Detail Project Report

Insurance Premium Prediction

Document Version Control

Date	Version	Description	Author
02-Feb-2024	1.0	Abstract Introduction General Description	Roshan Kshirsagar
03-Feb-2024	1.1	Technical Requirements Data Requirements Data Pre-Processing Design Flow	Roshan Kshirsagar
05-Feb-2024	1.2	Data from user and its validation Rendering the Results Conclusion	Roshan Kshirsagar

Contents

Abstract

To give people an estimate of how much they need based on their individual health situation. After that, customers can work with any health insurance carrier and its plans and perks while keeping the projected cost from our study in mind. We are considering variables as age, sex, BMI, number of children, smoking habits and living region to predict the premium amount. This can assist a person in concentrating on the health side of an insurance policy rather than the ineffective part.

1.Introduction

1.1 Why this DPR Document?

The main purpose of this DPR documentation is to add the necessary details of the project and provide the description of the machine learning model with written code. This also provides the detailed description on how the entire project has been designed end-to-end.

Key points:

1. Describes the design flow
2. Implementations
3. Software requirements
4. Architecture of the project
5. Non-functional attributes like
   - 5.1 Reusability
   - 5.2 Portability
   - 5.3 Resource Utilization

2.General Description

2.1 Problem Perspective

The insurance premium prediction is a machine learning model that helps users to understand their insurance premium based on some input data.

2.2 Problem Statement

The main goal of this model is to predict insurance premium price based on some input data like bmi, gender, age etc.

2.3 Proposed Solution

To solve the problem, we have created a user interface for taking the input from the user to predict insurance premium price using our trained ML model after processing the input and at last the predicted value from the model is communicated to the user.

3.Technical Requirements

As technical requirements, we don't need any specialized hardware for virtualization of the application. The should have a device that has the access to web and the fundamental understanding of providing the input.

3.1 Tools Used

Python programming language and framework such as numpy, pandas, scikit-learn, vs code are used to build the whole model.

1. VS code is as IDE .
2. For visualization of the plots, Matplotlib, seaborn and Plotly are used.
3. Streamlit is used for the user interface.
4. Scikit learn and Python are used for modular programming and model building.
5. For version controlling of project we used Github.

4. Data Requirements

The data requirements is fulfilled through online platform. The dataset we are using is easily available at different online platforms.

4.1 Data collection

The data for this project is collected from the Kaggle dataset, the url for the dataset: https://www.kaggle.com/datasets/noordeen/insurance-premium-prediction

