Machine Learning Engineer Nanodegree

Capstone Project
Subhakar K S
May 31st, 2018

I. Definition

Machine learning is a method of data analysis that automates analytical model building. It is a branch of artificial intelligence based on the idea that systems can learn from data, identify patterns and make decisions. Machine learning is being used in various domains like Financial services, Government, Health Care, Marketing and sales, Oil and gas, Transportation, etc.

Machine learning techniques can broadly be divided into 3 types

a. Supervised learning
b. Unsupervised learning
c. Reinforcement learning

Project Overview

Supervised learning is the machine learning task of learning a function that maps an input to an output based on example input-output pairs. It infers a function from labeled training data consisting of a set of training examples. Supervised learning is where we have input variables (x) and an output variable (Y) and we use an algorithm to learn the mapping function from the input to the output. Supervised learning is the process of an algorithm learning from the training dataset. Supervised learning problems can be further grouped into

a. Regression problems and
b. Classification problems

In classification, learning algorithms takes the input data and map the output to a discrete output like True or False
In regression, learning algorithms maps the input data to continuous output like weight, cost, etc.

In this project I will apply regression techniques of supervised learning to predict the medical insurance costs. Kaggle is a platform for machine learning and data science. Kaggle provides a number of open data sets catering to various domains.

This project uses an open data set for Medical Cost Personal Datasets from Kaggle. I used below languages, libraries and tools to implement this project.

a. Python 2.7
b. Jupyter
c. Pandas
d. NumPy
e. Seaborn
f. Matplotlib
g. Scikit learn

First the data is analyzed to understand the correlation between various fields from the input data. Input data is explored using various graphs and plots. Data is pre-processed and prepared for applying machine learning algorithms. Several machine learning models are evaluated using the regression related metrics. Selected model is fitted and fine-tuned for obtaining a best estimator for the data and model is applied to make predictions.

Problem Statement

Goal of this project is to apply regression learning techniques to the input data and predict the medical insurance costs. As mentioned in the previous section, this project uses the data from Medical Cost Personal Datasets. As seen in the project, input data consists of several fields described below.
Dataset consists of 1338 records. Each record contains the below data for specific person.

a. age – Age of the person
b. sex – Sex of the person
c. bmi – Body Mass Index(BMI) of the person
d. children – Number of children for the person
e. smoker – Smoking status of the person
f. region – Region of the person in US
g. charges – Medical Insurance costs per year for the person

Data was split into input features and output target. The fields age, sex, bmi, children, smoker, region are treated as input features and charges is treated as target that needs to be predicted.
Further, the data is split into training and testing data to apply regression learning techniques. Selected model is used to learn from the training data and predictions are made on the testing data.
Input features are explored using visualizations.

Metrics

Metrics are very important in evaluating the model. Regression calculates an equation that minimizes the distance between the fitted line and all of the data points. Regression model focuses on the relationship between a dependent variable and a set of independent variables. The dependent variable is the outcome, which we are trying to predict, using one or more independent variables. A model fits the data well if the differences between the observed values and the model's predicted values are small and unbiased. There are several metrics that can be used to gauge the performance of a regression model. I opted for below two metrics.

a. **Coefficient of determination**
   This metric is denoted by R^2 and also called are "R squared". R-squared is a statistical measure of how close the data are to the fitted regression line. This metric calculates the proportion of the variance in the dependent variable that is predictable from the independent variable(s).
   Best possible score is 1.0 and it can be negative (because the model can be arbitrarily worse). A constant model that always predicts the expected value of y, disregarding the input features, would get a R^2 score of 0.0.

b. **Explained variance score**
   While determining the dependent variable, first we need to understand how much variance is observed in it. This metric explains the fluctuation (or variance) of the dependent variable by using the independent variables. This measures how far a set of numbers are spread out from their mean value. Best possible score is 1.0, lower values are worse.

