TELECOMMUNICATIONS CUSTOMER CHURN ANALYSIS

INSTRUCTOR：Dr. Ruoqing Zhu
GROUP MEMBERS: Xiaoying Yang, Bo Yang, Yvonne Ho

CONTENTS

- INTRODUCTION
- DATA WRANGLING
- UNSUPERVISED LEARNING
- SUPERVISED LEARNING
- CONCLUSION AND DISCUSSION

INTRODUCTION

INTRODUCTION

CUSTOMER CHURN

Customer churn is a percentage of how many customers stopped using your company's product or service during a certain time frame.

REFERENCE
https://www.google.com/url?sa=i&url=https%3A%2F%2Frawalrameshr2.medium.com%2Fchurn-prediction-f21a0e1c198f&psig=AOvVaw13M3k8g82KaMfA10cj8E0S&ust=1650839550841000&source=images&cd=vfe&ved=0CAkQjRxqFwoTCMDTleGeq_cCFQAAAAAdAAAAABAD

INTRODUCTION

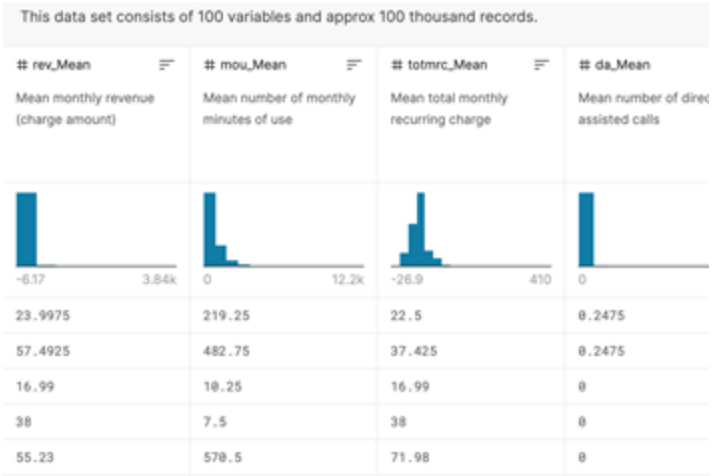
Ideas in Customer Churn Prediction

- Predicting churn is to create proactive marketing campaigns targeted at the customers.
- Forecasting customer churn with the help of machine learning is possible.
- Machine learning and data analysis are powerful ways to identify and predict churn.
- Churn is a one of the biggest problem in the telecom industry.
- Research has shown that the average monthly churn rate among the top 4 wireless carriers in the US is 1.9% - 2%.

DATASET

Telecommunication customer churn analysis

- This data set consists of 100 variables and approximates 100 thousand records.
- The variables explain the attributes of telecommunications industry and various important customer factors.
- Churn, the target variable, indicates whether the customer will keep using the products and services and can be predicted by associated variables.



TARGET

1. Identify if the churn variable (variable that characterizes the customers if they leave the telecom company or not) has any relationship with our analysis, and if we can classify the customers without using the churn variable and they correlate.
2. Select related variables and predict which characteristics of customers are most likely to lose and evaluate models.

