Predicting Telcom Customer Retention Using Machine Learning  
Bruno Xie

**Data Set:** The data we will be using is The *Telco Customer Churn* data.

<https://www.kaggle.com/blastchar/telco-customer-churn>

This database is provided by the IBM business analytics community. It is based on a telco company that provides home phone and Internet services to customers in California. The data contains information of 7043 customers, and 21 features for each customer, which can be categorized in the following groups:

* customers who left or stayed within this month (churn);
* services that each customer has signed up for (phone service, multiple lines service, internet, online security, online backup, device protection, tech support, and streaming TV and movies)
* customer account information (how long they’ve been a customer, contract, payment method, paperless billing, monthly charges, and total charges)
* demographics information about customers (gender, senior, partners, dependents).

**Project Idea**: The marketing department of the telco company proposes to start a marketing campaign to reduce the churn rate and keep their customers. Another data scientist team will conduct cost-benefit analysis to target most profitable customers to campaign on based on our result. Our main task is to predict the customer's behavior of retention based on their features and help identify customers who are most likely to stay and who are not. We will be using the churn column - a dummy variable - as our predictive variable, and the rest of the variables will be used to help with the prediction. We will go through some of the most prevalent machine learning models, such as logistic regressions, k-NN, classification tree, and random forest, to fit the data and help us predict customer behaviors. Since our data contains up to 20 features, we will be using variable selection techniques such as intrinsic method, filter method and wrapper method to choose the most influential features. In the end, we will use methods including MSE, misclassification rate, ROC curve and profit curve, to pick the best model as the final result.