Young Kyung Kim
Sep 12 2018

Machine Learning Engineer Nanodegree

Proposal

Domain Background

I am going to make a model to predict behavior of customers whether they would be retained or not. This data had been provided by IBM for teaching and data contains information about Telco customers. This kind of data and analysis is important for service provider in deciding marketing strategy to retain their customers. If they know which customers are not likely to be retained, then they could prevent this from happening by giving better offer to those customers. This kind of actions will prevent profit loss. I was personal motivated to investigate this problem because this dataset and problem seemed to be something I will be doing in company as a data scientist.

Problem Statement

One of the biggest concerns for many telephone service providers is retaining customers. To retain customers, the providers could provide retention program to current customers but it usually cost a lot of money. If the provider could know which customers will likely to leave, then it could save its cost on retention program by providing the program to them. In another word, identifying the possible leaving customers will increase efficiency for retention program.

My hypothesis is that there will be a model that could predict whether a customer will likely to leave or not.

Y = f(x)
Y = percent of customer leaving.
f(x) = prediction model

For example, assume percent of customer leaving could be explained by multiple linear regression or decision tree.

Datasets and Inputs

For this project, we will use dataset provided by IBM and Telco, which could accessed in the Kaggle(https://www.kaggle.com/blastchar/telco-customer-churn/home). The dataset contains information on whether customer leaving the service, what kind of services that customer has sign up for, customer's account information and customer's demographic information. For what kind of services that customer has sign up for, the data contains information whether they have signed up for phone or internet services. These kinds of information could be important. For example, customer who signed up for phone could have higher chance of leaving compared to customer who signed up for internet. Such information will be helpful for predicting whether customers will leave or not.

Similar to what kind of services that customer has sign up for, customer's account information could be important. Whether contract is monthly or yearly based could influence whether customers will leave or not. Also, customer's demographic whether they are married or not could influence a lot in predicting whether customers will leave or not. Even though it is not clear which kind of data would relevant in this project, it is our job to find out through exploration and testing.

Solution Statement

My hypothesis is that there will be a model that could predict whether a customer will likely to leave or not.

Y = f(x)
Y = percent of customer leaving.

For example, assume percent of customer leaving could be explained by multiple linear regression, then f(x) will be like

$f(x) = a_1X_1 + a_2X_2 \ldots.. + a_nX_n + e$

X= input variables
n= number of input variable
e = residual or error

I am also going to try out different methods as random forest or multiple linear regression. Among those model I am going to compare their prediction score and choose the best one as the solution.

Benchmark Model

Since this data has been posted on Kaggle, there are a lot of model posted online. One of the posted model used various methods to solve the problem. For example, it used regression and random forest. It achieved about 70% of accuracy score. The equation for accuracy score is

Acc = (True Positive + True Negative)/ (True Positive + False Positive + False Negative + True Negative)

True Positive = Whether the model had correctly predicted positive value
False Positive = When the model had predicted negatively when actual value is positive
True Negative = Whether the model had correctly predicted negative value
False Negative = Whether the model had predicted positively when actual value is negative

Evaluation Metrics

In this section, propose at least one evaluation metric that can be used to quantify the performance of both the benchmark model and the solution model. The evaluation metric(s) you propose should be appropriate given the context of the data, the problem statement, and the intended solution. Describe how the evaluation metric(s) are derived and provide an example of their mathematical representations (if applicable). Complex evaluation metrics should be clearly defined and quantifiable (can be expressed in mathematical or logical terms).

I will use accuracy score, which is explained in above, as an evaluation metrics.

The equation for accuracy score is

Acc = (True Positive + True Negative)/ (True Positive + False Positive + False Negative + True Negative)

True Positive = Whether the model had correctly predicted positive value
False Positive = When the model had predicted negatively when actual value is positive
True Negative = Whether the model had correctly predicted negative value
False Negative = Whether the model had predicted positively when actual value is negative

I am also going to create one more scoring metrics. Above accuracy score do not differentiate a model that predicts given customer will 51% leave and another model that predicts given customer will 99% leave as it both models will predict given customer will leave. Below scoring metrics will evaluate how much the model had accurately predicted in percentage.

Score = $-\sum_i^n(|\text{actual value } i - \text{ expected value } i|)$ /n

Actual value = whether customer left or not. Left is 1 and not left is 0
Expected value = percentage of whether customer will leave or not
n = number of customers

This score indicates that the score closer to the 0 means the model had accurately predicted in percentage.

Project Design

First I will explore the data, especially to check correlation among input variables. If there are correlation among input variables then I will measure which variable is more influential on predicting whether customer will leave. I will use variable with more influence as a input variable. After choosing input variables, I will preprocess the data for different models that I am going to implement. First I will try multiple linear regression and access its score in prediction. Then I will try random forest to build model. I will compare the result and choose the best one.