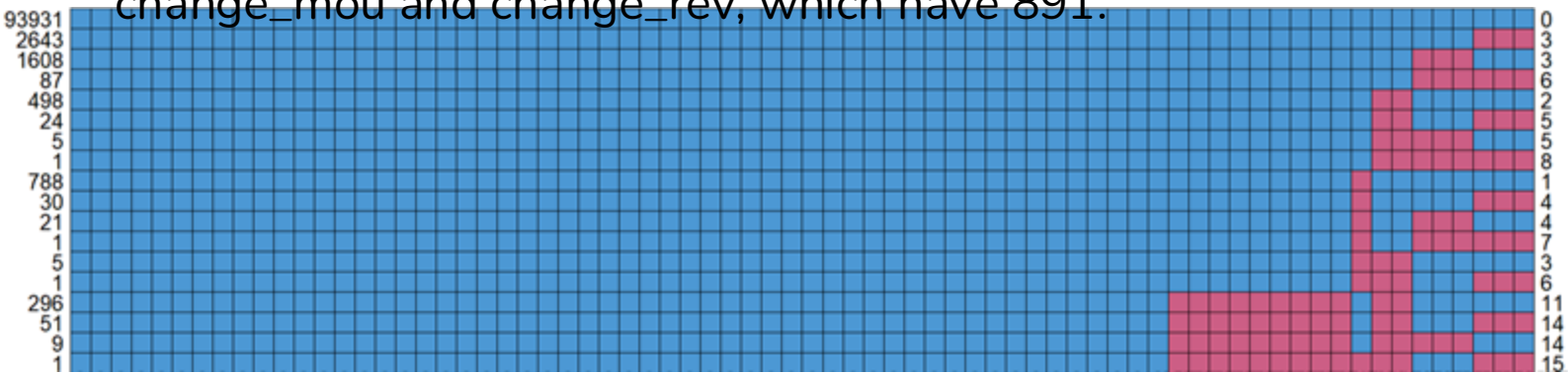
METHODOLOGY

Data Cleaning	Feature Engineering	Variable Selection	Clustering	Model Prediction
- Detected Missing Values - Used Random Forest to fill out the values	- Check Correlations - Check outliers and scaling	- Principal Component Analysis - Factor Analysis - Backward	- Normal K-Means - Mahalanobis distance K-Means	- Logistic Regression - Decision Tree - Random Forest - KNN

DATA WRANGLING

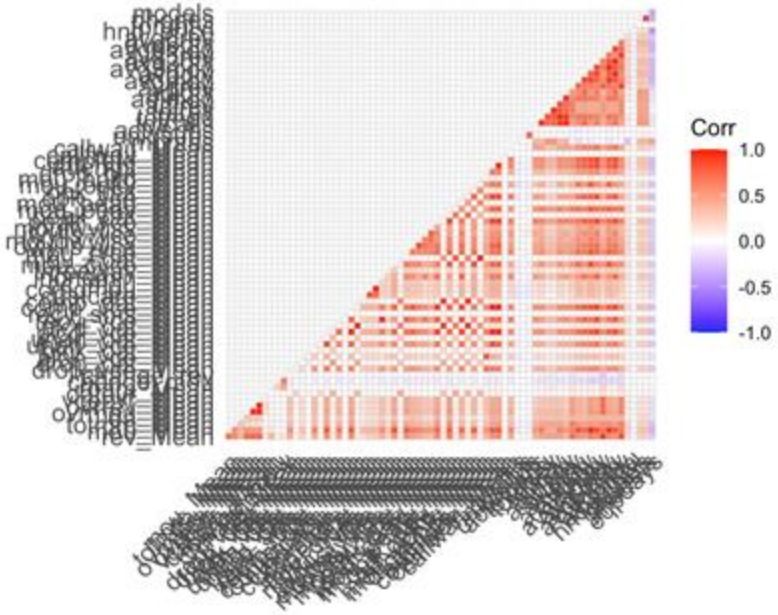
DATA WRANGLING

- A total of 33 variables have missing values.
- The variables with the most missing values are change_mou and change_rev, which have 891.



FEATURE ENGINEERING

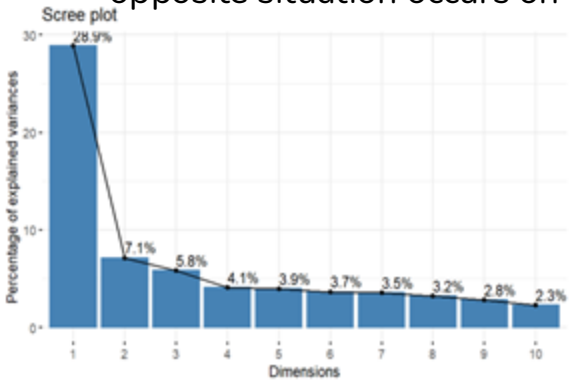
- Correlation between variables.
- PCA can be applied.