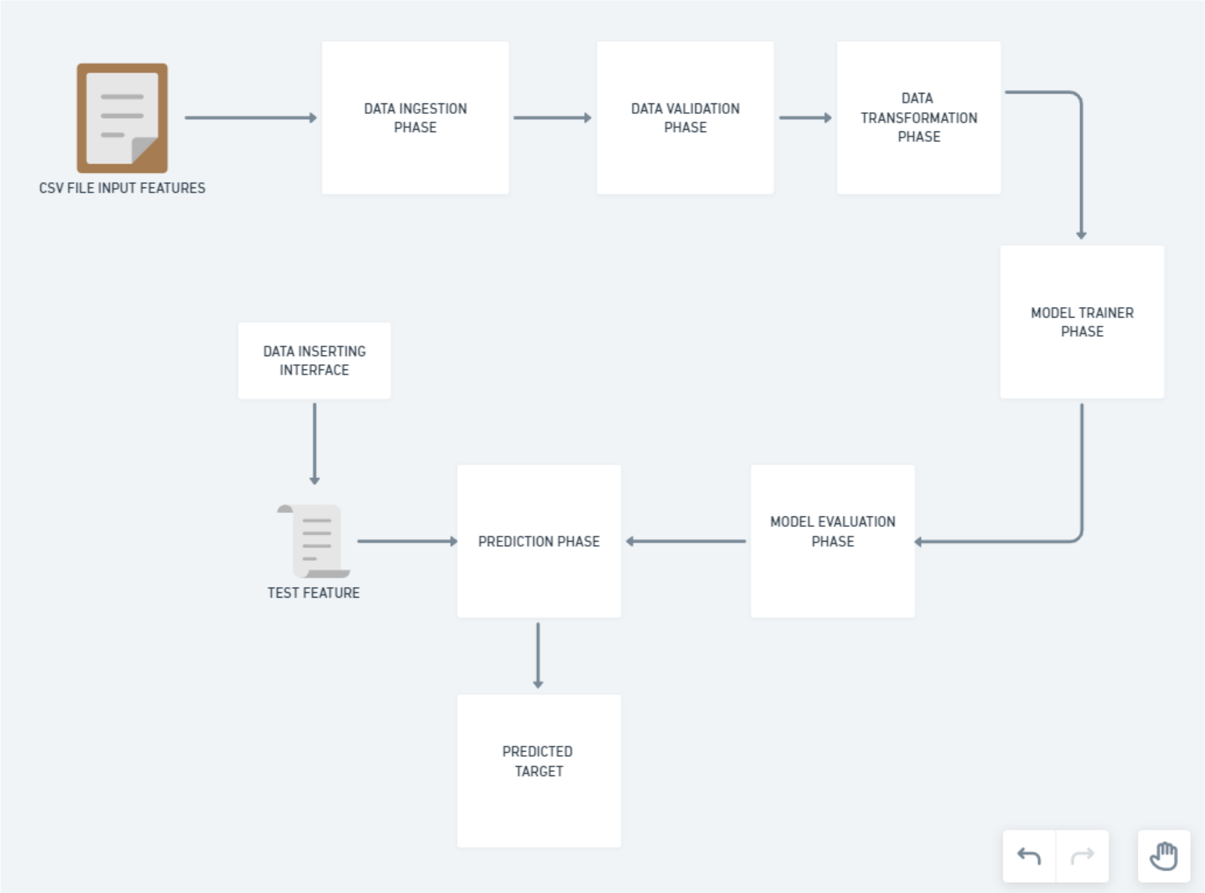
4.2 Data Description

The insurance premium data is publicly available on Kaggle. The information in the dataset is present in one csv files named as insurance.csv. Dataset contains 1339 rows which shows the information such as age, bmi, children, expenses, smoking habit etc.

4.3 Data Pre-processing

- Checked for info of the dataset, to verify the correct datatype of the columns.
- Checked for null values, because the null values can affect the accuracy of the model.
- Performed label encoding on the categorical columns
- Performed robust scaling for the transformation of the features.

5.Design Flow



5.1 Logging

In logging, at each time when an error or exception occurs, the event is logged into the system log file with reason and timestamp. This helps the developer to debug the system bugs and rectify the error.

5.2 Data from user

The data from the user is retrieved from the created streamlit web page.

5.3 Data Validation

The data provided by the user is then being processed by app.py file and validated. The validated data is then sent to the prepared model for the prediction.

5.4 Rendering the results

The data sent for the prediction is then rendered to the web page.

6.Conclusion

The Insurance Premium prediction system will predict the price for helping the customers with the trained knowledge with set of rules. The user can use this system to recognize the approximate values of their insurance premium.

7.Frequently Asked Questions (FAQs)

Que1. What's the source of data?

Ans. The data for training is provided by the client in multiple batches and each batch contain multiple rows. In many cases developer itself needs to collect the required data.

Que2. What was the type of data?

Ans. The data is the combination of numerical and categorical values. Basically it is tabular data which is collected over the time period.

Que3. What's the complete flow you followed in this project?

Ans. Firstly we ingested the data, then the validation of the data has done, pre-processing steps are implemented. Then we train multiple regression algorithms on the pre-processed data and build the final model on the best performed algorithms.

Que4. How logs are managed?

Ans. For logging we use our custom info logging class which basically track each and every activity of the project workflow.

Que5. What techniques were you using for data pre-processing?

Ans.

1. Removing unwanted attributes.
2. Visualizing relation of independent variables with each other and output variables.
3. Checking and changing distribution of continuous values.
4. Removing outliers.
5. Cleaning data imputing if null values are present.
6. Converting categorical data into numeric values.

Que6. How training was done or what models were used?

Ans.

1. Before dividing the data in training and validation set, we performed pre-processing over the dataset and made the final dataset.
2. As per the dataset training and validation data were divided.
3. Algorithms like linear regression, SVM, decision tree, random forest, xgboost were used based on the recall, final model was used on the dataset and we saved that model.

Que7. How was prediction done?

Ans. The testing files are shared by the client. We performed the same life cycle on the provided dataset. Then, on the basis of dataset, the model is loaded and prediction is performed.