Customer Churn Project
ISQA 8720 - Final Project

Dilanga Galapita Mudiyanselage
Ashwathy Ashokan
Anoop Mishra

Business Understanding

- ➔ 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.

Data Understanding

- ➔ 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.

United Communications dataset
26,827 rows x 13 columns(train)

Numeric:8

Categorial:5

United Communications dataset
8942 rows x 11 columns(test)

Numeric:7

Categorial:4

Firmographic dataset
55012 rows X 41 columns

Numeric:19

Categorial: 22

Train data

	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

Test data

	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

Firmographic dataset

	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

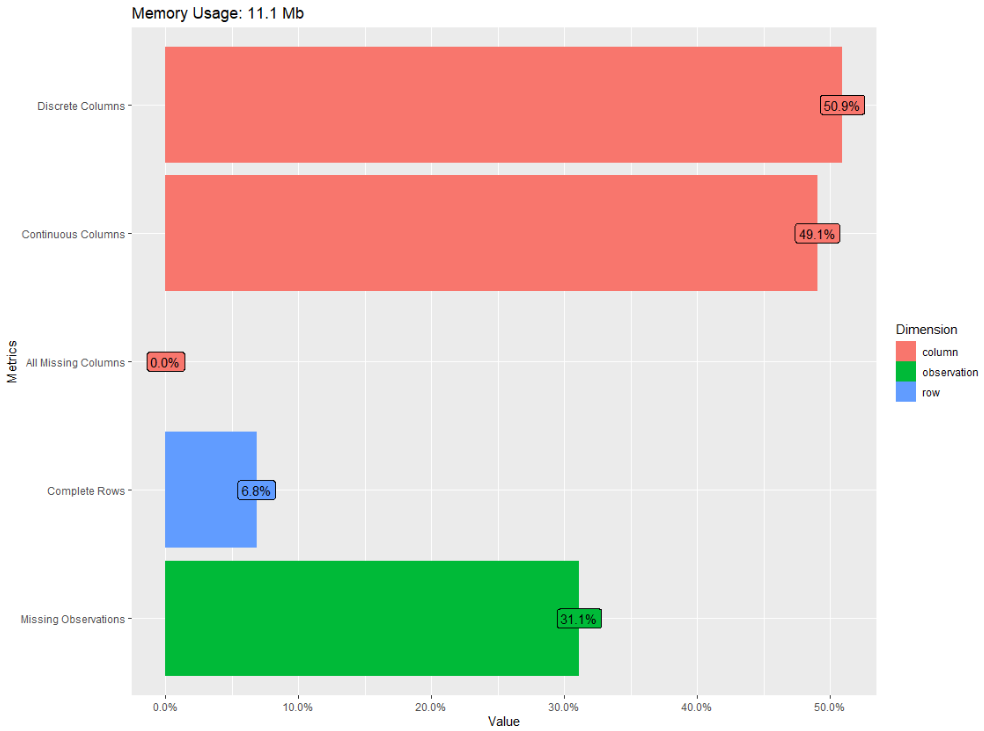
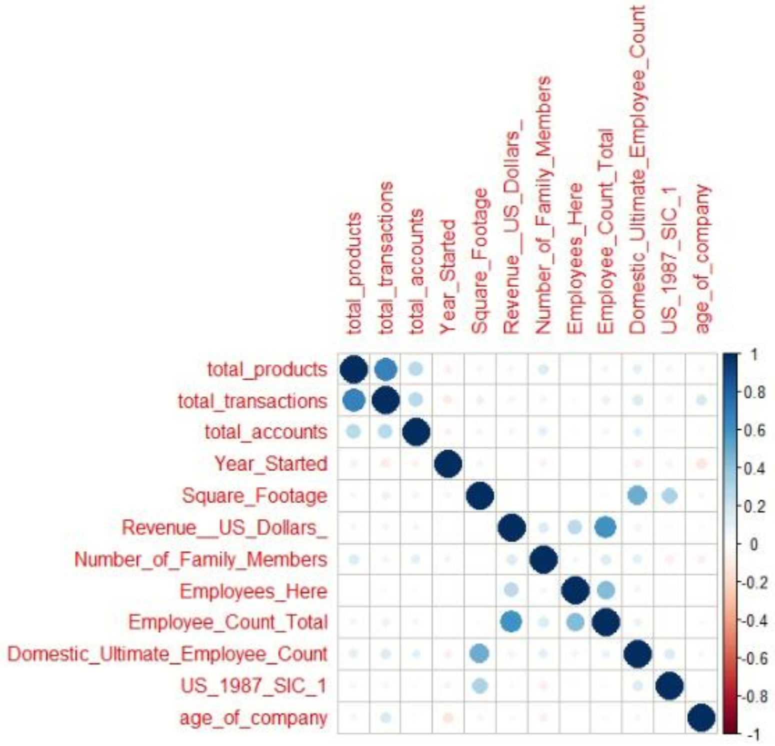
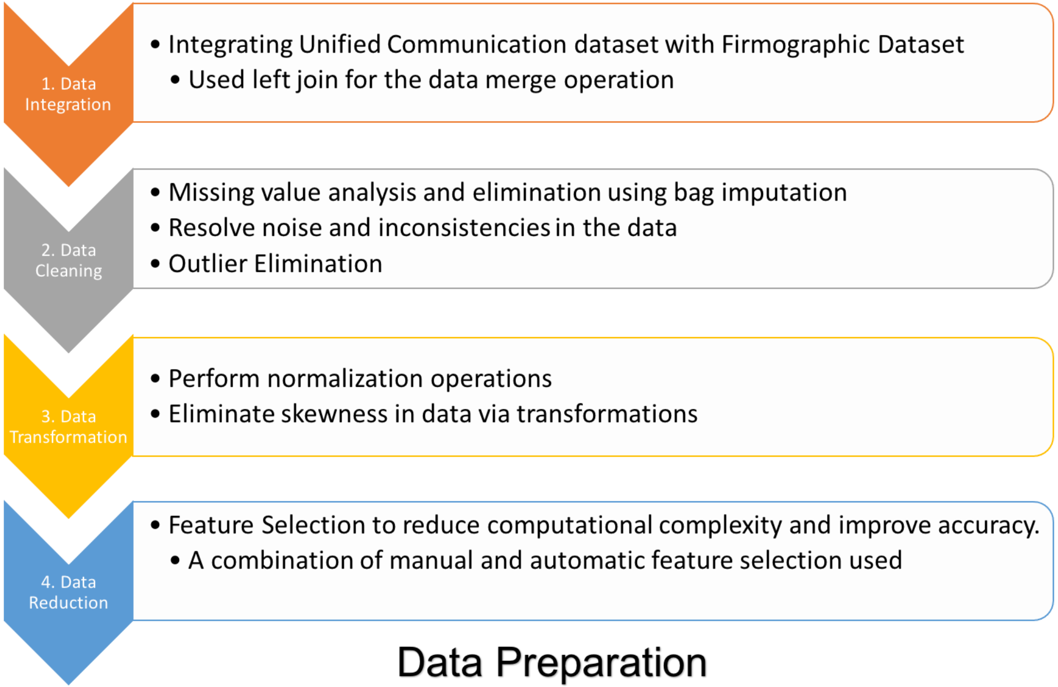