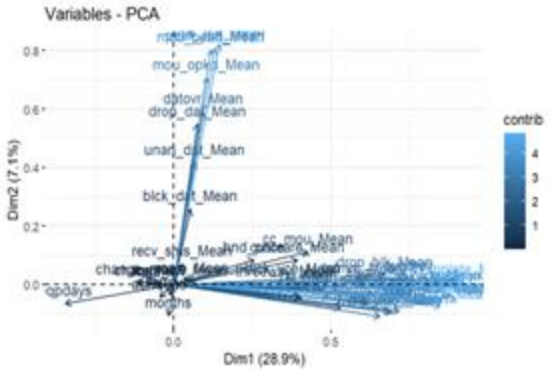


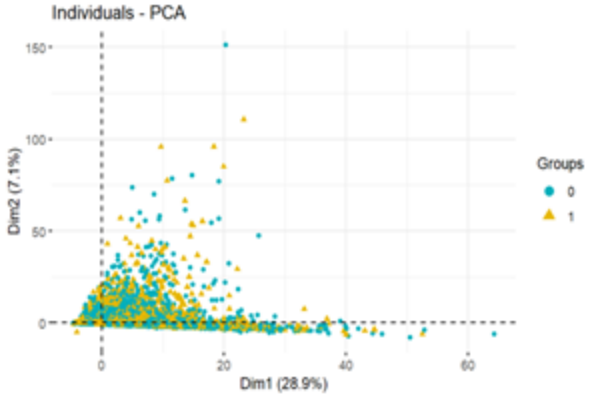
UNSUPERVISED LEARNING

PRINCIPAL COMPONENT ANALYSIS

- PCA analysis shows that the 1st principal component explains about 28% variance.
- The first PC is dominated by 21 variables, which the biggest contribution is from variable named complete_mean.
- The PC plot of 1st and 2nd dimension shows users who choose to retain cluster on the 1st PC axis, opposite situation occurs on the 2nd principal component axis.

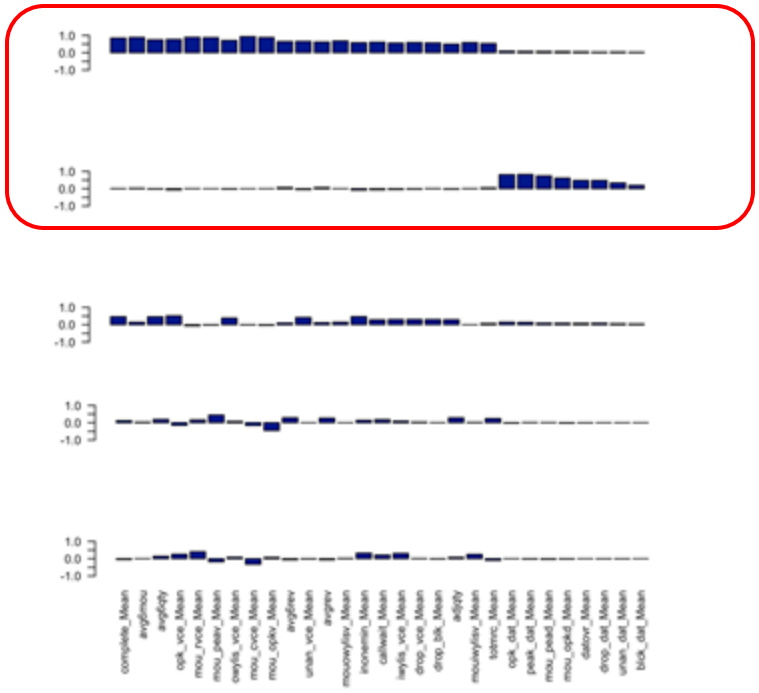






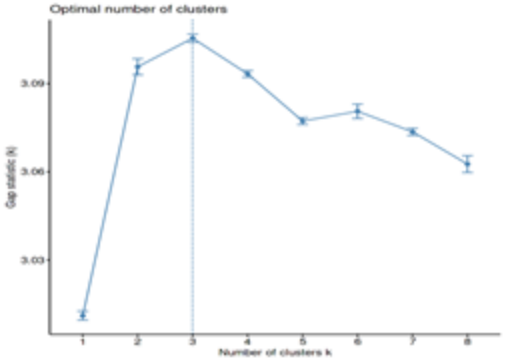
FACTOR ANALYSIS

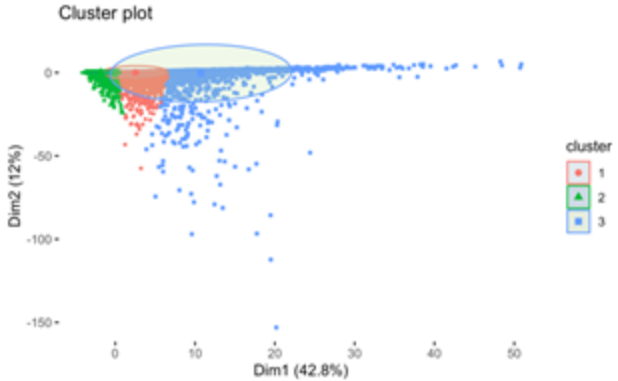
- Two of the factors show **inverse correlations**
- They differ on the exact variables and show that the characteristics of the two type of customers are opposite.
- First factor can be the customers that remain in the company, and the second factor the ones that leave the company



