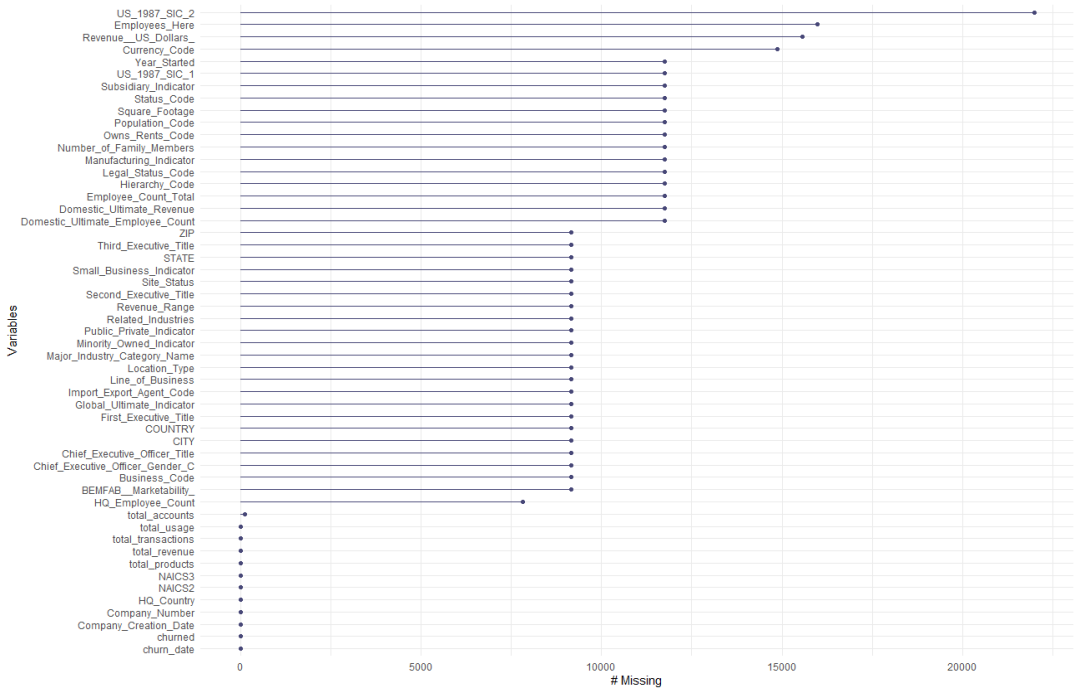


Figure above shows missing value count in the train set

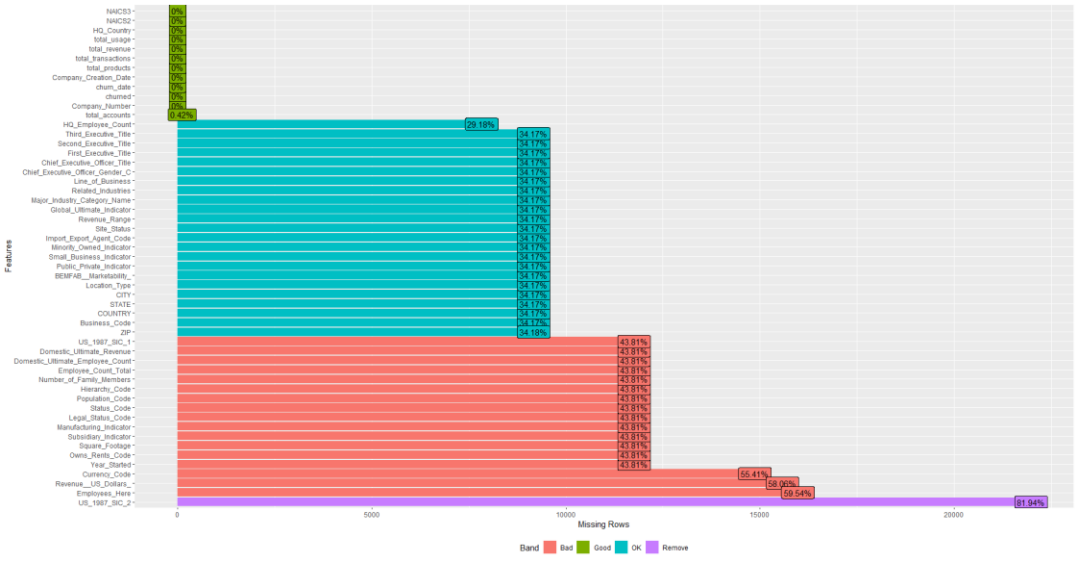


Figure above shows the % of missing values per variable in the train set

After the data merge, the final dataset contains a considerable number of missing data. Out of them, US_1987_SIC_2 and Employee_Here variables show a high percentage of missing data. These two variables whereas hence discarded from the dataset while the remaining missing values were dealt with using the various imputation techniques. We used the bag imputation strategy to deal with missing values. One of the reasons for using bag imputation is we have only a data dictionary as detail and in these scenarios bag imputation is an optimal technique to deal with missing values. This strategy replaces the missing values with derived/predicted values. Specifically, each variable containing missing values is predicted using a bagged tree of all other variables. It is superior to other commonly used imputation strategies such as KNN imputation and median imputation though is computationally more expensive. Also KNN performance is not good for large datasets(Raiwal *et al* 2012). To perform the imputation all of the categorical variables were converted to factors and a new category of NA was added to account for missing values. Numerical variables were then bag imputation and combined with categorical variables to eliminate all the missing values in the train set.

	na_count
US_1987_SIC_2	12815
Employees_Here	6806
Currency_Code	5698
Year_Started	2586
Owns_Rents_Code	2586
Square_Footage	2586
Subsidiary_Indicator	2586
Manufacturing_Indicator	2586
Legal_Status_Code	2586
Status_Code	2586
Population_Code	2586
Hierarchy_Code	2586
Number_of_Family_Members	2586
Employee_Count_Total	2586
Domestic_Ultimate_Employee_Count	2586
Domestic_Ultimate_Revenue	2586
US_1987_SIC_1	2586
total_accounts	83
ZIP	2

Figure showing the NA count per variables

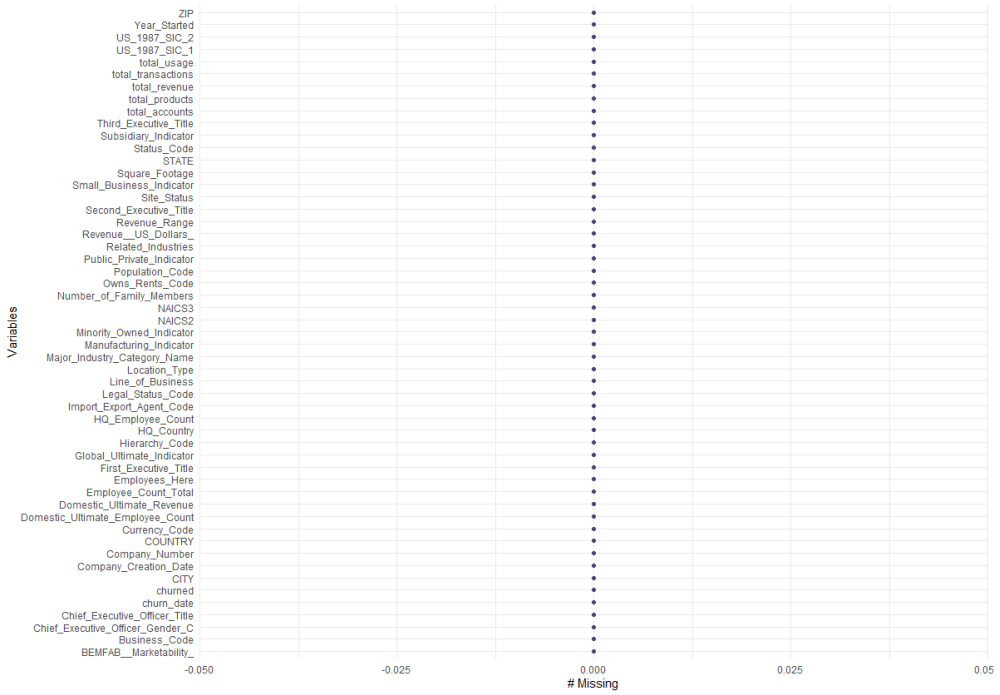


Figure above shows missing value count in the train set after bag imputation.

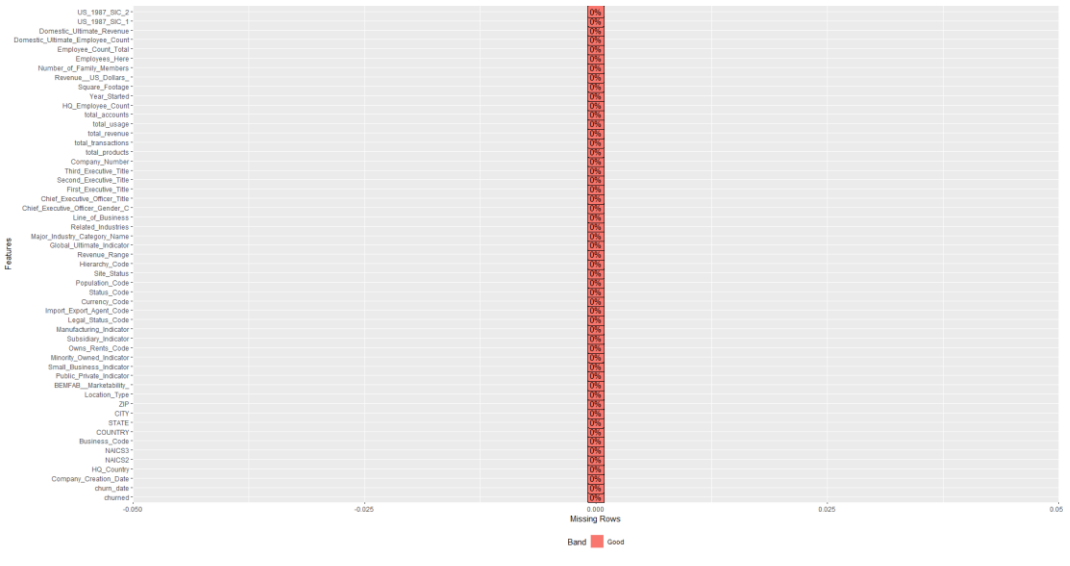


Figure above shows the % of missing values per variable in the train set after bag imputation.

The test set also received the same treatment as the train set and the statistics of the same is shown in the figures below.

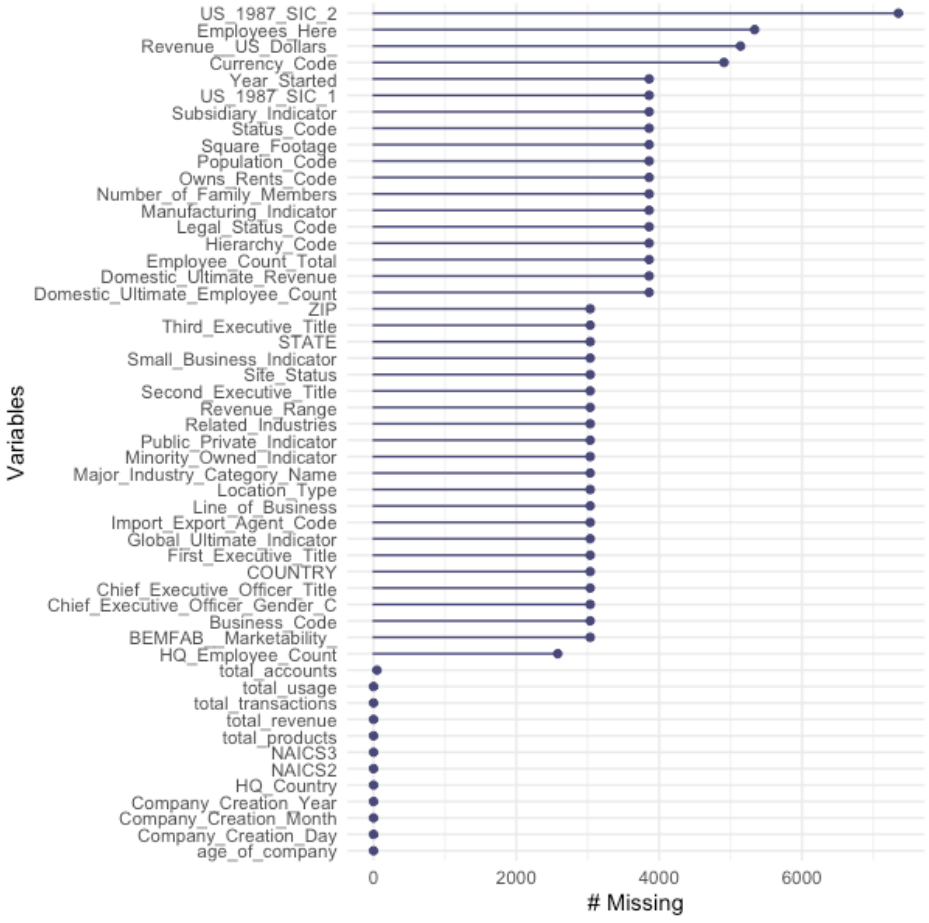


Figure above shows missing value count in the test set.