Figure above shows the various types of data from the train set



Correlation plot for numerical data from train set

Data Preparation



1. Data Integration

- Integrating Unified Communication dataset with Firmographic Dataset
  - Used left join for the data merge operation

2. Data Cleaning

- Missing value analysis and elimination using bag imputation
- Resolve noise and inconsistencies in the data
- Outlier Elimination

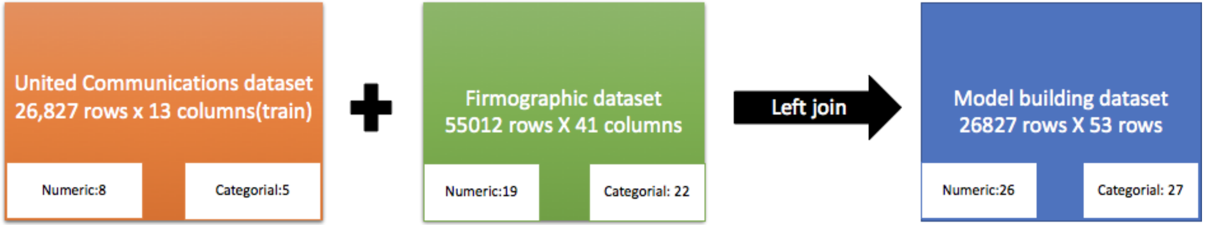
3. Data Transformation

- Perform normalization operations
- Eliminate skewness in data via transformations

