Machine Learning Engineer Nanodegree

Capstone Proposal

P.MANOJ RAJESH

23-06-2018

Domain Background

The domain of my proposal data is related medical insurance and I choose this because if an insurance company wants to make money, it should collect the more money in year premiums than that it spends on medical treatment for a beneficiary so, insurers should invest more time to predict the annual cost of an individual and charge a little more than that , as it is a very time taking process to predict each individual cost manually, and as time is the most precious factor in today's life, I am trying to develop a model that predicts estimated cost of an individual based on previous data. And this problem is to be solved because every day number people who are applying for health are increasing, insurance providing companies need a model that makes one of their task easier.

https://www.techemergence.com/machine-learning-at-insurance-companies/

https://cloud.google.com/blog/big-data/2017/03/using-machine-learning-for-insurance-pricing-optimization

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5134202/

Problem Statement

The problem statement is to predict the future medical expenses of individuals that help medical insurance to make decision on charging the premium and the medical expenses are difficult to estimate because many diseases are rare and random, but we can still predict the estimated that an individual needs in a annual year by observing some of the factors and help insurance company to

make decision on charging the premium. And this can be done by applying past data as training data as past data is the data and it contains the exact cost of an Individual that an insurance company spend on them so, by looking and analysing the features (smoker or not, body mass index, region) of individual of certain cost we can easily predict the future cost of an individual who is nearly with same features.

Datasets and Inputs

Columns in the data set are

Age : age of primary beneficiary

- sex: insurance contractor gender, female, male
- bmi: Body mass index, providing an understanding of body, weights that are relatively high or low relative to height, objective index of body weight (kg / m ^ 2) using the ratio of height to weight, ideally 18.5 to 24.9
- children: Number of children covered by health insurance / Number of dependents
- smoker: Smoking
- Region: the beneficiary's residential area in the US, northeast, southeast, southwest, northwest.
- charges: Individual medical costs billed by health insurance

In these data set inputs are sex, age, body mass index, number of children , smoker or not, region of the applying person.
Output is the individual medical cost.
In this data set I have 7 Rows and 1340 Columns.
I would like to split the data into training and testing set and evaluate my model in initial stage.
Link for dataset: https://www.kaggle.com/mirichoi0218/insurance

Solution Statement

The solution to this problem can be obtained by looking at the data set, even tough dataset is medium sized we can train our model very well as the data is very accurate and do good predictions and help the insurance policy companies and save their time and effort. As the dataset contains features like sex, age,

body mass index, number of children, smoker or not, region of the applying person we can estimate the how much cost is taken by a person who smokes and how much cost is taken by a person who has more than 1 children in this way we can get some relations between features and the cost so, by this we can predict the future cost of a new customer.

Benchmark Model

Initially I used decision tree regressor to fit the training data and then I will check the predicted the cost by dropping the cost axis of the testing then I will compare predicted values with the actual cost and I will calculate score to check the model working, first of all I need to convert and descriptive data whether a person is smoker or not values like 0 or 1(if smoker assign 1or assign 0) and similarly I need to assign numerical values for the region also as machine learning algorithms work on numerical data. So, by doing this we train our model and fit the data to decision tree regressor algorithm and then check the predicted values with the actual values.

Evaluation Metrics

Initially I used decision tree regressor to fit the training data and then I will check the predicted the cost by dropping the cost axis of the testing then I will compare predicted values with the actual cost and I will calculate score to check the model working, first of all I need to convert and descriptive data whether a person is smoker or not values like 0 or 1(if smoker assign 1or assign 0) and similarly I need to assign numerical values for the region also as machine learning algorithms work on numerical data. So, by doing this we train our model and fit the data to decision tree regressor algorithm and then check the predicted values with the actual values. So, now to increase the score of my I

will use algorithms like Gradient Boosting algorithms and SVC so then working performance of my model will increase and then it will predict cost more correctly. I would like to use score and RMSE(if needed) to check the performance of the model.

```
0<=score(truth, predictions)<=1
```

If score then our model is good.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

Comparison

Project Design

(approx. 1 page)

First of all I need to convert and descriptive data whether a person is smoker or not values like 0 or 1(if smoker assign 1or assign 0) and similarly I need to assign numerical values for the region also as machine learning algorithms work on numerical data and then I will fit the data to decision tree regressor and then I will calculate score and if my score is not good then I will use algorithms svc and gradient boosting algorithms so, by doing this we can improve the score and working performance of the model. And the estimation on the medical expenses of individual based on their personal/family traits is of interest in this exercise

and It is always helpful to have some idea of what the target feature is. Thus I will plot a histogram of the medical expenses. Having this in mind is helpful for me to do the further analysis.