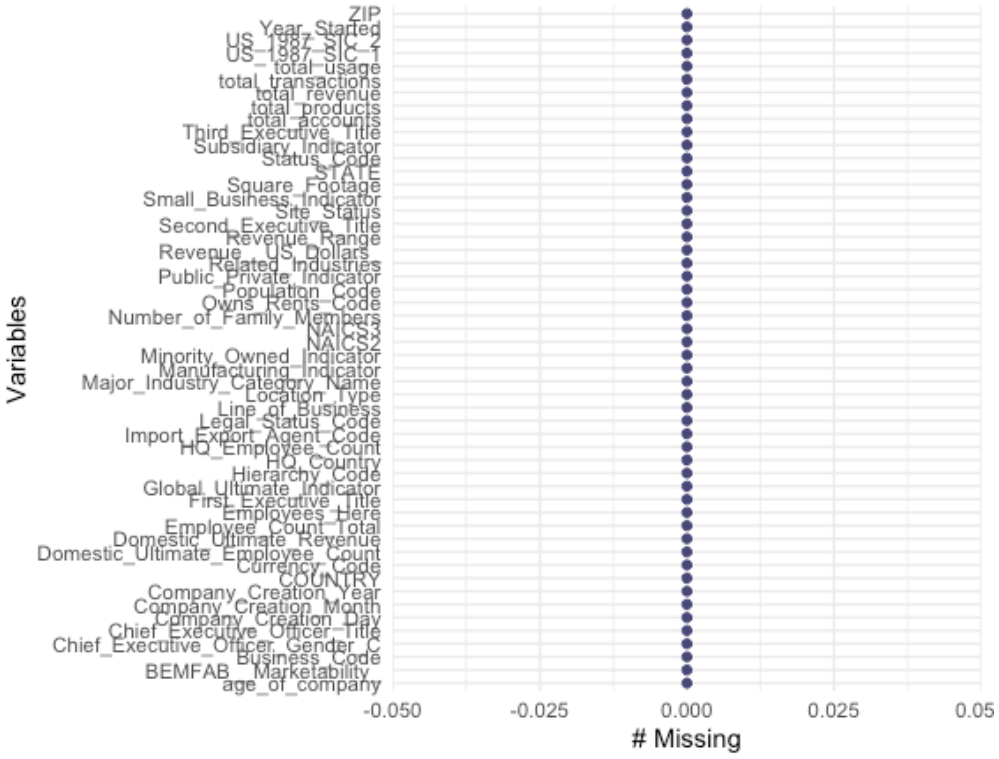


Figure above shows missing value count in the test set after bag imputation.

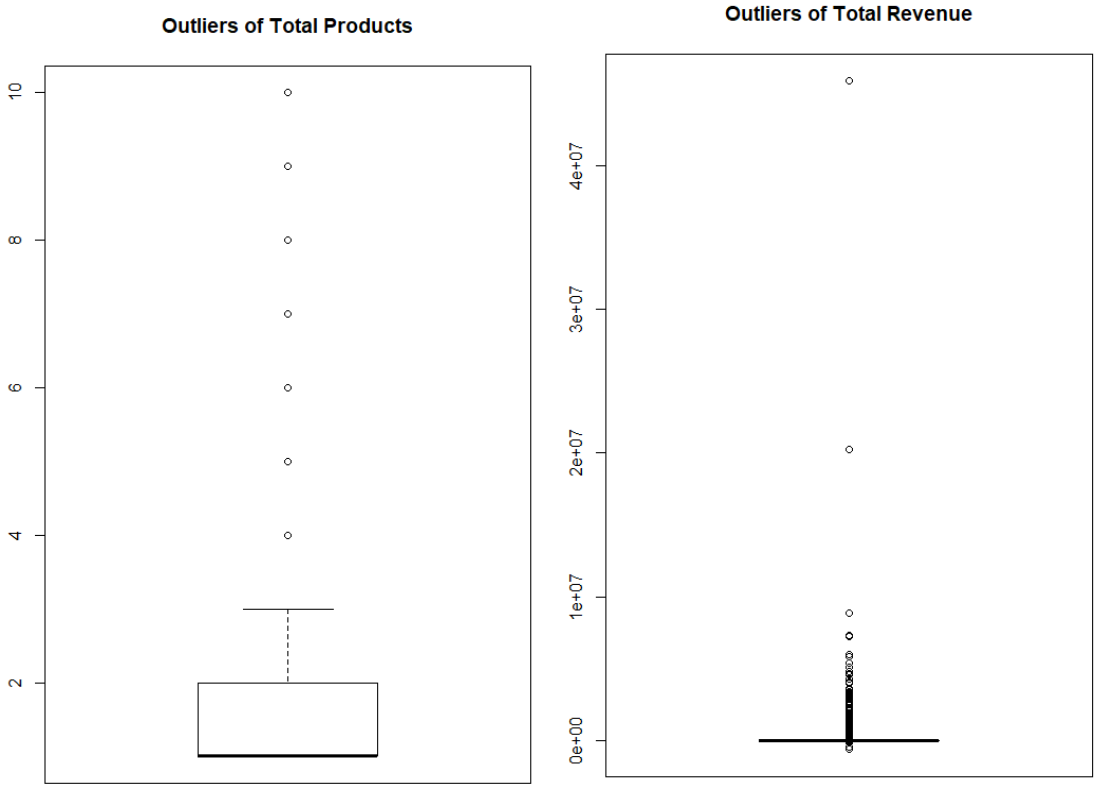
b. Resolving data noise and inconsistency

Most of the categorical data are in string format, hence they have been converted into factors format. For example, the column such as *Churn*, *HQ_Country*, *Country*, *State*, *City*, *ZIP*, *Location_Type*, *Year started, Public_Private_indiator, Small_Business_Indicator, Minority_ownedIINdicator, Owns_Rent_Code, Subsidiary_Indicator* have been changed to factors. The major data standardization and normalizations were performed during the model recipe stage.

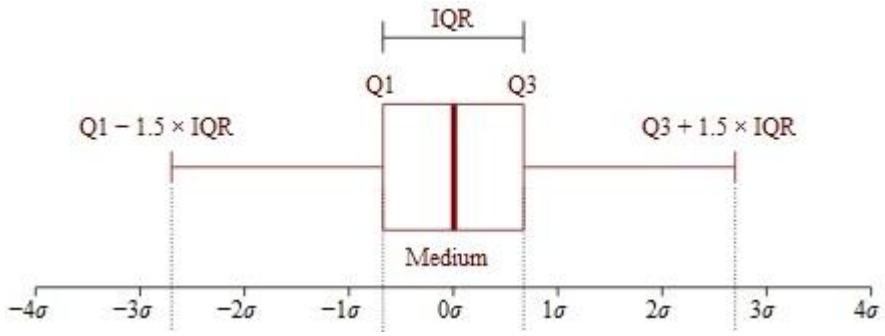
c. Outlier Elimination

Outlier is a term used to refer to the extreme values of a feature, i.e. a handful of values that fall outside the general range of value for that feature. Outliers in data can prevent the machine from effectively learning the key patterns in the data thus distorting predictions and decreasing the accuracy of the model. Hence detecting and handling outliers appropriately is an important step before the model building process especially when building regression models.

For the outlier detection, we used boxplots. It is apparent that most of the variables have a high amount of outliers. For example, following two box plots illustrate the *total_transactions*, and *total_revenue* variable distribution. After testing outliers for all the variables, we realized that some of the variables contain significant amounts of outliers, which require extra attention as removing the outliers reduces the size of the training dataset. Currently outlier removal techniques have been applied though we plan to investigate more on how to tackle outliers in an efficient manner. Winsorize method [5] is a highly used method to fix the outlier values instead of value removal in the literature. We plan to explore such techniques from literature for more efficient handling of outliers.



Common techniques used for treating outlier are: (1) imputation when the outlier values are imputed/replaces with the mean/median or mode and (2) Capping, where the values that lie outside the 1.5x*IQR(InterQuartile Range)* limit are capped by replacing with the 5th percentile, when the observation is outside the lower limit and by replacing with the 95th percentile, when the observation is above the upper limit. We used capping for treating outliers in the dataset.



3. Data Transformation

In the data preparation and understanding process we identified the variables that required transformations using conventional techniques. The transformation was performed in the light of both the classification and regression models' requirements. Some of the transformations used in this section are, variable elimination (unused, duplicate variables), data type transformations (string data to numerical data), Categorical encoding (one-hot-encoding, label encoding), scale and normalization of data and skewness adjustment (log, square root transformations).

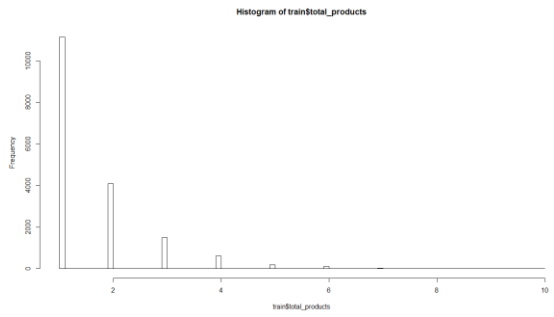
We have mainly derived Year, Month, Date separately from the given *Company_start_date* date-time string. We used string split and converted into numerical contents and factors to get more representativeness of the datetime variable. The main reason for transforming the DateTime value into multiple variables is to achieve more information from the variable than treating it as just a string. Similarly, we have derived new variables such as *company_age* and company categorizations. These company categorizations are done over revenue, employee density and administrative characteristics. These derived variables show more representative qualities than the existing variables.

We have considered using One-Hot encoding and Label encoding to convert the categorical data as numerical representation (dummy variables). However we did not use categorical variables with higher number of categories to convert into One-hot-

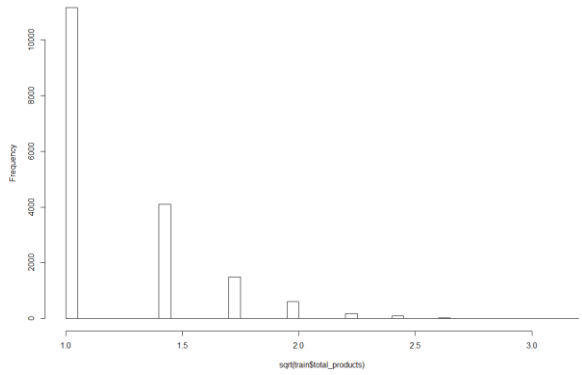
encoding as it increases the number of columns in the dataset. We used a label encoding approach appropriately. Variables such as *Revenue_Range* shows better representation with dummy variables. However, when we tested collinearity, we found out some of the dummy variables showed high correlation between each other (aka, dummy variable trap). We carefully removed some dummy variables to reduce the collinearity issues.

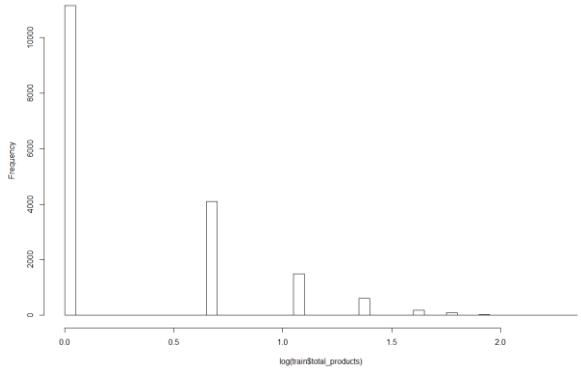
Finally, data normalizations were done on the numerical variables, as many variable value ranges are highly diverse. For example, total_revenue variable ranges from 0 to 1000000 whereas total_products variable ranges from 0 to 1000. Hence, data normalization and data scaling techniques were employed to reduce data range diversification issues.

Tests for normality shows that a lot of the variables have a skewed distribution. Plotting the histogram showed that most of the variables were skewed to the right. The conventional remedy to correct right skewed data is to consider square root, cube root, and log.



Plotting the histogram for *total_products* column showed a right skewed distribution(as seen from the first histogram on the left) with a skewness value of 2.227529.





After performing some of these transformations on the total_products column, we observed the following -

- When the *total_product* variable was transformed using the square root function, the histogram changed as shown in the second figure on the left and exhibited a reduced skewness value of 1.527186.

- When transforming *total_products* using a log function skewness value was seen to further reduce to 1.102961.

Similar tests and transformation treatment were performed on all of the variables with skewed distributions, and thus minimising the skewness value of the dataset.

Also, all the transformations performed on the train set are performed on the test set to ensure similar treatment of both.

4. Data Reduction

The final step in getting the data in shape for model building is feature selection. Feature selection is the process of selecting those features or columns from the dataset that contribute the most to the prediction of the response variable. The presence of irrelevant features increases the computational complexity of the model and decreases model accuracy [1]. Feature selection can be performed manually or

automatically. While manual feature selection enables created hand crafted features that are problem specifics and incorporated domain knowledge, they can be tedious, computationally time-consuming and error-prone as the size of the dataset increases [2]. Feature engineering can help reduce time, space and computational complexity of the machine learning pipeline scripts. Some of the popular ways of performing feature selection [3]when building machine learning models include:

- using Boruta, a feature ranking and selection algorithm based on random forest algorithms,
- using the variable importance computations after training machine learning models,
- feature selection using LASSO regression,
- using subset selection methods
- using the recursive feature elimination to determine variable importance before feeding the data into a machine learning model,
- using genetic algorithms,
- and using simulated annealing algorithms.

Our merged dataset had a total of 53 variables out of which 49% were quantitative (numerical) variables and 51% qualitative (categorical) variables. Even though we attempted to perform automatic feature engineering upfront, due to the space complexity of the dataset and the availability of limited computations resources, this did not produce any results due to out of memory issues. Hence, we used a combination of manual and automatic feature selection methods to reduce our dataset to have 29 (without dummy variables).