II. Analysis

Data Exploration

Given insurance data consists of 1338 records in it raw form as shown in the below output from Pandas Dataframe.

```
Int64Index: 1338 entries, 0 to 1337
Data columns (total 7 columns):
age         1338 non-null int64
sex         1338 non-null object
bmi         1338 non-null float64
children    1338 non-null int64
smoker      1338 non-null object
region      1338 non-null object
charges     1338 non-null float64
```

Further analysis of the target variable charges shows the below statistics.

```
Statistics for Medical Insurance dataset:

Minimum insurance cost: $1,121.87
Maximum insurance acost: $63,770.43
Mean insurance cost: $13,270.42
Median insurance cost $9,382.03
Standard deviation of insurance costs: $12,105.48
```

To understand correlation between the input or the independent variables, I classified the ages and BMI into specific categories. The below analysis shows the stats like number of male and female, smokers and non-smokers, etc. from the input data. Taking the BMI ranges from Medline Plus, I categorized the input samples into `below-weight`, `normal-weight`, `over-weight` and `obese`.

From the below data we can observe below

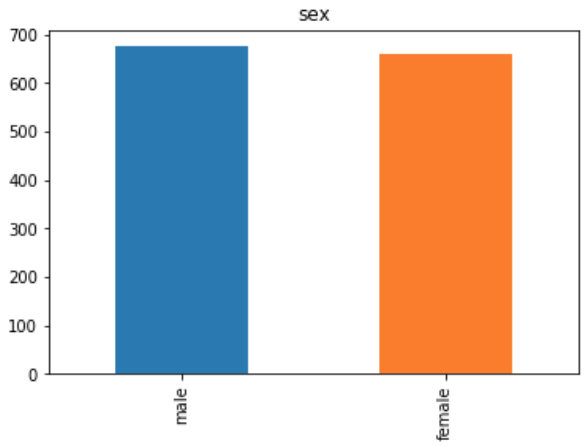
a. Input data contains an even distribution of male and female samples
b. Majority of them are non-smokers with 1064 samples
c. Majority of the samples are in age groups 20-29 and 40- 49 with the numbers 280 and 279 respectively.
d. A major sample of input data contains persons with no children with 574.
e. The data is evenly distributed across 4 regions with the region of `southeast` having slightly more samples.
f. Majority the sample fall under the category of `obese`

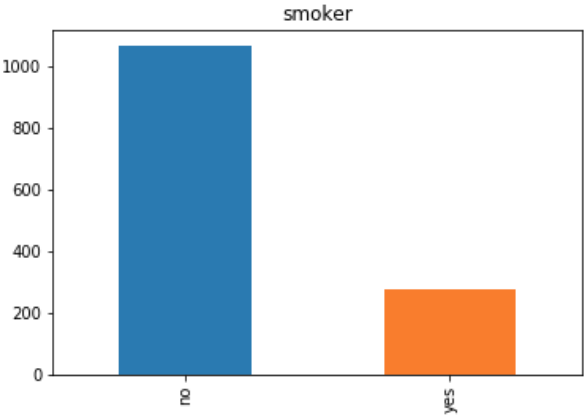
```
cage
10-19    137
20-29    280
30-39    257
40-49    279
50-59    271
60-69    114
dtype: int64
--------------------------------------------------------------------------------
sex
female    662
male      676
dtype: int64
--------------------------------------------------------------------------------
smoker
no     1064
yes     274
dtype: int64
--------------------------------------------------------------------------------
region
northeast    324
northwest    325
southeast    364
southwest    325
dtype: int64
--------------------------------------------------------------------------------
children
0    574
1    324
2    240
3    157
4     25
5     18
dtype: int64
--------------------------------------------------------------------------------
cbmi
normal-weight    222
obese            705
over-weight      380
under-weight      20
dtype: int64
--------------------------------------------------------------------------------
```