CLUSTERING

- The optimal number of clusters is **three**
- One group can correspond to customers that left the company and other to the ones that remain
- Cluster three are the most extreme customers and second the most common

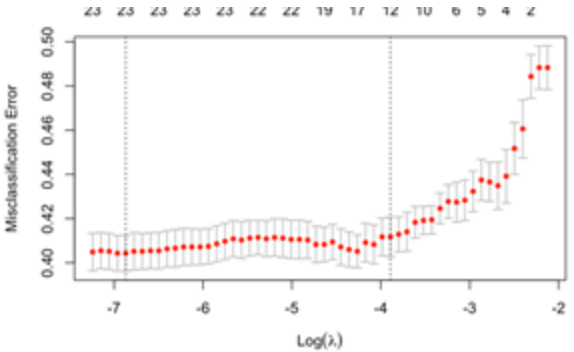


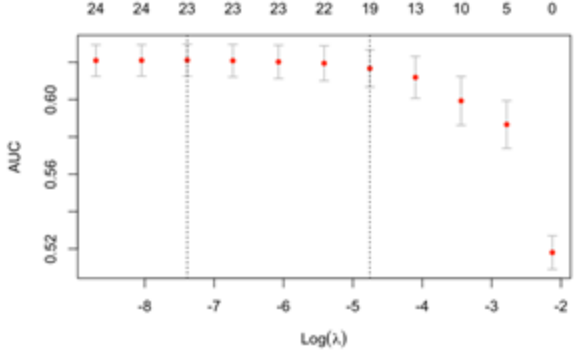


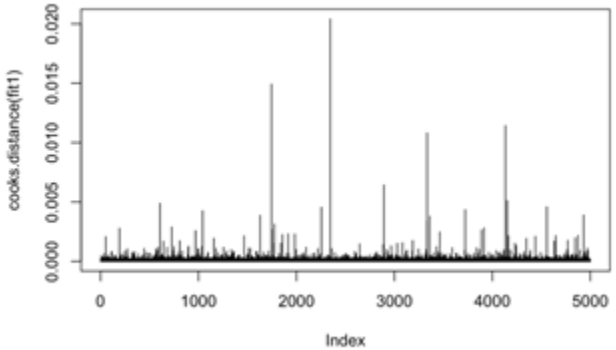
SUPERVISED LEARNING

LOGISTIC REGRESSION

- Backward elimination in additive model
- Residual Analysis: Leverages and outliers
- 10-fold cross validation to minimize the penalty
- Reconstructed a new logistic model with the smallest penalty for classification model evaluation