Manual selection helped eliminate a lot of the categorical variables that had more than 30 levels as well as the features that suffered from the problem of multicollinearity. Another criteria for manual elimination was the presence of unseen level for a categorical variable in the test set which was not in the train set. For automatic feature selection we used variable importance function after training a random forest classifier as well as the results of a LASSO regression model. Random Forest used the Gini Impurity metric, which gives a measure of variable importance to determine the splits during the training phase. Variable importance is calculated based on the mean decrease in Gini with the highest mean decrease in Gini corresponding to a higher importance to the variable. The Gini Impurity Score plot seen below ranks the variables in decreasing order of importance.

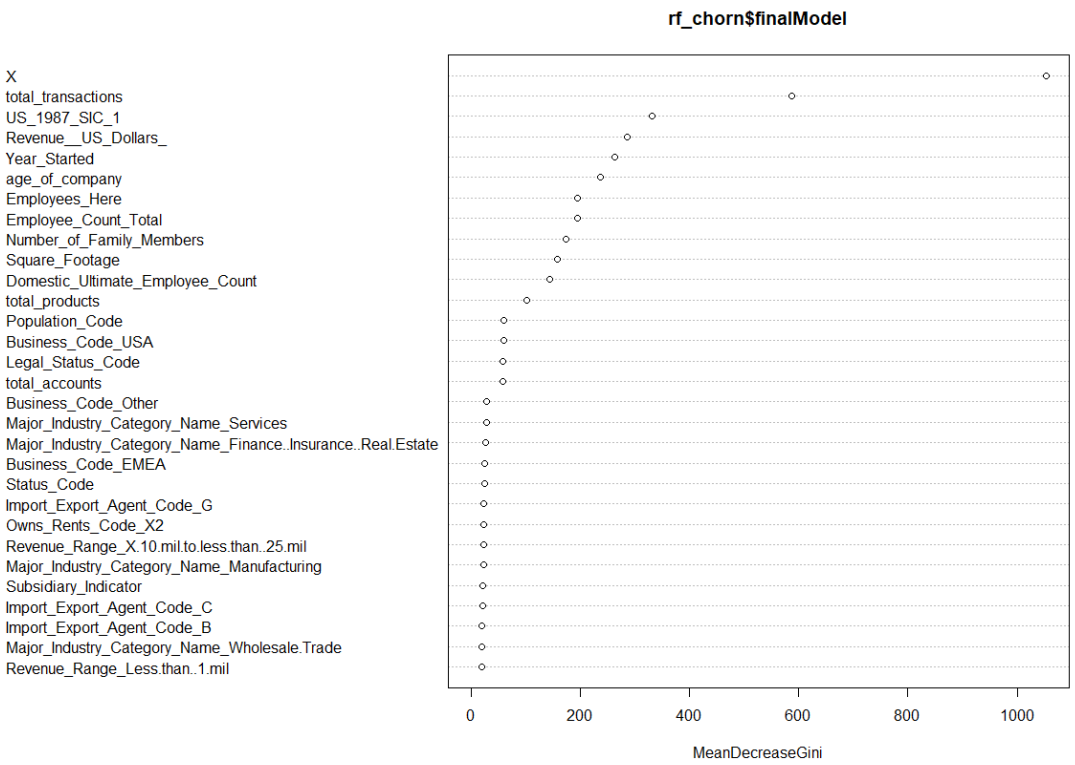
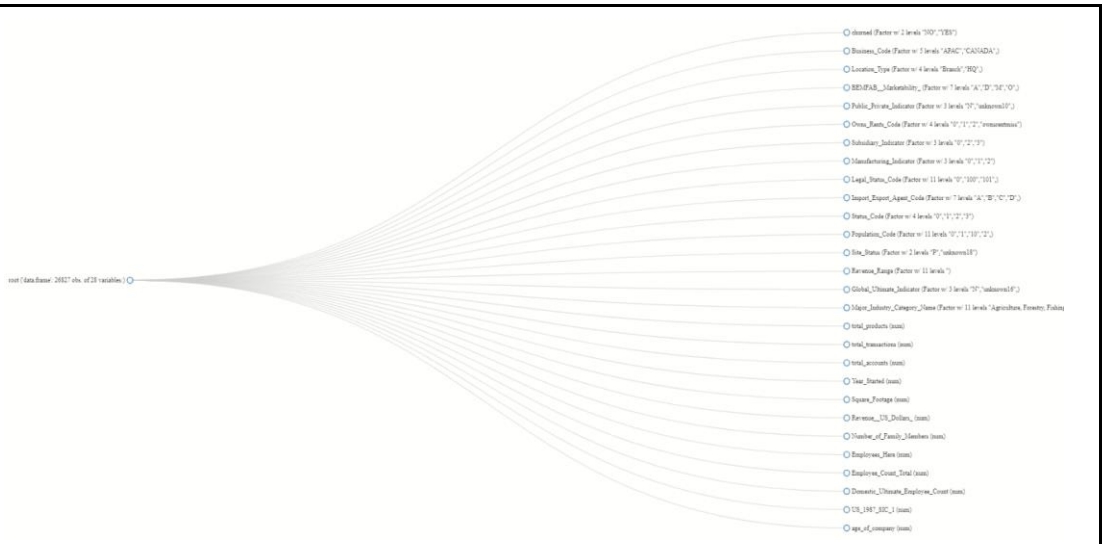
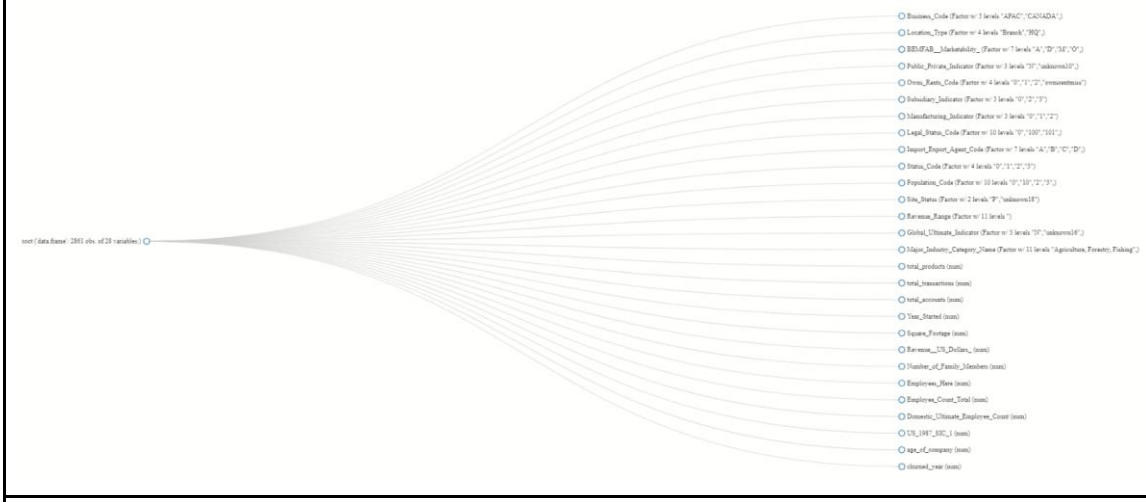


Figure about is the variable importance plot from the Random Forest model.



The final training dataset variables for classification after data preprocessing



The final set of training dataset regression for classification after data preprocessing

The final training and test data after data preparation have the following data distribution.

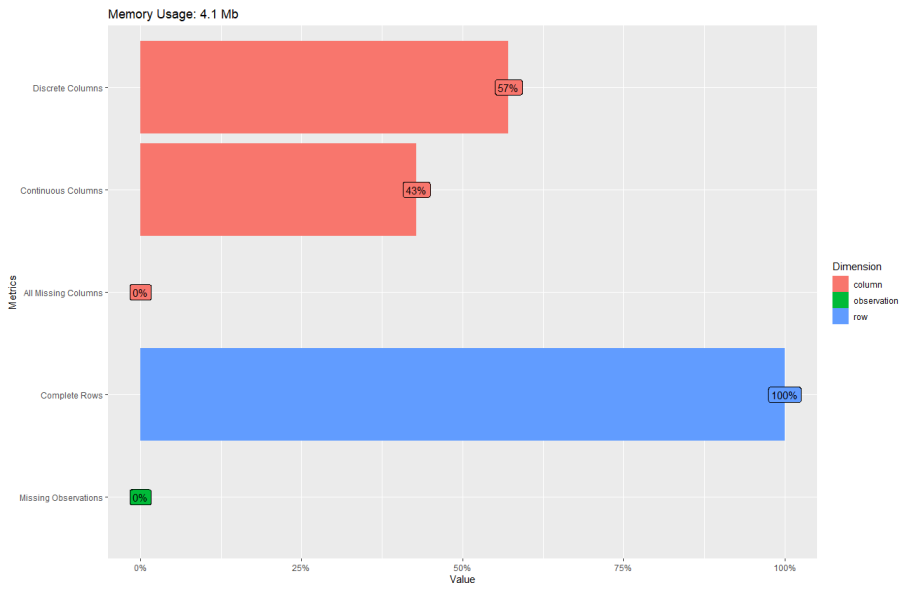


Figure above shows the various types of data from the train set after data treatment.

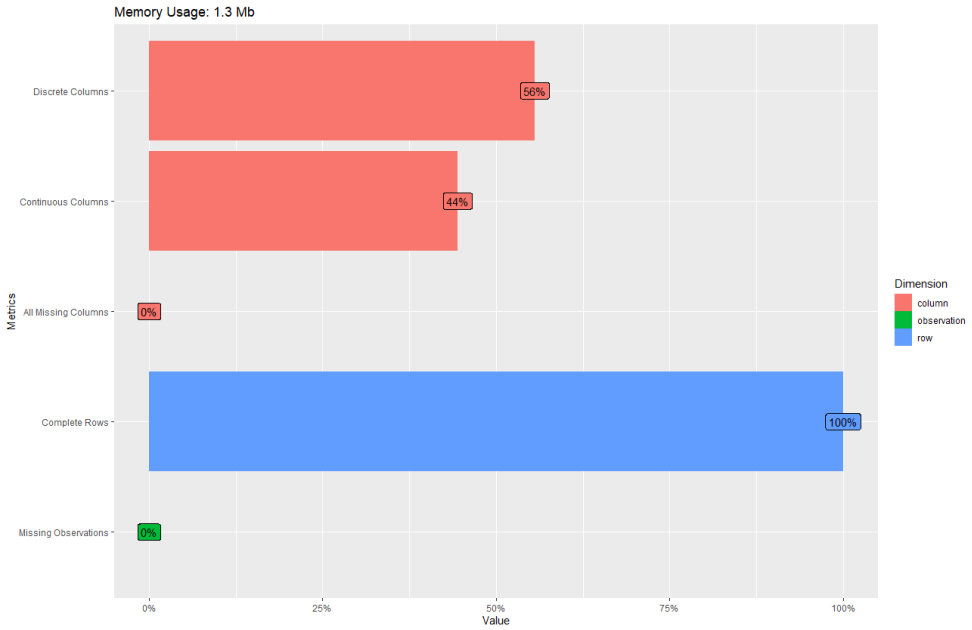


Figure above shows the various types of data from the test set after data treatment.

Table below shows the list of variables after the feature selection process that will be used for building the classification model.

Classification Model - List of Variables after feature selection				
#	**Variable**	**Type**	**Length**	**Description**
1	churned	Num	1	Target variable (1 = Churned; 0 = Not Churned)
2	total_products	Num	8	Number of products/services availed by the company
3	total_transactions	Num	8	Number of transactions / billings made by the company
4	total_accounts	Num	8	Number of accounts (can be department, unit, etc.) a company has
5	Business_Code	Char	10	Business region: USA, EMEA, CANADA, APAC
6	Location_Type	Char	15	HQ, Branch or Single Location

7	BEMFAB__Marketability_	Char	1	It indicates whether the record matched is a marketable record. M Matched, Full Marketing N Unmatched X Matched, Non Marketing A Undeliverable O Out of Business S Undetermined SIC D Delisted Record
8	Year_Started	Num	8	
9	Public_Private_Indicator	Char	1	Indicates whether ownership of the business is public or private.Indicates whether ownership of the business is publc or private. Only applicable on US records. Y=Publicly Held. Only the entity actively trading on US Stock Exchange is identified as publicly held and does not include all family members. N=Not Publicly held or Private
10	Owns_Rents_Code	Num	8	A code value that denotes if the business owns or rents the building it occupies. 0 - Unknown or not applicable

				1 - Business owns the building 2 - Business rents the building
11	Square_Footage	Num	8	The building space this entity operates from within a building as measured in square feet.
12	Manufacturing_Indicator	Num	8	Indicates whether or not manufacturing is done at this location. 0 - Manufacturing is done here. 1 - No manufacturing is done here.
13	Legal_Status_Code	Num	8	003 = corporation 008 = joint venture 012 = partnership of unknown type 013 = proprietorship 050 = government body 100 = cooperative 101= non profit organization 118 = local government body 120 = foreign company

14	Import_Export_Agent_Code	Char	1	A code value that identifies whether the business imports goods or services for re-manufacture or sale, exports products or services to a foreign country, and/or is an agent for goods. A - Import/Export/Agent B - Imports & Exports C - Imports D - Imports & Agent E - Exports & Agent F - Agent - keeps no inventory and does not take title goods G - None or Not Available H - Exports Blank - Not available
15	Status_Code	Num	8	A code value which describes the organizational status of the business. 0 = Single Location - no other entities report to the business 1 = Headquarter/Parent - branches and/or subsidiaries report to the business 2 = Branch - secondary location to a headquarter location

16	Population_Code	Num	1	A code value which describes the residential population for the geographical area where the business is located 0 = Under 1,000 1 = 1,000 to 2,499 2 = 2,500 to 4,999 3 = 5,000 to 9,999 4 = 10,000 to 24,999 5 = 25,000 to 49,999 6 = 50,000 to 99,999 7 = 100,000 to 249,999 8 = 250,000 to 499,999 9 = 500,000 and over
17	Site_Status	Char	1	Indicates the relationship to the business as either prospect or customer. **P** = Prospect **C** = Customer
18	Revenue__US_Dollars_	Num	8	
19	Revenue_Range	Char	30	

20	Number_of_Family_Members	Num	8	The number of family members including the global ultimate and all subsidiaries and branches of the entire family tree worldwide. All family members within a particular tree carry the same count.
21	Employees_Here	Num	8	The number of employees at this location.
22	Employee_Count_Total	Num	8	The total number of employees in the business organization; it should include subsidiary and branch locations.
23	Domestic_Ultimate_Employee_Count	Num	8	
24	Global_Ultimate_Indicator	Char	1	Indicates whether the site record is the Global Ultimate D-U-N-S® within the corporate family tree. Y - Is the global ultimate N - Is not the global ultimate
25	Major_Industry_Category_Name	Char	33	
26	US_1987_SIC_1	Num	8	
27	Subsidiary_Indicator	Num	8	Indicates whether the subject business is more than 50% owned by another organization. 0 - Not a subsidiary 3 - Is a subsidiary

28	Age of Company	Num	8	Age the company in years (Derived variable)

Table belows shows the list of variables after the feature selection process that will be used for building the regression model.

