- # Machine Learning Engineer Nanodegree

- ## **Capstone Proposal :** Customer Churn

Saeid Rostami
September 28st, 2018

- ## Domain Background

Machine learning is one of the hot-topics in both academia and industry. It is a field of engineering that uses statistical techniques to give computers the ability to learn and build analytical model from data. Although machine learning has been in academia for a long time, I have not used that since I started may Nanodegree program at Udacity. I am a master student at Electrical and Computer Engineering and I am about to finish my university. From the second year of my master I started thinking of my future career and based on my skills and my interest I choose to be a data analyst. Being a good data analyst was my primary reason that I decided to register at Udacity Machine Learning Engineer Nanodegree program. I believe that it is an essential part of every business in the world to collect and analyze data for their marketing purpose. It is where a data analyst or data scientist starts to shine. It is a data analyst job to translate and interpret mathematical data into a more understandable form. It is a data analyst job to predict a company's customer behavior based on the data that the company collected from the past. That is the main reason that I decided to do Telcom Customer Churn project.

Usually in business and marketing the term of customer churn refers to loss of customers, which can lead to loss of revenue. In the world of growing subscription base business it is extremely important to not loosing the customers. Customer churn is important since it can directly affect the amount of revenue of a company. In order to survive in this kind of business many companies are using data analyst techniques for understanding why some costumers are leaving the company and try to win them back. There is some research and papers on customer churn by different research group. For example, in [1] the authors show predictive modeling for customer churner. This paper demonstrates how to use decision tree for churning problem. In [2] the authors describes how to apply Naive Bayes algorithm with supervised learning for customer churn problem.

- ## Problem Statement

Customer churn refers to when clients decide to cut their relationships with a company. Many companies like telephone and Internet providers companies, Banks, insurance companies etc. usually use data analysts to predict and prevent losing customer. It is very important for every company to predict the risk of churning of a particular customer, especially when there is still time to prevent him/her to leave the company. Besides the direct loss of revenue because of churning, it is always more expensive to gain a new customer than it is to keep a current paying customer. It is worth to mention that there is a difference between voluntary churn and involuntary churn. Voluntary churn happens because of customers' own decision to leave the company and switch to another one, while involuntary churn occurs due to special circumstances like technical problem, customer relocating or death. Usually, involuntary churning is excluded from analysis and more focus on the factors that companies can control for avoiding churning.

The input data will be customers specifications and contract details such as, the customer is male or female, what kind of service he/she gets from the company, how he/she pays the bills, how often he/she pays the bill, is he/she senior citizen or not and so on. The output is a column of yes and no, which defines a customer keeps using the company services and pays or decides to leave the company. Customer churning is a classification problem since our output is a discrete type data. The output variable, Churn value, takes the binary form as "Yes" or "NO", it will be categorized under classification problem in the supervised machine learning.

- ## Datasets and Inputs

I will use Telcom Customer Churn dataset, which is available at http://www.kaggle.com. The data was downloaded from IBM Sample Data Sets https://www.ibm.com/communities/analytics/watson-analytics-blog/guide-to-sample-datasets/. The dataset provides 7043 customers information in 21 columns. We have both numerical and categorical type of information in this dataset. I am planning to use 25%-35% of data for testing purposes and 65% - 75% of data for training purposes. The dataset is not perfectly balanced since the proportion of the people that stayed and left the company is about 73/27. Some of the columns is listed as below;

CustomerID : Customer ID
Gender : Customer gender (female, male)
SeniorCitizen : Whether the customer is a senior citizen or not (1, 0)
Tenure : Number of months the customer has stayed with the company
PhoneService : Whether the customer has a phone service or not (Yes, No)
InternetService : Customer's internet service provider (DSL, Fiber optic, No)
TechSupport :Whether the customer has tech support or not (Yes, No, No internet service)
StreamingTV : Whether the customer has streaming TV or not (Yes, No, No internet service)

Contract : The contract term of the customer (Month-to-month, One year, Two year)
PaperlessBilling : Whether the customer has paperless billing or not (Yes, No)
PaymentMethod : The customer's payment method (Electronic check, Mailed check, Bank transfer (automatic), Credit card (automatic))
MonthlyCharges : The amount charged to the customer monthly

- **Benchmark**

The benchmark of this project is the output of decision tree model and I will try to beat the decision tree model.

- **Solution Statements**

Since this is labeled data set, I am planning to apply supervised technique on that. I want to use different supervised algorithms like, Logistic Regression, Decision Tree, S.V.M. and Random Forest and then compare the result together and see which one can better predict the customer churning. Since some columns of the data set have categorical data I will do One-Hot-Encoding technique to convert them to numerical data. For the next step I will check the data for any missing and null value. After cleaning data and checking the correlation between different features I will split them to train and test group and will apply the mentioned algorithms on them. For each algorithm I will calculate the accuracy, f1-score, recall and precision to compare with other algorithms.

- **Evaluation Metrics**

For the evaluation metrics it should be considered that the data is not perfectly balanced (73/27). Since we have a binary output, Yes and No, for churning, the first evaluation metric can be confusion matrix. Accuracy of that is average of the values lying across the main diagonal. Area Under Curve(AUC) is one of the most popular metrics when we have a binary classification problem. AUC has a range of [0, 1]. The greater the value, the better is the performance of our model. Finally I will use the f1-score, which is the Harmonic Mean between precision and recall. It tells you how precise your classifier is, as well as how robust it is.

- **Project Design**

For this project I will use Jupyter notebook with Python 3.6.4 and related libraries like Scikit-learn, Pandas, Numpy, Matplotlib, Seaborn etc. The most important part in the project was finding appropriate dataset, after a deep search on the Internet I found Telecom Customer Churn dataset from Kaggle.com, which is provided by IBM.

The first step will be exploring data to see what I have, number of columns, data type and so on. The second step will be data manipulation to bring data to more suitable shape and format. For example, there are some categorical type columns, which I need to convert them to numerical type data. I will use One-Hot-Encoding technique for converting categorical data to numerical data. Then I need to find missing values and if it is possible replace them with proper values, otherwise put null values there. Final step for this step is finding null and duplicate values and get rid of them.

After cleaning and preprocessing data, next step will be data visualization to see how the data distributed what kind of customers churn more and also find the correlations between different columns. For data visualization I will use Python Matplotib and Seaborn library. Also Seaborn Heatmap is a very proper way to see the correlation between columns. After finding the correlated data I will drop the columns that we do not necessarily need them.

The next step will be applying different supervised techniques on the data. For doing this I will use Python Scikit-learn library. At the first step I need to split the data into training and testing groups. The Churn column is my label column since it is in 'Yes', for the people stay with the company, and 'No', for the people leave company, format I need to convert them to 1 and 0 format for further calculations. Then I can apply Random Forest, Decision Tree, Logistic Regression and S.V.N. on the data and calculate accuracy, f1-score, recall and precision for each one of them.

At the final step I will compare the different algorithm, which I mentioned above to come up with the best model for my dataset.

- **References**

[1] K. B. Oseman, S.B.M. Shukor, N. A. Haris, F. Bakar, “Data Mining in Churn Analysis Model for Telecommunication Industry”, Journal of Statistical Modeling and Analytics, Vol. 1 No. 19-27, 2010.

[2] S.V. Nath, Customer Churn Analysis in the Wireless Industry: A Data Mining Approach, Technical Report, retrieved from http://download.oracle.com/owsf_2003/40332.pdf, April 14, 2014.