4. Data Reduction

- Feature Selection to reduce computational complexity and improve accuracy.
  - A combination of manual and automatic feature selection used

Data Preparation



United Communications dataset
26,827 rows x 13 columns(train)

Numeric:8

Categorial:5

Firmographic dataset
55012 rows X 41 columns

Numeric:19

Categorial: 22

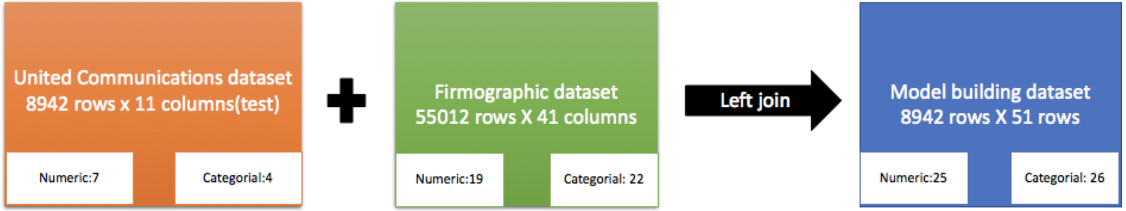
Left join

Model building dataset
26827 rows X 53 rows

Numeric:26

Categorial: 27

Data Integration – train data



United Communications dataset
8942 rows x 11 columns(test)

Numeric:7

Categorial:4

Firmographic dataset
55012 rows X 41 columns

Numeric:19

Categorial: 22

Left join

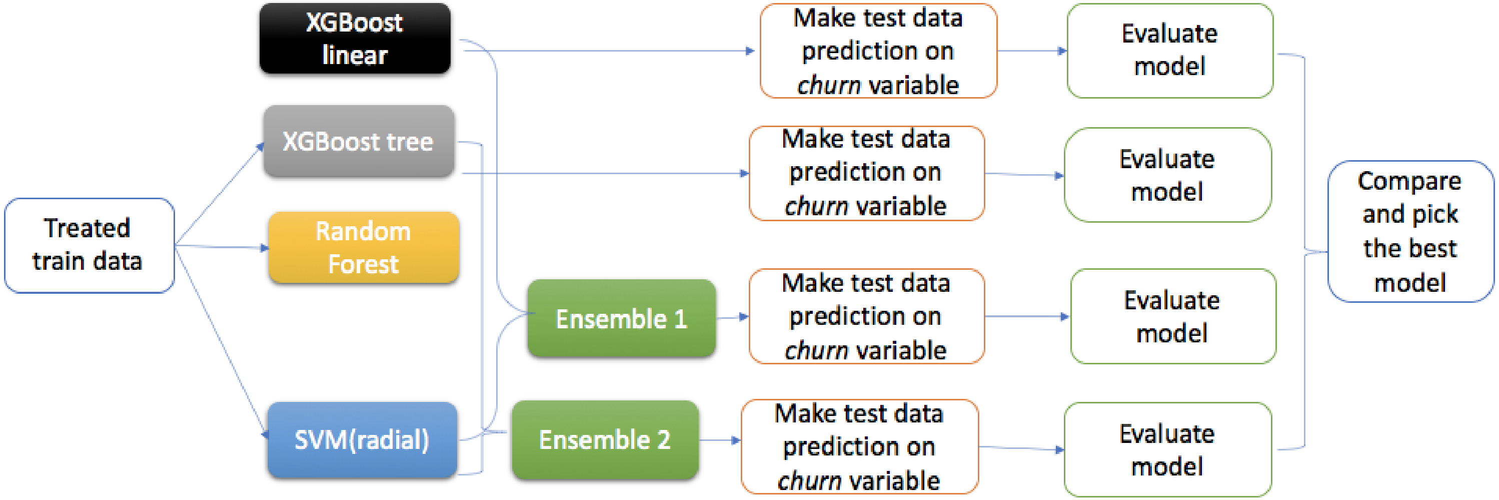
Model building dataset
8942 rows X 51 rows