Regression Model - List of variables after feature selection				
#	**Variable**	**Type**	**Length**	**Description**
1	churned	Num	1	Target variable (1 = Churned; 0 = Not Churned)
2	total_products	Num	8	Number of products/services availed by the company
3	total_transactions	Num	8	Number of transactions / billings made by the company
4	total_accounts	Num	8	Number of accounts (can be department, unit, etc.) a company has
5	Business_Code	Char	10	Business region: USA, EMEA, CANADA, APAC
6	Location_Type	Char	15	HQ, Branch or Single Location

7	BEMFAB__Marketability_	Char	1	It indicates whether the record matched is a marketable record. M Matched, Full Marketing N Unmatched X Matched, Non Marketing A Undeliverable O Out of Business S Undetermined SIC D Delisted Record
8	Year_Started	Num	8	
9	Public_Private_Indicator	Char	1	Indicates whether ownership of the business is public or private.Indicates whether ownership of the business is public or private. Only applicable on US records. Y=Publicly Held. Only the entity actively trading on the US Stock Exchange is identified as publicly held and does not include all family members. N=Not Publicly held or Private
10	Owns_Rents_Code	Num	8	A code value that denotes if the business owns or rents the building it occupies. 0 - Unknown or not applicable

				1 - Business owns the building 2 - Business rents the building
11	Square_Footage	Num	8	The building space this entity operates from within a building as measured in square feet.
12	Manufacturing_Indicator	Num	8	Indicates whether or not manufacturing is done at this location. 0 - Manufacturing is done here. 1 - No manufacturing is done here.
13	Legal_Status_Code	Num	8	003 = corporation 008 = joint venture 012 = partnership of unknown type 013 = proprietorship 050 = government body 100 = cooperative 101= non profit organization 118 = local government body 120 = foreign company

14	Import_Export_Agent_ Code	Char	1	A code value that identifies whether the business imports goods or services for re-manufacture or sale, exports products or services to a foreign country, and/or is an agent for goods. A - Import/Export/Agent B - Imports & Exports C - Imports D - Imports & Agent E - Exports & Agent F - Agent - keeps no inventory and does not take title goods G - None or Not Available H - Exports Blank - Not available
15	Status_Code	Num	8	A code value which describes the organizational status of the business. 0 = Single Location - no other entities report to the business 1 = Headquarter/Parent - branches and/or subsidiaries report to the business 2 = Branch - secondary location to a headquarter location

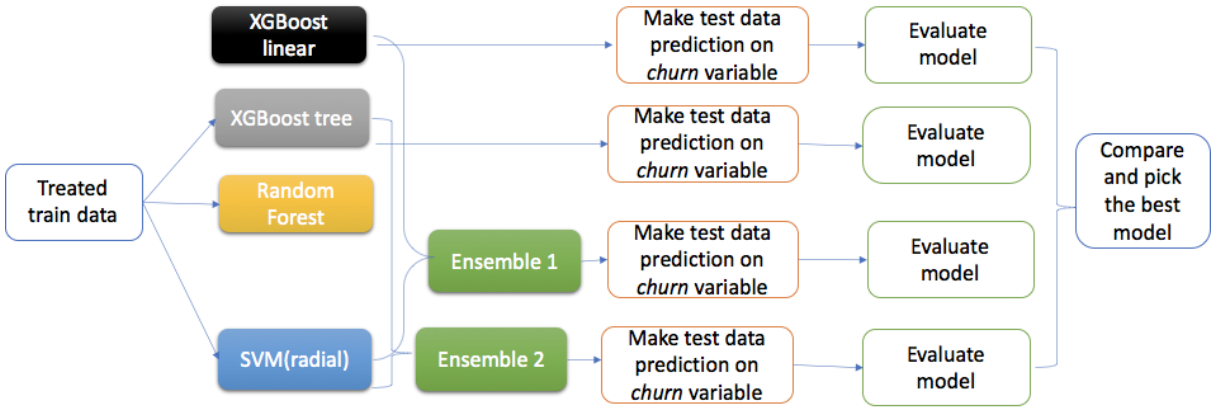
16	Population_Code	Num	1	A code value which describes the residential population for the geographical area where the business is located 0 = Under 1,000 1 = 1,000 to 2,499 2 = 2,500 to 4,999 3 = 5,000 to 9,999 4 = 10,000 to 24,999 5 = 25,000 to 49,999 6 = 50,000 to 99,999 7 = 100,000 to 249,999 8 = 250,000 to 499,999 9 = 500,000 and over
17	Site_Status	Char	1	Indicates the relationship to the business as either prospect or customer. **P** = Prospect **C** = Customer
18	Revenue__US_Dollars_	Num	8	
19	Revenue_Range	Char	30	

20	Number_of_Family_Members	Num	8	The number of family members including the global ultimate and all subsidiaries and branches of the entire family tree worldwide. All family members within a particular tree carry the same count.
21	Employees_Here	Num	8	The number of employees at this location.
22	Employee_Count_Total	Num	8	The total number of employees in the business organization; it should include subsidiary and branch locations.
23	Domestic_Ultimate_Employee_Count	Num	8	
24	Global_Ultimate_Indicator	Char	1	Indicates whether the site record is the Global Ultimate D-U-N-S® within the corporate family tree. Y - Is the global ultimate N - Is not the global ultimate
25	Major_Industry_Category_Name	Char	33	
26	US_1987_SIC_1	Num	8	
27	Subsidiary_Indicator	Num	8	Indicates whether the subject business is more than 50% owned by another organization. 0 - Not a subsidiary 3 - Is a subsidiary

28	Age of Company	Num	8	Age the company in years (Derived variable)
29	Churned Year	Num	8	Year a customer churned (Derived variable)

5. Data Modeling, Model Building

Model building is done in two phases in this project. First we attempt to build the optimal classification model to predict if a customer will churn or not. After this we attempt to build an optimal regression model that can predict the churn date for those customers that are predicted to churn.



Classification model building

Classification Model Building

Addressing the first part of the project involves building a classification model to predict if a customer will churn or not, i.e. the value of the *churned* response variable.

The feature engineered data was then fed into various machine learning models for training. We explored various models that included Naive Bayes, KNN, Logistic

Regression models, Discriminant Analysis models, Decision Tree models and Support Vector Machine models. Following are our findings related to each machine learning model.

Logistic Regression and Linear Discriminant Analysis (LDA) models

Logistic regression is a statistical machine learning model which can classify binary dependent variables. One of the major drawbacks of this model is that it uses liner boundaries to classify the dependent variables. The linearity assumption of this model narrows down most of the nonlinear dependencies between independent variables and dependent variables. In this project, we eliminated the logistic regression model as we found out that there are important non linear relationships. Similarly, we did not consider the LDA model in our classification task as its assumption of data linearity.

Decision Tree models

Decision Trees are popular in classification tasks which have better capabilities of model interpretability. Out of them rpart, and C4.5, C5.0 algorithms are commonly used in machine learning tasks. We considered multiple variants of decision tree algorithms such as rpart, C5.0 and its bagging and boosting versions of Random Forest and XGB Trees. As rpart and C5.0 models perform poorly with this data we consider eliminating these from both independent model process as well as the ensemble model (stacking) process. The final ensemble model contains both Random Forest and XGB Tree models.

Support Vector Machine (SVM) models

SVM is one of the most popular and has been a state of the art for many years. This is a mathematical based model, trying to separate two class variables with a linear hyperplane. This has variants such as linear SVM and Kernel based SVM. Kernel based SVM models are highly capable of separating two class data points when they do not show linear separable notion. In this project, we have considered the SVM radial model both independently as well as in the final ensemble model building.

While multiple models were considered, considering the nature of the data, their fit with various models and the scope of the this being a class project we chose report the results of the following models for the final analysis -

1. XGB tree
2. Random Forest
3. SVM with radial kernel
4. Ensemble model with XGB tree, XGB linear, Random Forest, SVM radial
5. Ensemble model with XGB tree, Random Forest, SVM radial

Resampling Strategies

The resampling strategies we used in this project are K fold cross validation, Repeated cross validation and hold out cross validation. For all the model building tasks in classification and regression we have used 10 fold cross validation other than the ensemble classification model. We have used Repeated cross validation with 5 folds and 3 repeats. Though the repeated cross validation task is highly computationally exhaustive, the final results seem consistent than K fold cross validation with 10 folds. We have used Holdout cross validation method to partition the training validation partitions. We have used random holdout partitions of 80% to 20% for training validation respectively. For each model evaluation, we generated these partitions with the same random seed for comparison purposes.

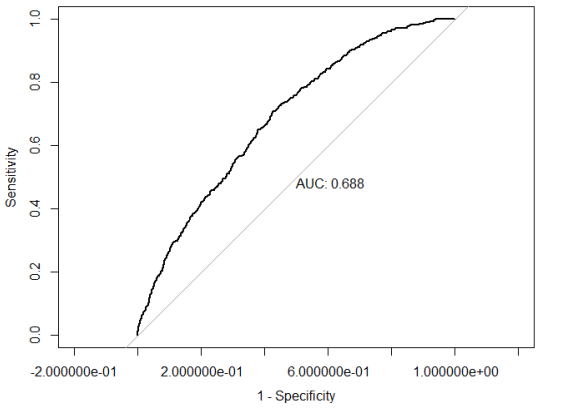
Evaluation Approaches

We use the confusion metrics to and related metrics such as accuracy, sensitivity, specificity, kappa value to compare performance of the various models and pick the optimal model. In addition we also measure the area under the ROC(Receiver Operating Characteristics) curve or AUC curve as a way to evaluate the models.

XGBoost Tree

XGBoost Tree is an implementation of gradient boosting, the most popular type of boosting algorithm. Boosting algorithms performs slow incremental learning. Boosting works as a combination of weak individual small trees. Boosting generates trees sequentially to address the weakness of the previous trees thereby improving performance. In the gradient boosting algorithm new models that predict the residuals or errors of prior models are created and then added together to make the final prediction. XGBoost was built with the goal of execution speed model performance in mind and performs considerably faster than some of the other implementations of gradient boosting. These models can be used for modelling predictive regression and classification problems.

```
Confusion Matrix and Statistics

          Reference
Prediction   NO  YES
       NO  4789  567
       YES    4    5

               Accuracy : 0.8936
                 95% CI : (0.885, 0.9017)
    No Information Rate : 0.8934
    P-Value [Acc > NIR] : 0.4935

                  Kappa : 0.014

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.999165
            Specificity : 0.008741
         Pos Pred Value : 0.894137
         Neg Pred Value : 0.555556
             Prevalence : 0.893383
         Detection Rate : 0.892637
   Detection Prevalence : 0.998322
      Balanced Accuracy : 0.503953

       'Positive' Class : NO
```



Performance measure for the XGBoost model

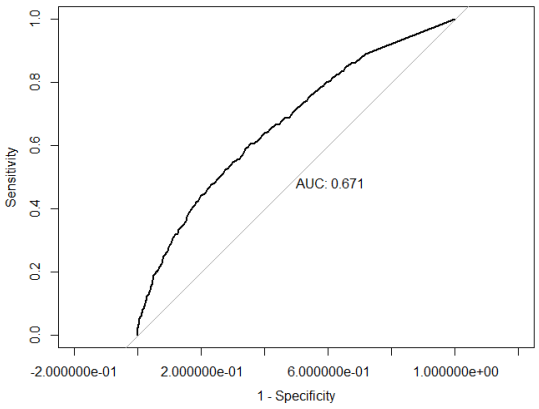
Random Forest

Random Forest is one of the most popular statistical learning methods built on the general idea of bagging. Fundamental idea behind the Random Forest algorithm is to go with the wisdom of the crowds. Individual decision trees that operate as an ensemble make up the *random forest*. Each of the individual trees outputs a class prediction and the class with the majority votes becomes the ensemble models prediction.

The random forest algorithms not only resample the data but also the predictor variables when splitting trees. The ability to de-correlate bagged trees to reduce variance sets it superior to the general bagging method. This is done by picking only a subset of the predictor variable to avoid the problem of correlation between the created trees.