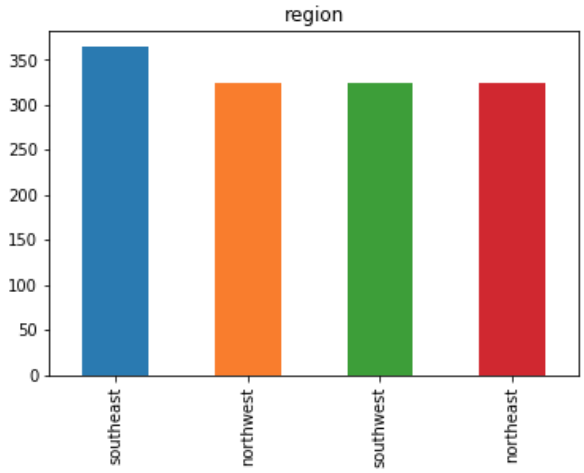
Exploratory Visualization

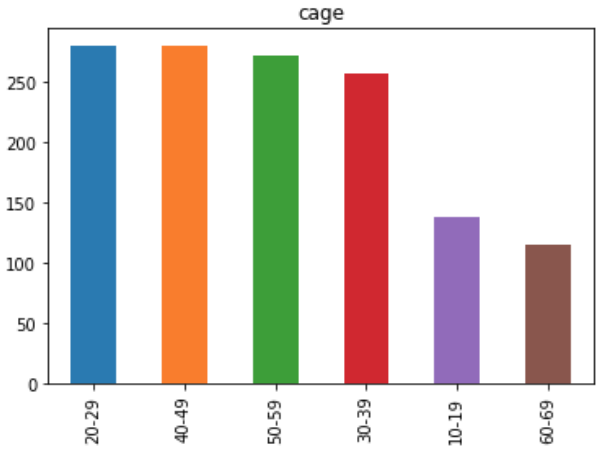
I used several bar graphs to pictorially understand the correlation among the input variables.

