**Overview**

Customer acquisition and retention is a key concern for many industries, especially acute in the strongly competitive and quick growth telecommunications industry. Meanwhile, since the cost of retaining a good customer is much lower than acquiring a new one, it is very profit-effective to input valuable resource on the Retention Campaign. Customers churn for various different reasons. Experience tells us that once the customer has made up their mind, retention becomes really hard. Therefore, managing churn in a proactive process is essential for the growth of the customer base and profit. The primary goal of churn analysis is usually to create a list of contracts that are likely to be cancelled in the near future. The customers holding these contracts are then targeted with special incentives designed to deter cancellation. At a more sophisticated level, the telecommunications company attempts to detect the reasons for an expected cancellation, because this information may help customize the offer.

A diagram of data analysis

Description automatically generated

**Dataset**

The Data Source chosen here is an open-source data (cell2cell) by Teradata center for customer relationship management at Duke University. Cell2Cell dataset is pre-processed, and a balanced version provided for analyzing Process that consists of 71,047 instances and 58 attributes. We can try with data set, try Machine learning algorithms and deep leaning algorithms find measures like accuracy, ROC, AUC. Telecom Industry faces fierce competition in satisfying its customers. The role of churn prediction system is not only restricted to accurately predict churners but also to interpret customer churn behavior. Churn management in the telecom services industry is used to securing the customers for a company. In essence, proper customer management presumes an ability to forecast the customer decides to move to another service provider, a measurement of customer profitability, and a set of strategic and tactic retention measures to reduce the movement. In practice, we can segment its customers by profitability and focus retention management only on those profitable segments and score the entire customer base with propensity to churn and prioritize the retention effort based on profitability and churn propensity.

**Extract the Transform Load**

Before Modeling and Visualization, first of all we need follow the data engineering process, specifically ETL process, which stands for Extract, transform, and load variables from identified data items. Extract is the process of reading data from a database. In this stage, the data is collected, often from multiple and different types of sources. Transform is the process of converting the extracted data from its previous form into the form it needs to be in so that it can be placed into another database. Transformation occurs by using rules or lookup tables or by combining the data with other data. Load is the process of writing the data into the target database.

A diagram of gears and text

Description automatically generated

**Prediction model training and evaluation**

The first step of this module is data preprocessing. We'll start from VectorAssembler, which is useful for combining raw features and features generated by different feature transformers into a single feature vector, in order to train ML models like Random Forest and Gradient-Boosted Trees. Next, since we're going to use K-Means and Neural Network for the modeling, we're going to use StandardScaler transforms the dataset of Vector rows, normalizing each feature to have unit standard deviation and/or zero mean. Last, we'll use down-samplings method to deal with our unbalance dataset.

Next Step is modeling, in this project we'll use Random Forest Classifier and GBTs Classifier to predict customer Churn; Use XGBoost Regressor and Neural Network to predict Revenue Loss.

In the last part of each modeling iteration is evaluate the performance of each prediction model, in order to do so, we will split the dataset in 2 parts: training set (80% of the entries in the dataset) and test set (20% of the entries in the dataset). We train the model on the training set and then we test the model on the test set. I decided to use the ROC curve (A receiver operating characteristic curve) to test the model performance for classification problem which is to predict customer Churn in our case; And use mean absolute error for regression program which is predicted revenue loss in our case, then We'll plot a chart of how many of total revenue loss can be covered with our predicted highest risk group.

**Analysis and Visualization**

Based on the industry experience, usually there're five dimensions and listed below are the significant variables.

* Customer demography: Age, Tenure, Gender, Location, Zip code, etc.
* Bill and payment: Monthly fee, Billing amount, Count of overdue payment, payment method, Billing type, etc.
* Network/TV/Phone usage records: Network use frequency, network flow, usage time and period, internet average speed, In-net call duration, Call type, etc.
* Customer care/service: Service call number, service type, service duration, account change count
* Competitors information: Offer detail under similar plan, etc

**Customer Segmentation**

In this module, we'll use K-Means to build a Customer Segmentation model, and then we can compare the different distributions of each cluster.

Customer segmentation is the practice of dividing customers into groups of individuals that are similar in specific ways relevant to marketing, such as age, months in service, services usage and services spending. Based on the customer segmentation, the telecommunications company can customize their marketing efforts for different cluster, and gain a deeper understanding of their customers' preferences in order to more accurately tailor marketing materials.

**Results and Discussion**

From the result of the segmentation model, we can see that all customers have been separated into different clusters perfectly. Customer segmentation begins with gathering and analyzing data and ends with acting on the information gathered in a way that is appropriate and effective. So the next step in real business is based on these clusters and distributions, following with analytical technique such as Uplift Modeling to models the difference in behavior between target and control groups, so the business owner can customize their up-sale and cross-sale strategy.