Random Forest - Performance measures

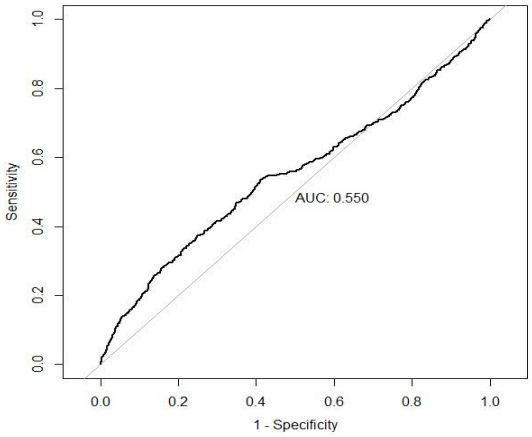
```
Confusion Matrix and Statistics

          Reference
Prediction   NO  YES
       NO  4784  566
       YES    9    6

               Accuracy : 0.8928
                 95% CI : (0.8842, 0.901)
    No Information Rate : 0.8934
    P-Value [Acc > NIR] : 0.5638

                  Kappa : 0.0151

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.99812
            Specificity : 0.01049
         Pos Pred Value : 0.89421
         Neg Pred Value : 0.40000
             Prevalence : 0.89338
         Detection Rate : 0.89171
   Detection Prevalence : 0.99720
      Balanced Accuracy : 0.50431

       'Positive' Class : NO
```



Performance measure for the Random Forest model

Support Vector Machines

```
Confusion Matrix and Statistics

          Reference
Prediction   NO  YES
       NO  4792  572
       YES    1    0

               Accuracy : 0.8932
                 95% CI : (0.8846, 0.9013)
    No Information Rate : 0.8934
    P-Value [Acc > NIR] : 0.5288

                  Kappa : -4e-04

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.9998
            Specificity : 0.0000
         Pos Pred Value : 0.8934
         Neg Pred Value : 0.0000
             Prevalence : 0.8934
         Detection Rate : 0.8932
   Detection Prevalence : 0.9998
      Balanced Accuracy : 0.4999

       'Positive' Class : NO
```



Performance measure for the SVM radial model

Support vector machines are another popularly used machine learning model for regression and classification. It works by finding a hyperplane in an N-dimensional feature space that can separate the data points into two classes. Real-life datasets are not always linearly separable, and Support Vector Machines allow for non-linear decision boundary by specifying kernels. Kernels are a specific way to enrich the original predictor space by taking a low dimensional input space and transforming it into a higher dimensional space. It can convert non-separable problems to a separable problem. The different types of kernel used are linear kernels, polynomial kernels, radial kernels, and sigmoid kernels. We use SVM with a radial kernel as a part of the ensemble model building.

Ensemble model

Ensemble modelling is the technique of combining several base models in an attempt to produce a better performing predictive model than the individual base models it is composed of. We use two types of ensemble models on our dataset. Ensemble1 model combines the results of a XGB Linear, XGB Tree, Random Forest and SVM model with radial kernel using a stacking algorithm. Since XGB Linear and XGB Tree are highly

correlated models we build another ensemble model, Ensemble 2 model that aggregated the results of XGB Tree, Random Forest and SVM model with radial kernel using a stacking algorithm.

Ensemble1 Model with four models (Random Forest, XGBoost, XGBoost Linear, SVM)– Performance measures

```
Confusion Matrix and Statistics

          Reference
Prediction   NO  YES
       NO  4782   11
       YES  549   23

               Accuracy : 0.8956
                 95% CI : (0.8871, 0.9037)
    No Information Rate : 0.9937
    P-Value [Acc > NIR] : 1

                  Kappa : 0.0647

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.89702
            Specificity : 0.67647
         Pos Pred Value : 0.99770
         Neg Pred Value : 0.04021
             Prevalence : 0.99366
         Detection Rate : 0.89133
   Detection Prevalence : 0.89338
      Balanced Accuracy : 0.78674

       'Positive' Class : NO
```

Performance measure for Ensemble1 model

Ensemble model correlation test

It is important to revisit the ensemble model with the inclusive models. If the model correlation is higher than the overall ensemble performance decreases. The model

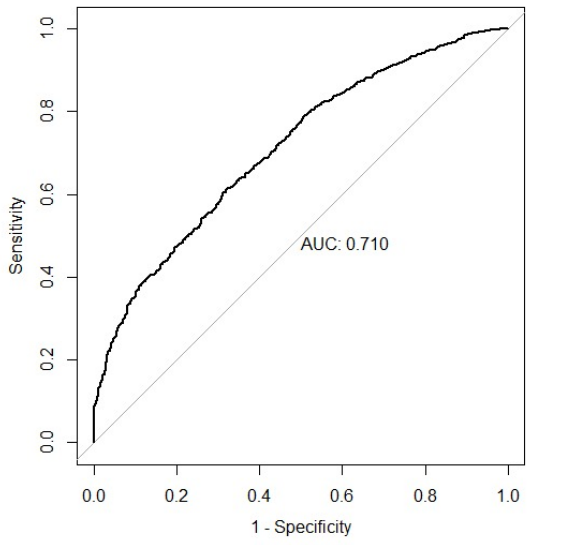
correlations of the ensemble stack is as follows. It is clear that the XGBLinear model is highly correlated with the Random Forest model. Hence, we decided to eliminate the XGBLinear model from the ensemble model.

```
> modelCor(resamples(models))
             svmRadial          rf   xgbTree  xgbLinear
svmRadial  1.00000000 -0.01438178 0.3055100 -0.1235746
rf        -0.01438178  1.00000000 0.6820293  0.7214776
xgbTree    0.30550997  0.68202933 1.0000000  0.6520271
xgbLinear -0.12357465  0.72147763 0.6520271  1.0000000
```

Ensemble2 Model with three models (Random Forest, XGBoost, SVM) – Performance measures

repeated cross validation with

```
Confusion Matrix and Statistics

          Reference
Prediction   NO  YES
       NO  4781   12
       YES  550   22

               Accuracy : 0.8952
                 95% CI : (0.8867, 0.9033)
    No Information Rate : 0.9937
    P-Value [Acc > NIR] : 1

                  Kappa : 0.0614

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.89683
            Specificity : 0.64706
         Pos Pred Value : 0.99750
         Neg Pred Value : 0.03846
             Prevalence : 0.99366
         Detection Rate : 0.89115
   Detection Prevalence : 0.89338
      Balanced Accuracy : 0.77194

       'Positive' Class : NO
```



Performance measure for Ensemble2 model

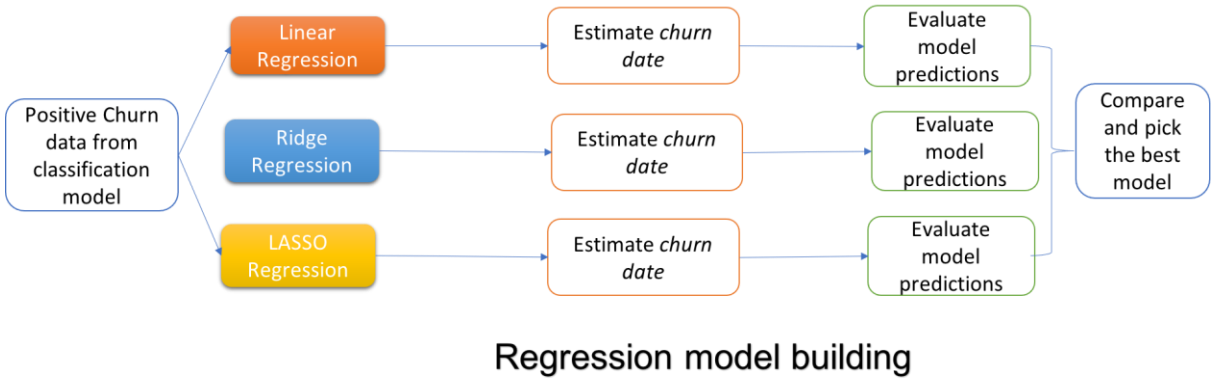
The model correlations have reduced, which validates the performance enhancement of the ensemble model. Even Though, the XGBoost and the Random Forest models show some correlations, the ensemble model performance is comparatively high.

```
> modelCor(resamples(models))
            svmRadial          rf   xgbTree
svmRadial  1.00000000 -0.01438178 0.3055100
rf        -0.01438178  1.00000000 0.6820293
xgbTree    0.30550997  0.68202933 1.0000000
```

Classification models parameters

	Model	**Parameters**
1	XGBoost Tree	nrounds = 50, max_depth = 3, eta = 0.4, gamma = 0, colsample_bytree = 0.8, min_child_weight = 1 and subsample = 0.75
2	Random Forest	ntry = 30
3	SVM radial	kernel = radial

Regression Model



Regression model building

The second part of the project aims to predict a churn date for those customers that are predicted to churn by the previously built classification model. For this a regression model is built using a subset of the dataset which has a *churned_date* and predicting the churned_date for those customers that are predicted to *churn.* As we identified the task of predicting the churn date is a regression related task, we explored various regression models for the task that included Linear Regression Models, Polynomial Regression Models, Ridge Regression Models, and LASSO Regression Models.

Linear Regression

As this is one of the most widely used regression models, and the nature of its simplicity. The basic idea is to establish a relationship between the dependent variable with the independent variables to fit a straight line. We decided to fit our data on a linear model as the training data is considerably less (~2800).

Polynomial Regression

We decided not to choose the polynomial regression in model development because of two reasons. Polynomial regression does not perform well if there are outliers present in the data [5]. Also, while applying polynomial regression a suitable degree should be selected for the dataset[6].

Ridge Regression

Ridge regression is the least square optimization function. This regression integrates all the variables in the model development. When the dataset is highly correlated this model is a better choice and also it reduces the bias.

LASSO regression

Similar to Ridge regression this is also the least square optimization function. This model model performs variable selection activity and filters out irrelevant variables for model development.

Selection of LASSO and Ridge regression as a model:

We used LASSO Regression and Ridge Regression for building the regression model due the computationally easy as it has inclusive feature selection capability while performing the prediction results above the linear regression model.

While multiple models were considered, considering the nature of the data, their fit with various models and the scope of the this being a class project we chose the following models for the final analysis

1. Linear Regression
2. Ridge Regression
3. LASSO Regression

Resampling Strategies

Similar to the classification model building pipeline, the regression model building tasks also resampled with 10 fold cross validation. We have used Holdout cross validation method to partition the training validation partitions. We have used random holdout partitions of 80% to 20% for training validation respectively. For each model evaluation, we generated these partitions with the same random seed for comparison purposes.

Evaluation Approaches

There are various metrics that are used to evaluate regression models. List and define then

The most popular means of evaluating regression models are[7]:

1. R2 or adjusted R2 - This is a measure of how much variability in dependent variables can be explained by the regression model. It is a good measure of how well the model fits the dependent variable.
2. MSE- Mean square error describes the difference between predicted values and original values obtained by squaring the average difference.
3. RMSE - It is the root mean squared error.
4. MAE - Mean absolute error describes the difference between the predicted and the original values obtained via averaging the absolute difference.

The choice of the best regression model will be done after evaluating the RMSE, MSE, R2 and MAE values of the models.

Regression models parameters

	Model	Parameters
1.	Linear	Simple linear regression
2.	Ridge	alpha = 0, lambda = 0.4641589
3.	LASSO	alpha = 1 , lambda = 0.02009233

We have conducted a tune grid process to achieve the optimal model parameters. For the lambda parameter we have considered the performance in the range 1 – 100,000 and retrieved the optimal parameter. We considered that the churn year can be predicted by the time duration of a company being a customer until they churned in historic data.

6. Evaluation, Recommendation, and Conclusion

Classification Models Evaluations Table

Summary of regression model performance measures on test sets of United Communications-Firmographic dataset					
	Accuracy	Kappa	Sensitivity	Specificity	AUCs
XGB tree	0.8936	0.01400	0.999165	0.008741	0.688
Random Forest	0.8928	0.05151	0.99812	0.01049	0.671
SVM	0.8932	-4e-04	0.9998	0.00	0.550
Ensemble 1	0.8956	0.0598	0.8960 (N)	0.6764 (N)	Not Available
Ensemble 2	0.8952	0.0614	0.8968 (N)	0.6471 (N)	0.710

According to the above classification model evaluation results, it is clear that Ensemble model 2 (with Random Forest, XGB Tree and SVM radial) performed better than other model buildings. Ensemble 1 model also performed almost upto the level of Ensemble model 2, still we preferred Ensemble model 2 as the best model, after consideration of model correlations of the ensemble model. The initial four models Random Forest, XGB Tree, SVM radial and XGB linear, show high correlations, hence we decided to drop the XGB Linear model from the ensemble model.

None of the independent models show better predictive performances than the ensemble model 2. It is important to see that independent models could not incorporate the class variable imbalance hence they show poor sensitivity and specificity balance, and abnormal accuracy results.

Regression Models – Comparison Table

Summary of regression model performance measures on test sets of United Communications-Firmographic dataset			
	RMSE	R^2	MAE
Linear Regression	0.7303308	0.9785898	0.5005579
Ridge Regression	0.9025569	0.9684252	0.6153540
LASSO Regression	0.7280747	0.9775723	0.4950664

The above table describes the results obtained for churned date prediction utilizing linear regression, ridge regression, and lasso regression. Based on comparing the metric in the table, lasso regression performs the best considering the RMSE and MAE. Lasso regression performs variable selection while model development. In our model development, Lasso has selected 17 variables, all these variables are available at appendix.

