# Insurance Cost Prediction of Individuals based on their Health Information using Machine Learning Methods

Hasan Kısa
*Computer Engineering*
*İstanbul Technical University*
İstanbul, Turkey
kisah16itu@gmail.com

Cedan Murat Zeynepli
*Computer Engineering*
*İstanbul Technical University*
İstanbul, Turkey
zeynepli19@itu.edu.tr

Ertuğ Erdoğan
*Computer Engineering*
*İstanbul Technical University*
İstanbul, Turkey
erdoganer19@itu.edu.tr

Kerim Genç
*Computer Engineering*
*İstanbul Technical University*
İstanbul, Turkey
genck18@itu.edu.tr

***Abstract*—This project aims to predict health insurance cost using health parameters of individuals.**

***Index Terms*—machine learning, cross validation, cost prediction**

## I. Introduction

In this project, our team, 190714_190804_180720_160021, implemented a function which used five-fold cross validation to predict a person's insurance charge, using this person's health data which consists of age, sex, BMI (body mass index), number of children, smoking status and region. Some of these data used directly and some of them used combined as another variable for inclusion enhancement purposes. Our team's final score is 20491183.03880 and rank is 6 as now.

## II. Datasets

### *A. Smoking Status*

Smoking status is 1 if person is smoker, 0 otherwise.

### *B. Sex*

Sex is 1 if person is female, 0 otherwise. Sex is not used as this raw in the function.

### *C. BMI*

BMI is 1 if BMI of person is above 29.72, 0 otherwise. Ideal range of the BMI is 18.5 to 24.9. However based on our data, our team decided to use whether the person is obese which means BMI is above 29.72, rather than the BMI of the person is in the ideal range or not. BMI is not used as this raw in the function.

### *D. Region*

In data, region can be either northeast, southeast, northwest, southwest which we can map into for different values. Yet our team decided to map into two values which gave less error value. Region is 1 if it is northeast or southeast, 0 otherwise.

### *E. Number of Children*

For number of children, our team decided to split data into 5 values. If person has no child, map value is 7. Otherwise, if person has $x$ children, map value is $max(6-x, 1)$.

### *F. Derivated Data*

We used some combined data which are age and BMI combined, smoking status and BMI combined, and smoking status, age and BMI combined. We will explain how we got these derivated data using raw data.

## III. Methods

In this section, first we will explain how did we derivate our processed data, data selection, and then our prediction function.

### *A. Derivation of Raw Data*

*1) Age & BMI:* We designed an algorithm to combine age and BMI data. Cases in this algorithm are depend on whether person's age is below 28, above 50, and in between these, and also person's obesity status which is based on BMI. We can clearly show our 6 cases as below:

$$(age < 28, obesity = 1), (age < 28, obesity = 0),$$
$$(28 \leq age \leq 50, obesity = 1), (28 \leq age \leq 50, obesity = 0),$$
$$(age > 50, obesity = 1), (age > 50, obesity = 0)$$

For each category, we calculated the mean of insurance cost, and used it as our mapping value.

*2) Smoking Status & BMI:* For this combination, we designed four cases depends on person's smoking status and obesity status. Since each data is either true or false, we easily designed four cases. Same as former combination, we used mean of insurance cost for our mapping value.

*3) Smoking Status & Age & BMI:* To derive values from this combination, we basicly used cartesian product of above combination categories (smoking status & BMI) and age categories of the first derivated data ($age < 28$ or $28 \leq age \leq 50$ or $age > 50$).

### *B. Data Selection*

Our team decided not to use sex and BMI data as raw since their absolute correlations (0.0541, 0.2061, respectively) are lower than absolute correlations of our derivated data (0.3327, 0.8712, 0.8920, respective to the prior subsection).

### C. Gradient Descent Algorithm Using 5-Fold Cross Validation

In our function, we used gradient descent algorithm using 5-fold cross validation. We used KFold from sklearn, with 5 splits, 1 as random seed and enabled shuffle. We decided learning rate as 0.0003, and maximum number of iterations as 300000. We used common gradient descent algorithm $((X^TX)^{-1}X^TX)$ for each fold, and took the mean square error for each fold after calculated the beta vector. Beta vector of the minimum of the mean square errors is our finalized beta vector. We used this beta vector to calculate our predictions. Our function saves the predictions as .csv file and printed the mean square error for the corresponding input.

## IV. Results and Conclusions

Our beta vector from our five-fold cross validation gradient descent function is

[
$2.03634144 \cdot 10^2$,
$-4.83471390 \cdot 10^2$,
$1.41798223 \cdot 10^3$,
$1.01591098 \cdot 10^2$,
$2.47487716 \cdot 10^{-1}$,
$6.96867091 \cdot 10^{-1}$,
$-5.56433113 \cdot 10^{-2}$,
$-4.62000400 \cdot 10$
]

for corresponding data

[
age,
number of children,
smoking status,
region,
derivation of smoking status & age & BMI,
derivation of smoking status & BMI,
derivation of age & BMI,
$beta_0$
]
,respectively.

In Kaggle competition, our team got $6^{th}$ place with 20491183.03880 score.

In Fig.1, our reasons for selecting and processing data can be seen visually.

Also, in Fig.2, our derivated data plots can be seen, derivation of smoking status & age & BMI, derivation of smoking status & BMI, derivation of age & BMI respectively.

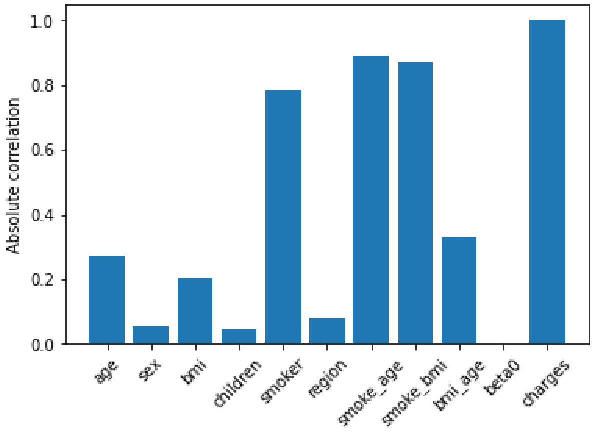

Fig. 1. Absolute Correlation Chart.

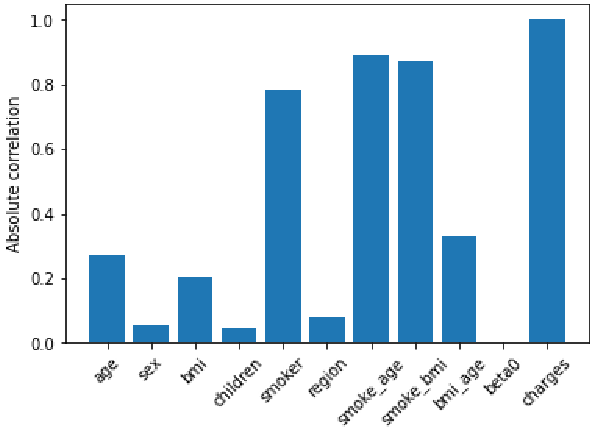

Fig. 2. Plots for Derivated Data