# Customer Churn Model

Customer churn rate is a key metric in the telecom industry as the cost of retaining an existing customer is far less than acquiring a new one. Customers may churn voluntarily (e.g., switching to a different provider) or involuntarily (e.g., relocation to a different country). Under most circumstances, companies focus on voluntary churns as they can be addressed and controlled more easily via relationship management strategies.

The goal of this project is to develop a churn model to predict customers’ propensity of risk to churn for a telecom company, and then generates a prioritized list of potential defectors to facilitate the marketing and sales team’s retention programs.

## Data

### Input

The project uses a sample telco customer data from Kaggle: <https://www.kaggle.com/blastchar/telco-customer-churn>

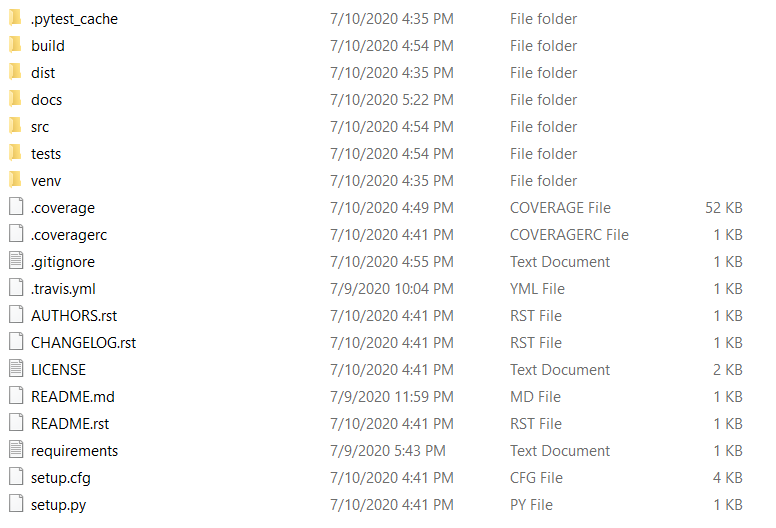
The raw data contains 7043 customer and 21 features:

* Customers who left within the last month – the column is called Churn
* Services that each customer has signed up for – phone, multiple lines, 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
* Demographic info about customers – gender, age range, and if they have partners and dependents

### Output

* Priority list of customers who have the highest likelihood to churn (Customer ID & propensity score)
* Model performance plots (ROC)

## Project structure



* doc: Sphinx documentation module
* src: core module
  + data: data extraction
  + features: feature engineering
  + models: train models
  + post: post-processing (e.g., KPI visualization, priority list)
* test: test functions for different modules

# ETL

After the raw data has been read in, the following procedures will be implemented:

* Check missing/invalid data
* Encoding for categorical/ordinal features

# Model

The following algorithms will be tested for the model and the best performed one will be used for prediction:

* Regularized logistic regression
* Random forest
* Gradient boosting tree
* Support vector machine