Recommendation

Based on the above discussions and results from the performance table we recommend the following:

Use an ensemble model with random forest, XGBoost, and SVM(kernel:radial) for classification models to predict if a customer will churn or not.

Use a LASSO regression model to predict the churn date of a customer who has been classified to churn.

These recommendations are provided considering the computational complexity as well as performance measures of the models. A list of variables that can highly influence if a customer will churn and the churn date is provided in Appendix. United Communication should consider using this information to reduce the churn rate of customers.

Conclusion

In this project classification and regression models are used to predict if a customer will churn and the churned date, if the customers are classified to churn.

We merged two datasets for the final prediction. The merging results in a high number of missing values. We performed imputations techniques. We found that initially the dataset was quite computational costly. Hence, we performed both automatic and manual feature selection to reduce the computational load for model development. Our assumption is that due to this practice of feature selection we have to eliminate some variables that affect our final prediction results both in the classification and the regression models. From the dataset, we predict whether the accounts have churn or not churn where we utilize classification based machine learning models. In addition to that an estimated churn date is also predicted utilizing regression models.

We ensemble 2 (Random Forest, XGB Tree, SVMRadial) models for churn classification tasks. Here, both the ensemble models performed well with sensitivity, specificity and AUC metrics. As the data is highly imbalanced, XGBoost, Random Forest models individually performed less.

For the churn year prediction, we have used LASSO regression as the final model. The main reason was the interpretability with less number of variables (17) were easy and yet the prediction performances are above the level of linear regression model. As per the classification results, we would like to acknowledge the clients that the churn rate is lower according to the provided test data.

In conclusion, we could have used time-series analysis to predict the customer churn date but due to less number of customer churned data it was not an optimal approach. We propose to consider this as a future work with required dataset.

1. **REFERENCES**

1) Raheel Shaik, Feature Selection Techniques in Machine Learning with Python, 28 October 2018, retrieved from https://towardsdatascience.com/feature-selection-techniques-in-machine-learning-with-python-f24e7da3f36e
2) Will Koehrsen, Why Automated Feature Engineering Will Change the Way You Do Machine Learning, 9 August 2018, retrieved from https://towardsdatascience.com/why-automated-feature-engineering-will-change-the-way-you-do-machine-learning-5c15bf188b96
3) Selva Prabhakaran, Feature Selection – Ten Effective Techniques with Examples, retrieved from https://www.machinelearningplus.com/machine-learning/feature-selection/
4) Raikwal, J. S., & Saxena, K. (2012). Performance evaluation of SVM and k-nearest neighbor algorithm over medical data set. *International Journal of Computer Applications*, *50*(14).
5) https://towardsdatascience.com/introduction-to-linear-regression-and-polynomial-regression-f8adc96f31cb
6) https://www.geeksforgeeks.org/advantages-and-disadvantages-of-different-regression-models/
7) https://towardsdatascience.com/what-are-the-best-metrics-to-evaluate-your-regression-model-418ca481755b

2. **APPENDIX**

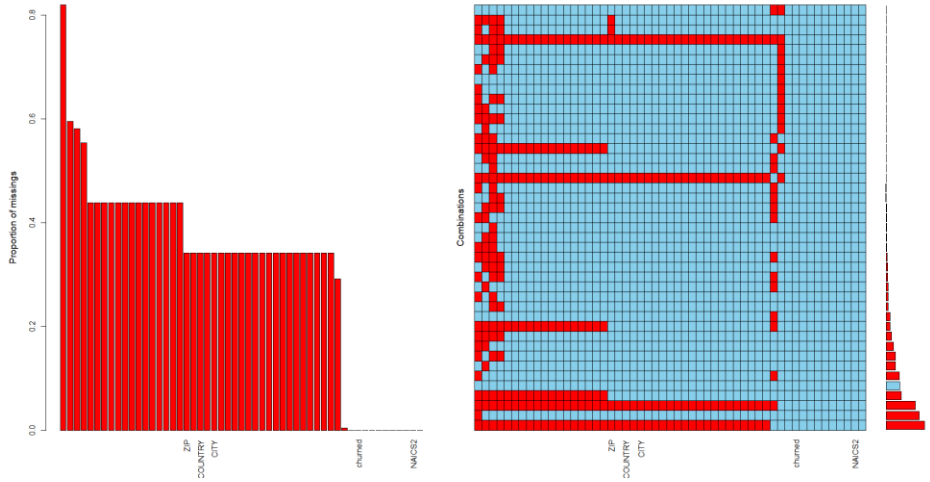


Fig 1 - Figure showing proportion of missing values in the train set

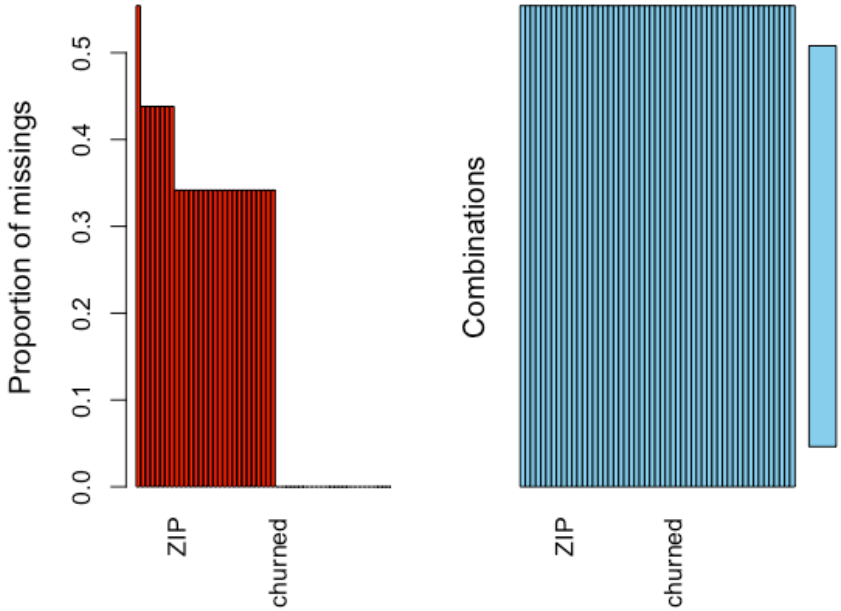


Fig 2 - Figure showing proportion of missing values in the train set after bag imputation

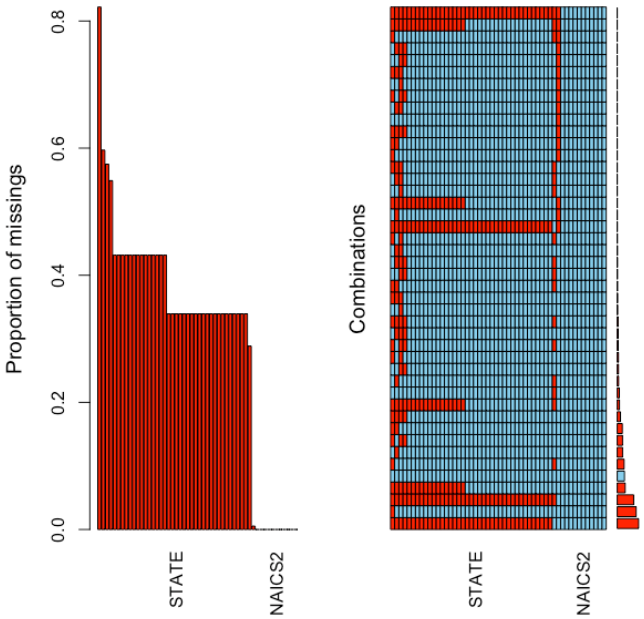


Fig 3 - Figure showing proportion of missing values in the test set

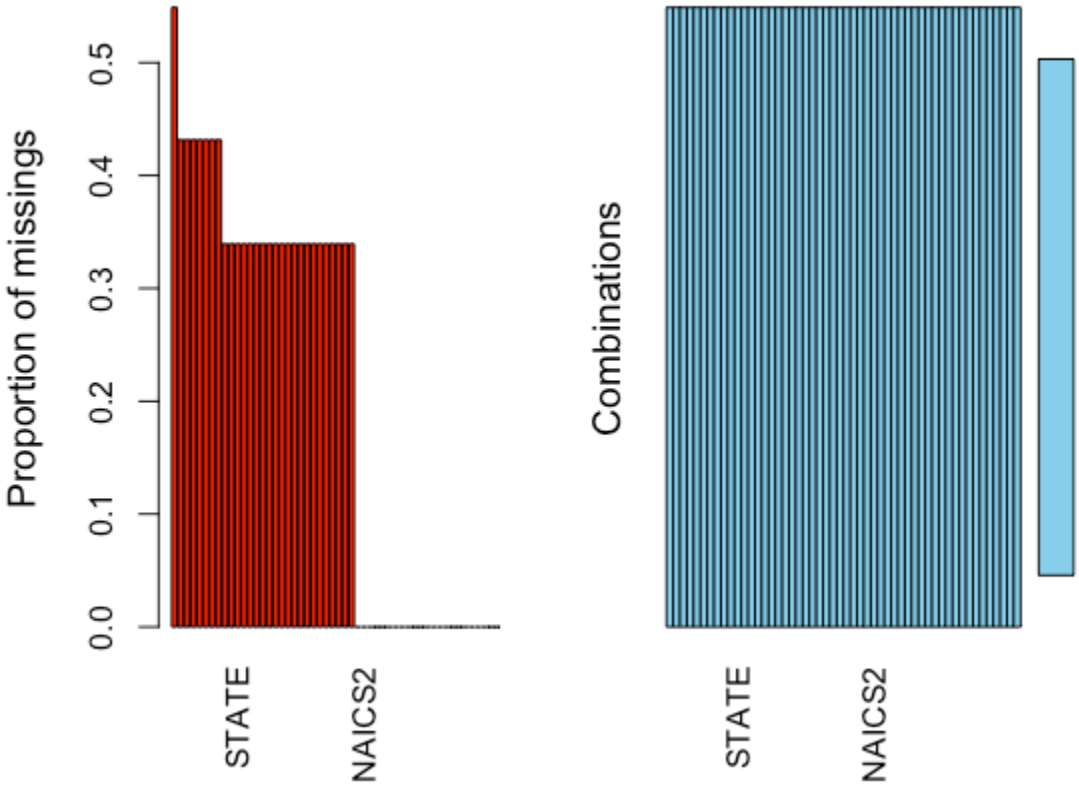


Fig 4 - Figure showing proportion of missing values in the test set after bag imputation.