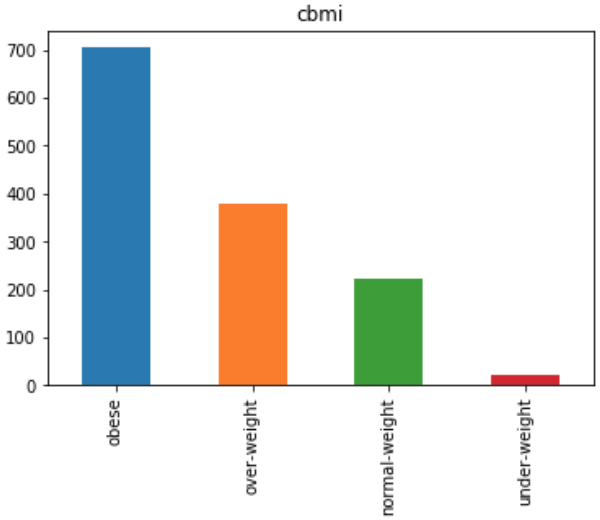
```
Data distribution analysys
```

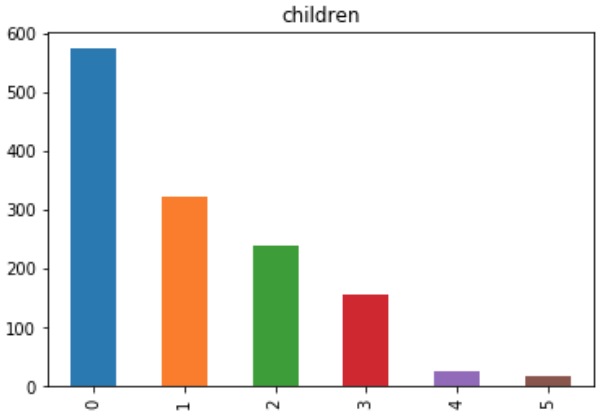




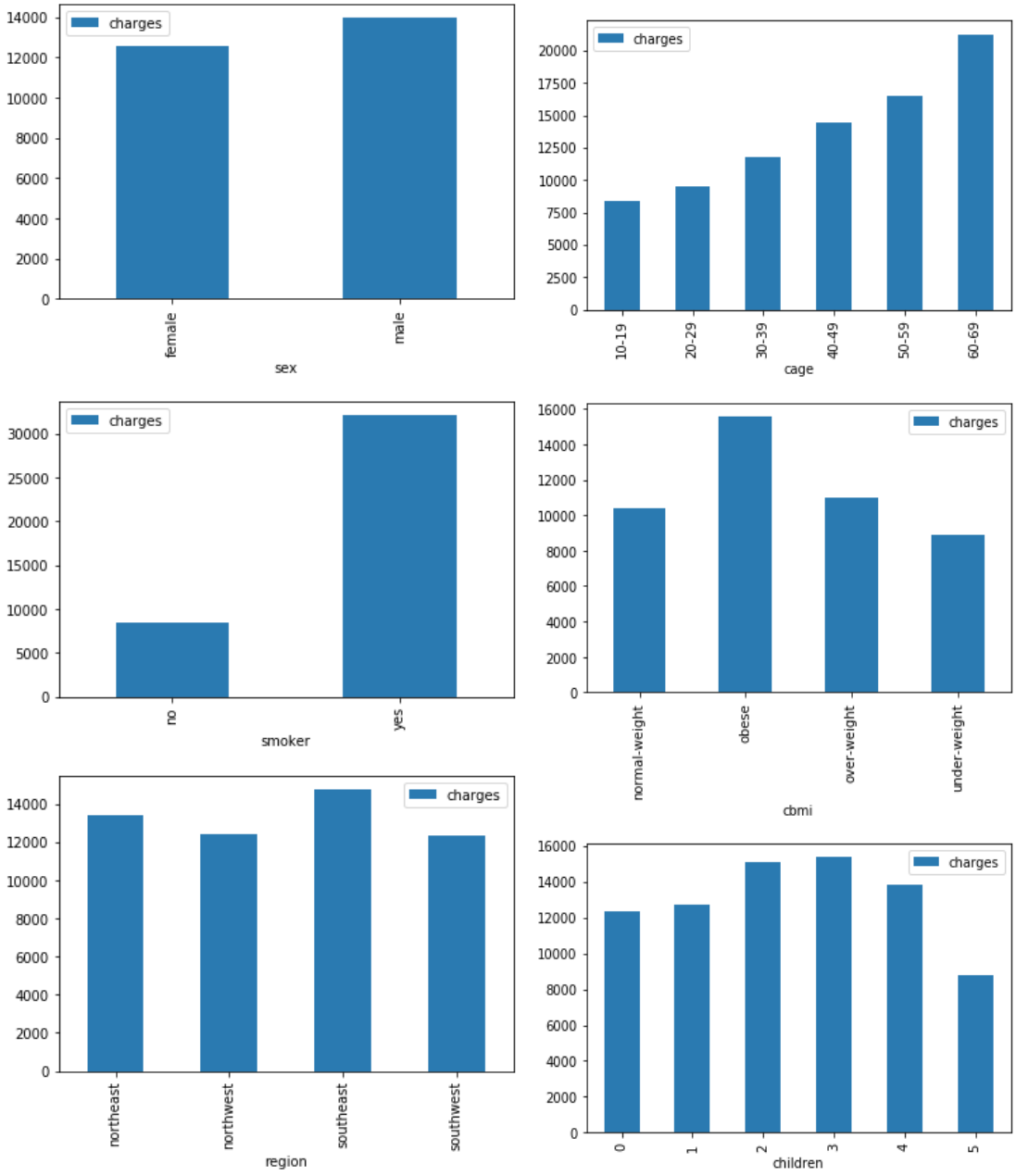








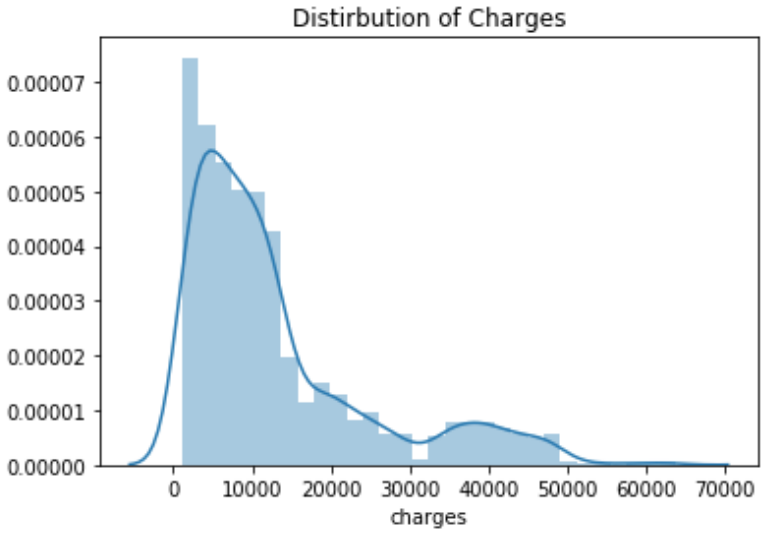
I also used several visualizations to understand the relationship between the independent features and target variable `charges`. Intention is to find out how of the input features have their impact on the target.



