Machine Learning Engineer Nanodegree
Capstone Proposal

Mihir Rajput
May 5th, 2020

Churn Modeling (from Kaggle)

Domain Background

The churn rate is the percentage of subscribers to a service who discontinue their subscriptions to the service within a given time period. For a company to expand its clientele, its growth rate, as measured by the number of new customers, must exceed its churn rate.

The goal of Churn prediction is to detect customers is intended to leave a service provider or not. Retaining existing customer costs an organization from 5 to 10 times than gaining a new customer.

Predictive models can provide correct identification of possible churners in the near future in order to provide a retention solution.

So that organizations can predict the insights, loss, gains and plan their strategies. Also in addition this will help a large organization who delivers services to the customers to make important decisions based on predicted insights.

There was a research paper available from the past which was published to tackle these specific types of problems. I have attached its link in a reference section.

Problem Statement

When I was interning with a Multinational Telecom Company, there the Data Science Team was working on this problem. But their dataset was a huge one. Here I have provided one simple dataset.

So, here the Company wants a model which can predict that how likely its current customers will leave the company in near future and hence calculate its churn rate.
This is a Classification Problem in which you'll classify a customer based on his/her Credit Score, Region, Gender, Age, Tenure, Balance, Salary etc. whether he/she will EXIT(1) or NOT(0).

Datasets and Inputs

The datasets are provided by Kaggle on its official competition website. They are free to download.

This dataset contains customer's data like salary, gender, age, and credit score along with the customer's current status, existed or not. That's why this dataset is very nice match to solve churning problem.

The dataset consist of 10,000 rows, I will be using 80% of them to train the model and 20% to evaluate the model.

This dataset is imbalanced, because class 1(Exited) has 7500 rows and the class 0(Not Existed) has 2500 rows.

Although there are some null entries in the dataset by removing them I might be able to balance the dataset properly.

Input Data fields

- CustomerId – Id of the customer
- Surname – Surname of the customer
- CreditScore – Customer's credit score
- Geography – Geo-location of the customer
- Gender – Gender of the customer
- Age – Age of the customer
- Tenure – Tenure of the customer
- Balance – Customer's current balance
- NumOfProducts – Number of products purchased by the customer
- HasCrCard – 0 if customer has credit card 1 otherwise.
- IsActiveMember – 0 if customer is not active 1 otherwise.
- EstimatedSalary – Estimated salary of the customer.
- Exited – 1 if existed 0 otherwise.

Sample data

RowNumb	Customer	Surname	CreditSco	Geograph	Gender	Age	Tenure	Balance	NumOfPr	HasCrCard	IsActiveM	Estimated	Exited
1	15634602	Hargrave	619	France	Female	42	2	0	1	1	1	101348.9	1
2	15647311	Hill	608	Spain	Female	41	1	83807.86	1	0	1	112542.6	0
3	15619304	Onio	502	France	Female	42	8	159660.8	3	1	0	113931.6	1
4	15701354	Boni	699	France	Female	39	1	0	2	0	0	93826.63	0
5	15737888	Mitchell	850	Spain	Female	43	2	125510.8	1	1	1	79084.1	0

Solution Statement

The solution will be predictions of either customer will leave the company or will not. I will start with dataset exploration; I will try to understand the pattern. Then I will perform feature engineering to extract the useful information or features. For the training purpose I may start with state-of-the-art ML algorithms but for the final submission I definitely will leverage the Deep Learning ANNs (Artificial Neural Networks).

Benchmark Model

Since this is a binary classification problem, the benchmark model will be SVC based model. I will try to outperform this model's accuracy which is near 80%.

Evaluation Metrics

Predictions can be evaluated using the confusion matrix. Confusion matrix consist of the four major aspects, number of true positives, number of false positives, number of true negatives and number of false negatives. Using these four aspects we can calculate accuracy of the model very easily.

The confusion matrix would be generated on test set.

Project Design

I will start with dataset exploration part where I will closely analyze all the columns which are present into the dataset. This will give me important insights about the data. Then I will perform basic EDA (Exploratory Data Analysis) operations in order to derive more useful insights. This would help me to understand the dataset features.

Then I will perform feature engineering on the data, I will remove unnecessary columns and out-liars from the data. This is very important data pre-processing step. This will improve the model training efficiency significantly.

Then I will start training the model with ANN, I will try different hyper-parameters in order to get accurate model.

I would be spending more than 65% of the time on data cleaning and data pre-processing part and 35% of the time on training models and tweaking different hyper-parameters. The final accuracy will be calculated against the test data set generated before the training.

Reference

- Kaggle - https://www.kaggle.com/adammaus/predicting-churn-for-bank-customers
- ANN –

    - https://towardsdatascience.com/introduction-to-artificial-neural-networks-ann-1aea15775ef9?gi=27bc4a183c41
    - https://www.researchgate.net/publication/251626579_Introduction_to_Artificial_Neural_Network_ANN_Methods_What_They_Are_and_How_to_Use_Them
- Hyper Parameters –
    - https://towardsdatascience.com/what-are-hyperparameters-and-how-to-tune-the-hyperparameters-in-a-deep-neural-network-d0604917584a
- Previous research –
    - https://www.researchgate.net/publication/236625937_A_Proposed_Churn_Prediction_Model