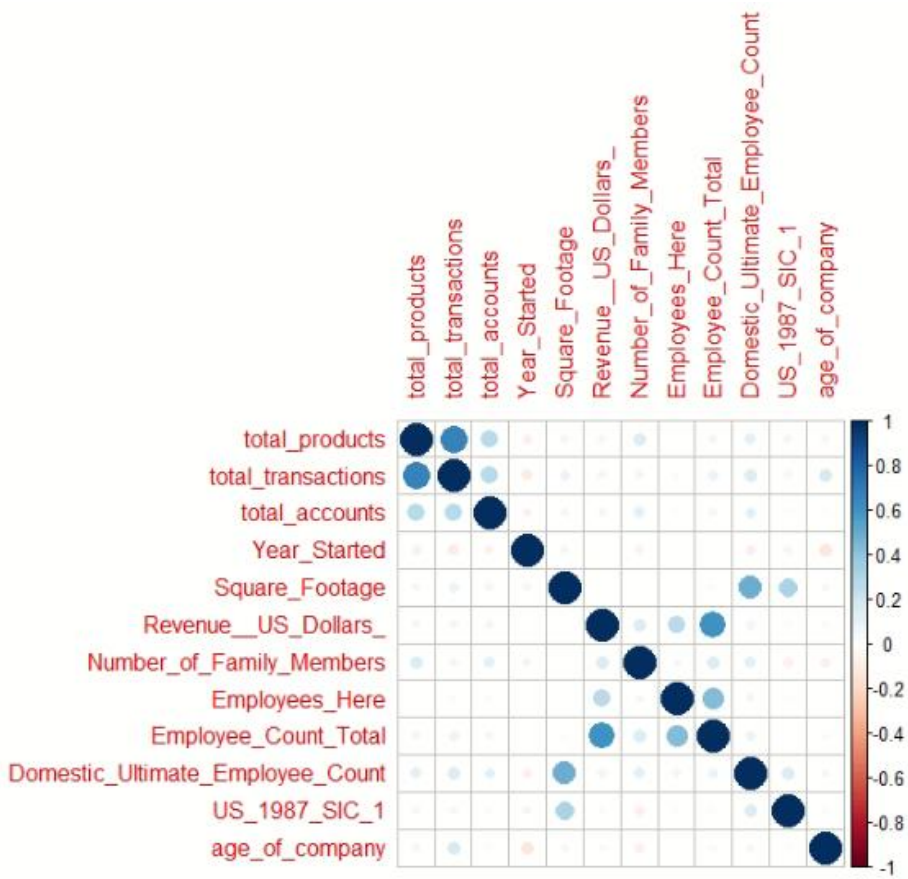
```
-- Data Summary ------------------------
                           Values
Name                       train_data
Number of rows             26827
Number of columns          53
_____________________
Column type frequency:
  factor                   27
  numeric                  26
_____________________
Group variables            None
```

```
-- Variable type: factor ---------------------------------------------------------------------------------------------------------------
# A tibble: 27 x 6
   skim_variable                 n_missing complete_rate ordered n_unique top_counts
 * <chr>                             <int>         <dbl> <lgl>      <int> <chr>
 1 churn_date                            0             1 FALSE        493 emp: 23966, 5/2: 179, 3/6: 57, 12/: 41
 2 Company_Creation_Date                 0             1 FALSE      24020 18A: 255, 03A: 151, 09J: 41, 18F: 32
 3 HQ_Country                            0             1 FALSE         99 Uni: 8193, emp: 7829, CON: 2554, Uni: 1943
 4 NAICS2                                0             1 FALSE         29 emp: 7829, 54-: 3533, 33-: 2574, 52-: 2069
 5 NAICS3                                0             1 FALSE        100 emp: 7829, 541: 3533, 999: 1482, 522: 899
 6 Business_Code                      9167         0.658 FALSE          4 USA: 8623, EME: 5736, APA: 2917, CAN: 384
 7 COUNTRY                            9167         0.658 FALSE          6 Uni: 8412, emp: 6338, Uni: 2040, Fra: 687
 8 STATE                              9167         0.658 FALSE        299 emp: 6195, CA: 993, NY: 717, TX: 567
 9 CITY                               9167         0.658 FALSE       5423 Lon: 640, New: 447, Sin: 286, Was: 211
10 ZIP                                9169         0.658 FALSE       9300 -: 904, .: 392, emp: 222, 0: 137
11 Location_Type                      9167         0.658 FALSE          4 Sin: 7249, HQ: 6363, emp: 2586, Bra: 1462
12 BEMFAB__Marketability_             9167         0.658 FALSE          7 M: 12662, emp: 2586, O: 669, X: 617
13 Public_Private_Indicator           9167         0.658 FALSE          3 N: 14740, emp: 2586, Y: 334
14 Small_Business_Indicator           9167         0.658 FALSE          3 emp: 9960, N: 4624, Y: 3076
15 Minority_Owned_Indicator           9167         0.658 FALSE          3 emp: 9929, N: 7511, Y: 220
16 Import_Export_Agent_Code           9167         0.658 FALSE          8 G: 11441, emp: 2586, B: 1520, C: 1126
17 Site_Status                        9167         0.658 FALSE          2 P: 15074, emp: 2586
18 Revenue_Range                      9167         0.658 FALSE         11 Les: 6220, emp: 2586, $1 : 1758, $10: 1395
19 Global_Ultimate_Indicator          9167         0.658 FALSE          3 N: 11769, Y: 3267, emp: 2624
20 Major_Industry_Category_Name       9167         0.658 FALSE         11 Ser: 7088, emp: 2586, Man: 2239, Fin: 2151
21 Related_Industries                 9167         0.658 FALSE       2675 emp: 2586, Oth: 971, Law: 909, Non: 799
22 Line_of_Business                   9167         0.658 FALSE        806 emp: 2586, Leg: 927, Bus: 739, Man: 568
23 Chief_Executive_Officer_Gender_C   9167         0.658 FALSE          4 emp: 11139, M: 5280, F: 961, B: 280
24 Chief_Executive_Officer_Title      9167         0.658 FALSE        195 emp: 4968, Pre: 3307, Dir: 1902, Chi: 1799
25 First_Executive_Title              9167         0.658 FALSE        228 emp: 12061, Pre: 1022, Chi: 676, Vic: 542
26 Second_Executive_Title             9167         0.658 FALSE        255 emp: 12681, Chi: 718, Vic: 646, Pre: 438
27 Third_Executive_Title              9167         0.658 FALSE        264 emp: 13169, Vic: 703, Chi: 393, Sen: 235
```

```
-- Variable type: numeric --------------------------------------------------------------------------------------------------------------------
# A tibble: 26 x 11
   skim_variable                 n_missing complete_rate     mean       sd       p0       p25       p50       p75          p100 hist
 * <chr>                             <int>         <dbl>    <dbl>    <dbl>    <dbl>     <dbl>     <dbl>     <dbl>         <dbl> <chr>
 1 Company_Number                        0             1 3.96e+5 1.34e+ 5      23   317602     435687     496236.       558988 ▁▁▁▁▁▁▁▁▁▇
 2 churned                               0             1 1.07e-1 3.09e- 1       0        0          0          0            1 ▇▁▁▁▁▁▁▁▁
 3 total_products                        0             1 1.51e+0 8.85e- 1       1        1          1          2           10 ▇▁▁▁▁▁▁▁▁
 4 total_transactions                    0             1 3.88e+1 6.30e+ 1       1        8         19         46         2099 ▇▁▁▁▁▁▁▁▁
 5 total_revenue                         0             1 2.79e+4 3.64e+ 5 -657715.     33.1      592.      4573.     45899982. ▇▁▁▁▁▁▁▁▁
 6 total_usage                           0             1 5.98e+5 1.90e+ 7       0      142       4496      37444.    2758029464 ▇▁▁▁▁▁▁▁▁
 7 total_accounts                      114         0.996 2.71e+0 5.03e+ 1       1        1          1          1         6087 ▇▁▁▁▁▁▁▁▁
 8 HQ_Employee_Count                  7829         0.708 3.11e+4 1.06e+ 5       0       15        501      15973      3600000 ▇▁▁▁▁▁▁▁▁
 9 Year_Started                      11753         0.562 1.78e+3 6.07e+ 2       0     1973       1997       2008         2017 ▁▁▁▁▁▁▁▁▁▇
10 Owns_Rents_Code                   11753         0.562 4.67e-1 7.91e- 1       0        0          0          1            2 ▇▁▁▁▁▁▁▁▁▂
11 Square_Footage                    11753         0.562 2.02e+4 1.58e+ 5       0        0          0       6400.     11000000 ▇▁▁▁▁▁▁▁▁
12 Subsidiary_Indicator              11753         0.562 9.36e-1 1.39e+ 0       0        0          0          3            3 ▇▁▁▁▁▁▁▁▂▁
13 Manufacturing_Indicator           11753         0.562 4.38e-1 4.96e- 1       0        0          0          1            1 ▇▁▁▁▁▁▁▁▁▆
14 Legal_Status_Code                 11753         0.562 6.21e+0 1.86e+ 1       0        3          3          3          120 ▇▁▁▁▁▁▁▁▁
15 Currency_Code                     14865         0.446 9.12e+2 1.97e+ 3      10       20         20        160         9450 ▇▁▁▁▁▁▁▁▁
16 Status_Code                       11753         0.562 6.16e-1 6.56e- 1       0        0          1          1            2 ▇▁▁▁▁▁▁▁▁▃
17 Population_Code                   11753         0.562 4.29e+0 4.26e+ 0       0        0          5          9            9 ▇▁▁▁▁▁▁▁▁▆
18 Hierarchy_Code                    11753         0.562 1.45e+0 1.85e+ 0       0        0          1          2           19 ▇▁▁▁▁▁▁▁▁
19 Revenue__US_Dollars_              15576         0.419 5.50e+8 4.23e+ 9       1  1500000   14847384  101133930. 226000000000 ▇▁▁▁▁▁▁▁▁
20 Number_of_Family_Members          11753         0.562 3.98e+2 1.90e+ 3       0        0          4        105        61720 ▇▁▁▁▁▁▁▁▁
21 Employees_Here                    15973         0.405 2.51e+2 3.72e+ 3       1        6         25        100       250000 ▇▁▁▁▁▁▁▁▁
22 Employee_Count_Total              11753         0.562 1.31e+3 1.02e+ 4       0        3         30        210       380300 ▇▁▁▁▁▁▁▁▁
23 Domestic_Ultimate_Employee_Count  11753         0.562 1.94e+2 2.54e+ 3       0        0          6         50       250000 ▇▁▁▁▁▁▁▁▁
24 Domestic_Ultimate_Revenue         11753         0.562 1.83e+9 1.11e+10       0   133063.   6000000  118338822. 239000000000 ▇▁▁▁▁▁▁▁▁
25 US_1987_SIC_1                     11753         0.562 2.23e+7 3.41e+ 7     172     6411       8742   50850401.     99998999 ▇▁▁▁▁▁▁▁▂
26 US_1987_SIC_2                     21982         0.181 1.49e+7 2.76e+ 7     172     5632       8052   17319903     96210302 ▇▁▁▁▁▁▁▁▁
```

```
Rows: 26,827
Columns: 53
$ Company_Number                      <int> 23, 182, 246, 282, 291, 344, 428, 510, 524, 597, 791, 897, 1067, 1255, 1303, 1596, 1614, 1618, 229
$ churned                             <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
$ churn_date                          <fct> , , , , , , , , , , , , , , , , , , , , , 3/2/2020, , 7/1/2019, , , , , , , , , , , , , , , , , , , ,
$ Company_Creation_Date               <fct> 03APR1999:04:21:00, 03APR1999:04:21:00, 03APR1999:04:21:00, 03APR1999:04:21:00, 03APR1999:04:21:00,
$ total_products                      <int> 1, 1, 5, 3, 1, 2, 5, 1, 6, 1, 3, 1, 1, 3, 1, 5, 2, 2, 3, 1, 1, 4, 2, 1, 1, 4, 2, 5, 4, 1, 1, 4, 2,
$ total_transactions                  <int> 6, 28, 534, 27, 2, 3, 333, 1, 584, 75, 362, 2, 1, 110, 2, 263, 79, 63, 113, 5, 6, 346, 19, 4, 88, 1
$ total_revenue                       <dbl> 48.19, 1508.09, 804847.30, 3388.50, 2.40, 18.20, 384343.74, 0.05, 751244.47, 23025.63, 289624.02, 8
$ total_usage                         <dbl> 13, 2955, 22965108, 170623, 228, 2, 412012, 2, 39788592, 38016, 2725932, 2, 2, 53225, 6, 24626320,
$ total_accounts                      <int> 1, 1, 4, 1, 1, 2, 1, 1, 2794, 1, 1, 1, 1, 2, 1, 2901, 1, 1, 3, 1, 1, 5, 1, 2, 2, 1, 1, 1, 1, 1, 4,
$ HQ_Employee_Count                   <int> 1, 27, 8500, 90000, 60000, 459000, 18000, 264000, 2400, 78600, NA, 8149, 99827, 15800, 309000, 540
$ HQ_Country                          <fct> United States, United States, United States, United States, United States, CONFLICT, United States,
$ NAICS2                              <fct> "0", "81-Other Services (except Public Administration)", "42-Wholesale Trade", "72-Accommodation ar
$ NAICS3                              <fct> "0", "813-Religious, Grantmaking, Civic, Professional, and Similar Organizations", "424-Merchant Wh
$ Business_Code                       <fct> NA, USA, USA, USA, USA, USA, USA, USA, USA, USA, USA, NA, NA, USA, NA, USA, USA, USA, USA, NA, USA,
$ COUNTRY                             <fct> NA, United States, United States, United States, United States, United States, United States, Unite
$ STATE                               <fct> NA, DC, WA, KY, MA, IL, MA, NY, IL, CA, CA, NA, NA, MA, NA, OH, CA, KY, CA, NA, GA, CA, TX, CA, GA,
$ CITY                                <fct> NA, Washington, Seattle, Louisville, Littleton, Chicago, North Andover, Purchase, Chicago, Pleasant
$ ZIP                                 <fct> NA, 20002, 30091, 60515, 60008, 80502, 2368, 10589, 60601, 94560, 94066, NA, NA, 2062, NA, 75240, 9
$ Location_Type                       <fct> NA, Single Location, HQ, HQ, Branch, Branch, Branch, HQ, Branch, Branch, , NA, NA, Branch, NA, , HQ
$ BEMFAB__Marketability_              <fct> NA, M, M, M, M, M, M, M, O, A, , NA, NA, M, NA, , M, M, O, NA, M, M, M, A, O, O, M, A, M, NA, A, M,
$ Year_Started                        <int> NA, 2014, 1986, 2000, 0, 0, 0, 1919, 0, 0, NA, NA, NA, 0, NA, NA, 1982, 0, 0, NA, 0, 1922, 2014, 20
$ Public_Private_Indicator            <fct> NA, N, N, N, N, N, N, Y, N, N, , NA, NA, N, NA, , Y, N, N, NA, N, N, N, N, N, N, N, N, N, NA, N, N,
$ Small_Business_Indicator            <fct> NA, N, N, Y, N, N, N, N, N, N, , NA, NA, N, NA, , N, N, N, NA, N, N, N, Y, N, Y, N, N, N, NA, Y, N,
$ Minority_Owned_Indicator            <fct> NA, N, N, N, N, N, N, N, N, N, , NA, NA, N, NA, , N, N, N, NA, N, N, N, N, N, N, N, N, N, NA, N, N,
$ Owns_Rents_Code                     <int> NA, 0, 0, 2, 0, 0, 0, 1, 2, 0, NA, NA, NA, 1, NA, NA, 1, 0, 0, NA, 0, 2, 2, 0, 0, 0, 2, 0, 0, NA, 0
$ Square_Footage                      <int> NA, 2941, 90000, 1000, 40289, 15519, 59947, 36586, 0, 0, NA, NA, NA, 5660, NA, NA, 15782, 24181, 0,
$ Subsidiary_Indicator                <int> NA, 0, 3, 0, 0, 0, 0, 0, 0, 0, NA, NA, NA, 0, NA, NA, 0, 0, 0, NA, 0, 3, 3, 0, 0, 0, 3, 0, 0, NA, 0
$ Manufacturing_Indicator             <int> NA, 1, 1, 1, 0, 1, 0, 0, 1, 1, NA, NA, NA, 0, NA, NA, 0, 1, 1, NA, 1, 0, 1, 1, 1, 1, 1, 0, 1, NA, 1
$ Legal_Status_Code                   <int> NA, 3, 3, 3, 0, 0, 0, 3, 0, 0, NA, NA, NA, 0, NA, NA, 3, 0, 0, NA, 0, 3, 3, 0, 0, 0, 3, 0, 0, NA, 0
$ Import_Export_Agent_Code            <fct> NA, G, B, G, G, G, G, B, G, G, , NA, NA, G, NA, , B, G, G, NA, G, B, C, G, G, G, B, G, G, NA, G, C,
$ Currency_Code                       <int> NA, 20, 20, 20, NA, NA, NA, 20, NA, NA, NA, NA, NA, NA, NA, NA, 20, NA, NA, NA, NA, 20, 20, 20, NA,
$ Status_Code                         <int> NA, 0, 1, 1, 2, 2, 2, 1, 2, 2, NA, NA, NA, 2, NA, NA, 1, 2, 0, NA, 2, 1, 1, 0, 2, 0, 0, 2, 2, NA, 0
$ Population_Code                     <int> NA, 9, 9, 9, 9, 9, 9, 9, 9, 9, NA, NA, NA, 9, NA, NA, 9, 9, 9, NA, 9, 9, 9, 9, 6, 6, 9, 8, 8, NA, 9
$ Site_Status                         <fct> NA, P, P, P, P, P, P, P, P, P, , NA, NA, P, NA, , P, P, P, NA, P, P, P, P, P, P, P, P, P, NA, P, P,
$ Hierarchy_Code                      <int> NA, 0, 2, 1, 1, 2, 2, 1, 0, 1, NA, NA, NA, 1, NA, NA, 1, 1, 0, NA, 2, 2, 2, 0, 0, 0, 3, 1, 2, NA, 0
$ Revenue__US_Dollars_                <dbl> NA, 391772, 4915255552, 1454773, NA, NA, NA, 62799000000, NA, NA, NA, NA, NA, NA, NA, NA, 87119200
$ Revenue_Range                       <fct> NA, Less than $1 mil, $1 bil and above, $1 mil to less than $5 mil, Less than $1 mil, Less than $1
$ Number_of_Family_Members            <int> NA, 0, 403, 3, 887, 867, 331, 2331, 0, 1372, NA, NA, NA, 111, NA, NA, 56, 10927, 0, NA, 253, 2220,
$ Employees_Here                      <int> NA, 1, 350, 5, 52, 46, 110, 1500, 2, 46, NA, NA, NA, 1, NA, NA, 277, 1000, NA, NA, 10, 402, 1000, 1
$ Employee_Count_Total                <int> NA, 1, 4369, 20, 0, 0, 0, 264000, 0, 0, NA, NA, NA, 0, NA, NA, 1700, 0, 0, NA, 0, 17718, 11900, 1,
$ Domestic_Ultimate_Employee_Count    <int> NA, 1, 25, 5, 2500, 4, 277, 1500, 2, 1500, NA, NA, NA, 1000, NA, NA, 277, 250, 0, NA, 82, 1329, 100
$ Domestic_Ultimate_Revenue           <dbl> NA, 391772, 8073700000, 1454773, 48238000000, 5493146777, 3990487000, 62799000000, 0, 12866757000,
$ Global_Ultimate_Indicator           <fct> NA, N, N, Y, N, N, N, Y, N, N, , NA, NA, N, NA, , Y, N, N, NA, N, N, N, N, N, N, N, N, N, NA, N, N,
$ Major_Industry_Category_Name        <fct> NA, "Services", "Wholesale Trade", "Retail Trade", "Manufacturing", "Services", "Manufacturing", "M
$ Related_Industries                  <fct> NA, "Non-Profit & Professional Orgs.", "Chemicals; Industrial Machinery & Equipment; Management Cor
$ Line_of_Business                    <fct> NA, "Religious organizations", "Chemicals and allied products, nec", "Eating places", "Computer per
$ US_1987_SIC_1                       <int> NA, 86610106, 51699907, 58120601, 3577, 8742, 3561, 20960000, 8111, 5651, NA, NA, NA, 367402, NA, N
$ US_1987_SIC_2                       <int> NA, NA, 51910101, NA, NA, NA, NA, 20870000, NA, NA, NA, NA, NA, NA, NA, NA, 23290100, 6794, NA, NA,
$ Chief_Executive_Officer_Gender_C    <fct> NA, M, M, M, , B, M, F, M, , , NA, NA, , NA, , M, M, , NA, M, M, M, , M, , , , M, NA, M, M, M, M, N
$ Chief_Executive_Officer_Title       <fct> NA, President, Chief Executive Officer, Member, , Branch Manager, Branch Manager, Chairman of the B
$ First_Executive_Title               <fct> NA, Vice-President, President, Manager, General Manager, Officer, Administrator, Vice-Chairman, , ,
$ Second_Executive_Title              <fct> NA, Administrative Assistant, Chief Financial Officer, , Manager, Computer Specialist, Technology,
$ Third_Executive_Title               <fct> NA, , Chief Marketing Officer, , Manager, Operations Staff, , Chief Executive Officer, , , , NA, NA
```