From the above bar graphs, we can deduce the below facts.

a. Insurance costs are higher among male population
b. Insurance costs are highest among the population of age groups 60-69.
c. Insurance costs increases among the smokers
d. Insurance costs increases among the obese population
e. Insurance costs are higher among the population in southwest region
f. Surprisingly, insurance costs are higher among the individuals with 2 or 3 children rather than with individuals with 4 or 5 children.

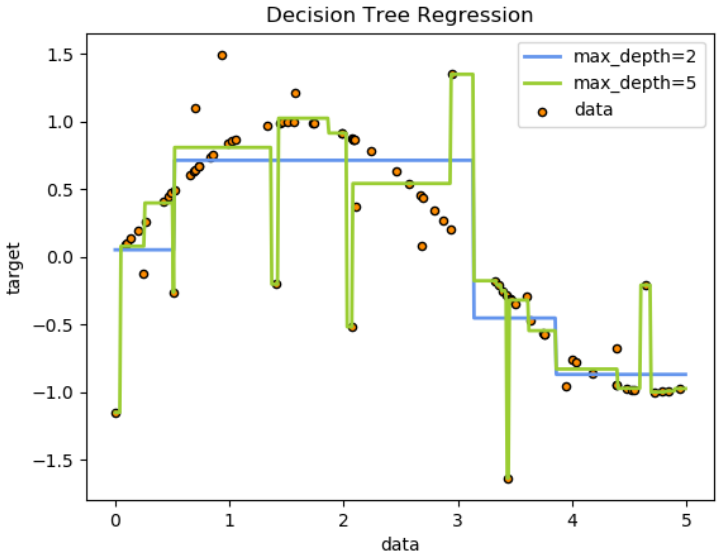
Further below distribution plot shows that the insurance costs of majority of population lies between 0 to 10,000 USD.



Algorithms and Techniques

As part of evaluation, I considered multiple regression algorithms like decision trees, Support Vector Machines for regression, etc. Based on the metrics, I choose to use decision tree technique for this project

The goal of decision tree is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. For instance, in the example below, decision trees learn from data to approximate a sine curve with a set of if-then-else decision rules. The deeper the tree, the more complex the decision rules and the fitter the model.



Decision Trees (DT) breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. Decision Trees(DT) takes a non-linearly separable decision surface and turns it into a linear decision surface by asking `Yes` or `No` or `True` or `False` questions.

Decision Trees are easy to understand and interpret. They require little data preparation like data normalization, dummy variables need to be created and blank values needs to be removed.

There are several terms involved with the decision trees.

Entropy:

a. Controls how a DT decides where to split the data.
b. It measures the level impurity in a bunch of samples

Information Gain:

a. The information gain is based on the decrease in entropy after a dataset is split on an attribute
b. It measured the level of purity in a bunch of samples

A decision tree is built top-down from a root node and involves partitioning the data into subsets that contain instances with similar values (homogenous). DT uses entropy to calculate the homogeneity of a sample. If the sample is completely homogeneous the entropy is zero and if the sample is an equally divided it has entropy of one. Constructing a decision tree is all about finding attribute that returns the highest information gain (i.e., the most homogeneous branches).

III. Methodology

Data Preprocessing

As mentioned above, DTs require some data pre-processing. I applied the below processing on the data.