Numeric:25

Categorial: 26

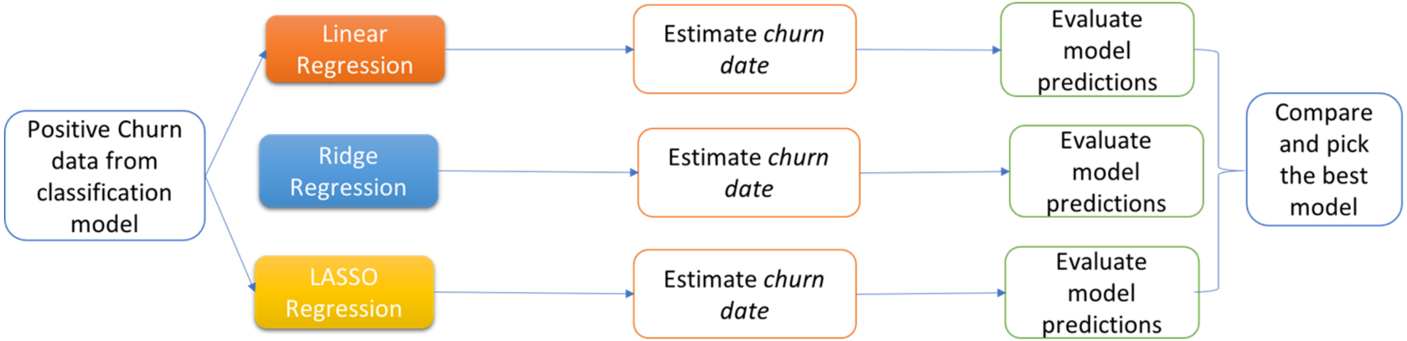
Data Integration – test data

Model Building


Treated train data
XGBoost linear
XGBoost tree
Random Forest
SVM(radial)
Ensemble 1
Ensemble 2
Make test data prediction on *churn* variable
Make test data prediction on *churn* variable
Make test data prediction on *churn* variable
Make test data prediction on *churn* variable
Evaluate model
Evaluate model
Evaluate model
Evaluate model
Compare and pick the best model

Classification model building

Model Building



Positive Churn data from classification model
Linear Regression
Ridge Regression
LASSO Regression
Estimate *churn date*
Estimate *churn date*
Estimate *churn date*
Evaluate model predictions
Evaluate model predictions
Evaluate model predictions
Compare and pick the best model

Regression model building

Evaluation - Classification

Summary of regression model performance measures on test sets of United Communications-Firmographic dataset

	Accuracy	Kappa	Sensitivity	Specificity	AUCs
XGB tree	0.8943	0.034	0.021	0.9985	0.688
Random Forest	0.8941	0.0612	0.04021	0.9960	0.671
SVM	0.8932	-4e-04	0.9998	0.00	0.550
Ensemble 1	0.8956	0.0647	0.8970	0.6764	Not Available
Ensemble 2(Yes)	0.8952	0.0614	0.8968	0.6471	0.710

Evaluation - Regression

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

Model Recommendation

- ➔ Due to the highly imbalanced nature of the data, accuracy cannot be used directly to choose the best model.
- ➔ AUC, sensitivity and kappa values will be used to pick the final model
- ➔ Churned: Ensemble learning-2 Classification   Churned date: Lasso Regression

Limitation and Challenges

- ➔ We did not consider most of the geographical informations, as they have high number of factor levels
- ➔ Computational resource limitations
- ➔ Most of the factor categories were merge together to one category to enhance the computational efficient

Conclusion

- ➔ Merged two datasets for the final prediction; high number of missing values; imputations techniques
- ➔ Computation was very costly; both automatic and manual feature selection are performed
- ➔ We eliminated variables that affect our final prediction results both in the classification and the regression models
- ➔ Churned: Ensemble learning-2 Classification   Churned date: Lasso Regression
- ➔ Future Work: TIme series Analysis