Figure 5 - Summary Statistics of train set



Correlation plot of numeric columns in train set after data preparation.

Random Forest Model Results

```
> rf_chorn
Random Forest

21462 samples
   27 predictor
    2 classes: 'NO', 'YES'

Recipe steps: novel, unknown, dummy, zv
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 19316, 19316, 19315, 19316, 19315, 19316, ...
Resampling results across tuning parameters:

  mtry  ROC        Sens       Spec
   2    0.5999960  1.0000000  0.00000000
  43    0.6818880  0.9967143  0.03058492
  85    0.6793075  0.9868568  0.05460622

ROC was used to select the optimal model using the largest value.
The final value used for the model was mtry = 43.
```

Add Description

> importance(rf_chorn$finalModel)

	MeanDecreaseGini
total_products	8.643668e+01
total_transactions	4.929721e+02
total_accounts	4.596630e+01
Year_Started	2.341718e+02
Square_Footage	1.363909e+02
Revenue__US_Dollars_	2.453420e+02
Number_of_Family_Members	1.584949e+02
Employees_Here	1.698558e+02
Employee_Count_Total	1.711528e+02
Domestic_Ultimate_Employee_Count	1.251078e+02
US_1987_SIC_1	2.982065e+02
age_of_company	3.487229e+02
Business_Code_CANADA	1.019800e+01
Business_Code_EMEA	2.323531e+01
Business_Code_Other	2.412905e+01
Business_Code_USA	4.877882e+01
Location_Type_HQ	1.345739e+01
Location_Type_Single.Location	1.560833e+01
Location_Type_unknown8	9.736721e-01
BEMFAB__Marketability__D	7.032749e+00
BEMFAB__Marketability__M	1.732868e+01
BEMFAB__Marketability__O	8.843976e+00

Add Description

BEMFAB__Marketability__S	4.092094e+00	
BEMFAB__Marketability__unknown9	9.527930e-01	
BEMFAB__Marketability__X	1.364903e+01	
Public_Private_Indicator_unknown10	8.256458e-01	
Public_Private_Indicator_Y	5.060397e+00	
Owns_Rents_Code_X1	1.275764e+01	
Owns_Rents_Code_X2	2.006305e+01	
Owns_Rents_Code_ownsrentmiss	1.010515e+00	
Subsidiary_Indicator_X2	1.063539e+00	
Subsidiary_Indicator_X3	1.776722e+01	
Manufacturing_Indicator_X1	1.206079e+01	
Manufacturing_Indicator_X2	9.277709e-01	
Legal_Status_Code_X100	9.072897e-01	
Legal_Status_Code_X101	2.732624e+00	
Legal_Status_Code_X118	7.095238e-03	
Legal_Status_Code_X12	1.637133e+01	
Legal_Status_Code_X120	5.844285e+00	
Legal_Status_Code_X13	4.805020e+00	
Legal_Status_Code_X200	1.015080e+00	
Legal_Status_Code_X3	2.369222e+01	
Legal_Status_Code_X50	9.648597e-01	
Legal_Status_Code_X8	1.645055e+00	
Import_Export_Agent_Code_B	1.683211e+01	
Import_Export_Agent_Code_C	1.718773e+01	
Import_Export_Agent_Code_D	1.610708e+00	
Import_Export_Agent_Code_unknown13	3.034337e+00	
Import_Export_Agent_Code_G	2.036626e+01	
Import_Export_Agent_Code_H	1.567308e+01	
Status_Code_X1	1.290612e+01	
Status_Code_X2	7.754272e+00	
Status_Code_X3	8.511990e-01	
Population_Code_X10	1.050004e+00	
Population_Code_X2	7.025000e-01	

Population_Code_X3 4.114161e+00 Population_Code_X4 3.257827e+00 Population_Code_X5 1.022228e+01 Population_Code_X6 8.895020e+00 Population_Code_X7 1.316357e+01 Population_Code_X8 1.485978e+01 Population_Code_X9 1.883565e+01 Site_Status_unknown18 9.299794e-01 Revenue_Range_X.1.mil.to.less.than..5.mil 1.586367e+01 Revenue_Range_X.10.mil.to.less.than..25.mil 2.061090e+01 Revenue_Range_X.100.mil.to.less.than..250.mil 1.289256e+01 Revenue_Range_X.25.mil.to.less.than..50.mil 1.586799e+01 Revenue_Range_X.250.mil.to.less.than..500.mil 9.944837e+00 Revenue_Range_X.5.mil.to.less.than..10.mil 1.462903e+01 Revenue_Range_X.50.mil.to.less.than..100.mil 1.550329e+01 Revenue_Range_X.500.mil.to.less.than..1.bil 6.562587e+00 Revenue_Range_Less.than..1.mil 1.756626e+01 Revenue_Range_unknown19 9.651271e-01 Global_Ultimate_Indicator_unknown16 3.989304e+00 Global_Ultimate_Indicator_Y 1.493685e+01 Major_Industry_Category_Name_Construction 7.465030e+00 Major_Industry_Category_Name_Finance..Insurance..Real.Estate 2.315844e+01 Major_Industry_Category_Name_Manufacturing 1.887792e+01	

Major_Industry_Category_Name_Mining 5.523749e+00 Major_Industry_Category_Name_Public.Administration 7.321613e+00 Major_Industry_Category_Name_Retail.Trade 1.393460e+01 Major_Industry_Category_Name_Services 2.392090e+01 Major_Industry_Category_Name_Transportation...Public.Utilities 1.280798e+01 Major_Industry_Category_Name_unknown17 9.342820e-01 Major_Industry_Category_Name_Wholesale.Trade 1.849089e+01	

Variable importance list from the Lasso regression task	
	1
(Intercept)	7.617467249
X	.
Subsidiary_Indicator	-0.023141541
Manufacturing_Indicator	.
Legal_Status_Code	.
Status_Code	.
Population_Code	.
total_products	.
total_transactions	0.303480611

total_accounts	.
Year_Started	.
Square_Footage	.
Revenue__US_Dollars_	-0.045544818
Number_of_Family_Members	.
Employees_Here	-0.005353137
Employee_Count_Total	-0.005746445
Domestic_Ultimate_Employee_Count	-0.033530241
US_1987_SIC_1	0.019713635
age_of_company	4.775065023
Business_Code_CANADA	.
Business_Code_EMEA	.
Business_Code_Other	.
Business_Code_USA	.
Location_Type_HQ	-0.005212210
Location_Type_Single.Location	.
Location_Type_unknown8	.
BEMFAB__Marketability__D	.
BEMFAB__Marketability__M	.
BEMFAB__Marketability__O	-0.065994554
BEMFAB__Marketability__S	.
BEMFAB__Marketability__unknown9	.
BEMFAB__Marketability__X	.
Public_Private_Indicator_unknown10	.
Public_Private_Indicator_Y	.
Owns_Rents_Code_X1	0.019977239
Owns_Rents_Code_X2	.
Owns_Rents_Code_ownsrentmiss	.
Import_Export_Agent_Code_B	.
Import_Export_Agent_Code_C	.
Import_Export_Agent_Code_D	.
Import_Export_Agent_Code_F	.
Import_Export_Agent_Code_G	.
Import_Export_Agent_Code_H	.
Import_Export_Agent_Code_unknown13	.
Site_Status_unknown18	.
Revenue_Range_X.1.mil.to.less.than..5.mil	0.024860871
Revenue_Range_X.10.mil.to.less.than..25.mil	.
Revenue_Range_X.100.mil.to.less.than..250.mil	-0.005667086
Revenue_Range_X.25.mil.to.less.than..50.mil	.

Revenue_Range_X.250.mil.to.less.than..500.mil .
Revenue_Range_X.5.mil.to.less.than..10.mil .
Revenue_Range_X.50.mil.to.less.than..100.mil .
Revenue_Range_X.500.mil.to.less.than..1.bil -0.001273996
Revenue_Range_Less.than..1.mil .
Revenue_Range_unknown19 .
Global_Ultimate_Indicator_unknown16 .
Global_Ultimate_Indicator_Y .
Major_Industry_Category_Name_Construction .
Major_Industry_Category_Name_Finance..Insurance..Real.Estate 0.007070872
Major_Industry_Category_Name_Manufacturing .
Major_Industry_Category_Name_Mining .
Major_Industry_Category_Name_Public.Administration .
Major_Industry_Category_Name_Retail.Trade .
Major_Industry_Category_Name_Services -0.029586601
Major_Industry_Category_Name_Transportation...Public.Utilities .
Major_Industry_Category_Name_unknown17 .
Major_Industry_Category_Name_Wholesale.Trade .