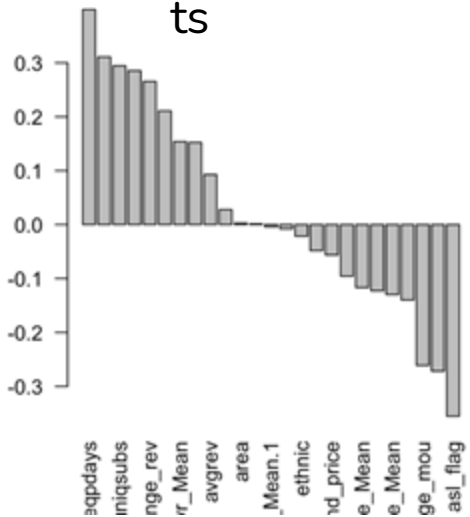




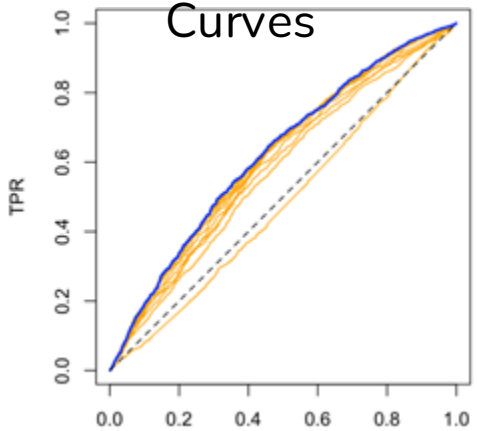


LOGISTIC REGRESSION

Coefficients

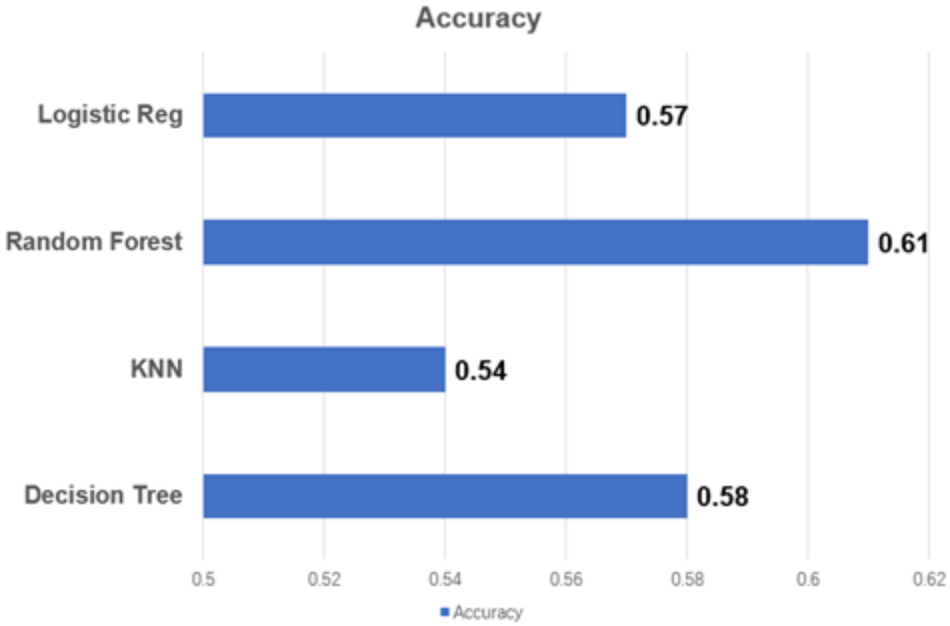


ROC Curves



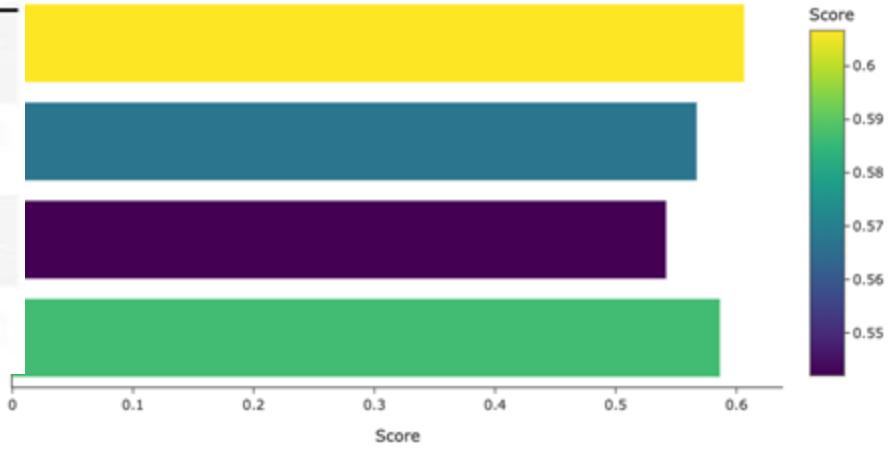
MODEL COMPARISONS IN R

- The random forest model performs the best, with 61 percent accuracy.
- The accuracy of each model is over 50 percent.
- Helps to predict whether a telecom company will retain a designated customer.



MODEL COMPARISONS WITH PYTHON

Model	Score
Random Forest Classifier	0.606667
Logistic Regression	0.586667
Decision Tree Classifier	0.567333
KNN	0.542000



CONCLUSION & DISCUSSION

CONCLUSION

- All variables are grouped into two clusters, and one can correspond to customers that left the company and other to the ones that remain

- The random forest model has the best performance, according to its highest accuracy rate. However, the logistic model has better interpretations on associations among the selected variables.

RECOMMENDATION

- Diagnostic analysis on leverages and outliers
- Transformations on numerical variables instead of simple scaling

THANK YOU

INSTRUCTOR：Dr. Ruoqing Zhu

GROUP MEMBERS: Xiaoying Yang, Bo Yang, Yin Tip Ho