a. As most of the models work well with numeric data, I need to convert the non-numeric data into numeric data. Feature `smoker` is a simple field with values `yes` or `no`. This can be easily represented as numeric 1 or 0 respectively.
b. There are couple of categorical features `sex` and `region` which take multiple values. To handle this situation, I created as many columns as possible values. For example, the feature `sex` takes two possible values `male` or `female'. Hence, I replaced the feature `sex` with two columns `sex_male` and `sex_female` for each row. For a value of `male`, the column `sex_male` will have a value of 1 and `sex_female` will have a value of 0.
   In a similar way, feature 'region` takes 4 possible values of `northeast`, `northwest`, `southeast` or `southwest`. Hence, the column `region` is split into 4 columns of `region_northeast`, `region_northwest`, `region_southeast` and `region_southwest`.
c. In the given data, there are two fields `bmi` and `age` which can take arbitrary values. There is no relation between any of two values among these features. I decided to convert them into categorical variables by mapping their values to specific set of values.
   For example, I classified the arbitrary value of `bmi` into categories `under-weight`, `normal-weight`, `over-weight` and `obese`. After that by applying the similar concept mentioned in b. I replaced the `bmi` column into multiple columns, one for each category.
   A similar mechanism is applied for `age` feature by dividing them into categories like `10-19`, `20-29`, `30-39` and so on.

Implementation

First, I divided the features and target data in to training sets and testing sets. I used 80% of the data for training and remaining 20% for testing set. This will provide a reasonable amount of data for training the model.

After splitting the data, I applied this data to make the regression models to learn for the data and measure their performance using the defined metrics.

Once after deciding the model, I used a random permutation cross-validator to split the features into multiple data-sets.

I used grid-search technique to fit the model against various parameters of the DT algorithm like

a. criterion: This parameter measures the quality of split with supported values of `mse`, `friedman_mse` and `mae`

b. splitter: This parameter is used to choose the strategy of split at each node with values of `best` and `random`
c. max_depth: This parameter defines the maximum depth of the decision tree. I used values from 1 thru 10.

IV. Results

Model Evaluation and Validation

The performance of various regression algorithms has been compared against the defined metrics. Training data was provided in chunks of 300, 600 and 900 data points. Coefficient of determination and explained variance scores are compared for this data points for below regression algorithms.

R-Squared	300	600	900
Decision Tree	0.9692	0.9559	0.9413
Support Vector	-0.0950	-0.0872	-0.0927
K-Neighbors	0.6266	0.6269	0.5828
NuSVR	-0.0569	-0.0700	-0.0537

Explained-Variance	300	600	900
Decision Tree	0.9692	0.9559	0.9413
Support Vector	0.0004	0.0007	0.0009
K-Neighbors	0.6285	0.6324	0.5895
NuSVR	0.0003	0.0005	0.0008

Justification

Clearly, metrics from the DT are closer to 1, which implies that DT is able to predict more reliably.
As further improvement, I fine-tuned the DT regressor by varying parameters like criterion, splitter and max-depth. I used grid-search algorithm to fit the DT against various cross-validation data-sets and above parameters. Grid-search algorithm has returned the best DT with below parameters.

a. criterion: `mae`, which is mean-absolute-error
b. splitter: `best`
c. max-depth: 6

I then trained and predicted using this model with above parameters and obtained the defined metrics as below.

- R-Squared score for training data: 0.8608
- R-Squared score for testing data: 0.8494
- Explained-variance score for training data: 0.8694
- Explained-variance score for testing data: 0.8573

V. Conclusion

I randomly selected few records from the original data-set and predicted the insurance costs for these records using the best DT estimator. The results are pretty much in-line with the actual values from the data set. The below are the actual and predicted values.

S No	Age	Sex	BMI	Children	Smoker	Region	Actual	Predicted
1	19	Female	27.900	0	Yes	Southwest	16884.9240	17081.08
2	18	Male	33.770	1	No	Southeast	1725.55230	2219.4451
3	19	Female	34.700	2	Yes	Southwest	36397.5760	36149.4835
4	30	Female	23.655	3	Yes	Northwest	18765.8754	19361.9988
5	19	Male	24.600	1	No	Southwest	1837.2370	2219.4451
6	31	Male	36.300	2	Yes	Southwest	38711.0000	38728.6775

Reflection

As seen from the above results, selected DT model seems to be predicting the costs accurately at the higher end of the costs scale.

From the above data, for records 1, 3, 4 and 6, the margin between the actual and predicted cost significantly less compared to the records 2 and 5.

For records 2 and 5, the actual insurance cost is less than 2000 USD and the model predicted the values to be above 2000 USD.

In case of other records, the actual insurance cost is more than 2000 USD and the model has predicted the values more accurately.