High Level Design (HLD)

Insurance Premium Prediction

Revision Number: 1.1

Last date of revision: 23/11/2021

Document Version Control

Date Issued	Version	Description	Author
09/11/2021	1	Initial HLD-V1.0	Raja Arvindan R
23/11/2021	1.1	Model Training and Evaluation-V1.1	Raja Arvindan R

Contents

Abstract

Insurance is a policy that eliminates or decreases loss costs occurred by various risks. Various factors influence the cost of insurance. These considerations contribute to the insurance policy formulation. Machine learning (ML) for the insurance industry sector can make the wording of insurance policies more efficient. This study demonstrates how different models of regression can forecast insurance costs. And we will compare the results of models, for example, Multiple Linear Regression, Generalized Additive Model, Support Vector Machine, Random Forest Regressor, CART, XGBoost, k-Nearest Neighbors, Stochastic Gradient Boosting, and Deep Neural Network.

1 Introduction

1.1 Why this High-Level Design Document?

The purpose of this High-Level Design (HLD) Document is to add necessary details to the current project description to represent a suitable model for coding. This model is also intended to help detect contradictions prior to coding, and can be used as a reference manual for how the modules interact at a high level.

The HLD will:

- Present all of the design aspects and define them in details.
- Describe the user interface being implemented
- Describe the hardware and software interfaces
- Describe the performance and requirements
- Include design features and the architecture of the project
- List and describe the non-functional attributes like:
    - Security
    - Reliability
    - Maintainability
    - Portability
    - Reusability
    - Application compatibility
    - Resource utilization
    - Serviceability

1.2 Scope

The HLD documentation presents the structure of the system, such as the database architecture, application architecture, application flow (Navigations), and technology architecture. The HLD uses non-technical to mildly-technical term which should be understandable to the administrator of the system.

1.3 Definitions

Term	*Description*
UGV	Unmanned Ground Vehicle
Database	Collection of all the information monitored by this system
IDE	Integrated Development Environment
AWS	Amazon Web Services

2 General Description

2.1 Product Perspective

The Insurance Premium Prediction is machine learning based regression model which will help us to predict the health insurance premium and take the necessary action.

2.2 Problem Statement

The goal of this project is to give people an estimate of how much the 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. This can assist a person in concentrating on the health side of an insurance policy rather than the effective part.

2.3 Proposed Solution

The solution proposed here is a health premium prediction can be implemented some cases. First, we need to collect all given features like age, sex, BMI, children, region, smoker etc. If we provide all this features to model then model can able to predict insurance premium.

2.4 Data Requirement

Name	Description
Age	Age of the client
BMI	Body mass index
The number of kids	Number of children the client has
Gender	Male / Female
Smoker	Weather the client is smoker or not
Region	Where the client lives southwest, southeast, northwest, northeast.

2.5 Tools used

Python programming language and frameworks such as NumPy, Pandas, Scikit-learn, Flask, Git.

- PyCharm is used as IDE.
- For visualization of the plots, Matplotlib, Seaborn and Plotly are used.
- Frontend development is done using HTML/CSS
- Python Flask is used for backend development.
- GitHub is used as version control system.

2.6 Constraints

The Insurance Premium Prediction must be user friendly, as automated as possible and users should not be required to know any of the workings.

2.7 Assumption

Judicious use of predictive analysis has empowered health insurers to improve their premium pricing accuracy, create customized health insurance plans and services, and build stronger customer relationships. Thus, the main goal of this project is to predict the insurance premiums based on the behavioral data collected from the individuals so that insurance companies can make useful and accurate predictions. Based on these predictions, they can then evaluate the following decisions and make better judgement calls:

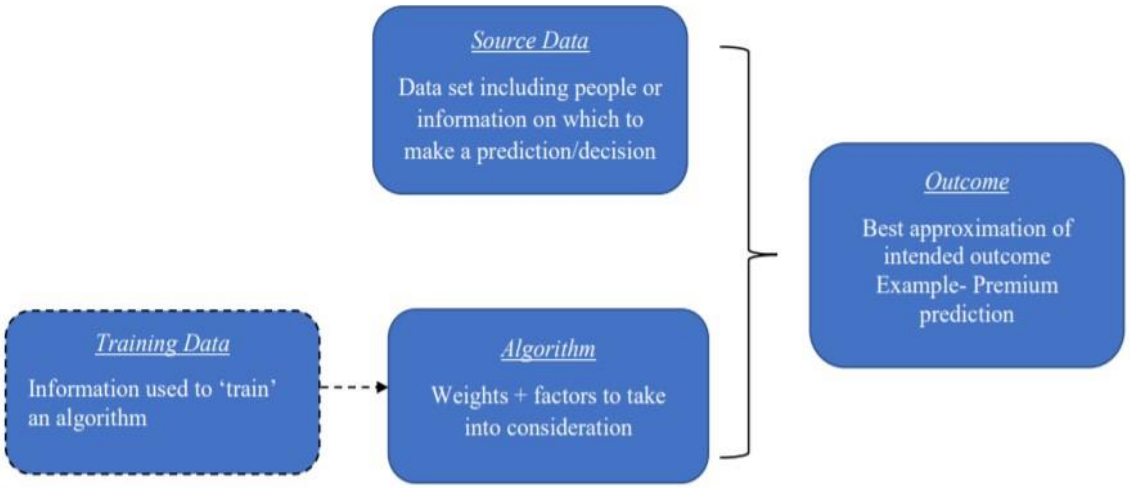
- Which individuals deserve which kind of insurance plan?
- Based upon an individual's behavior, predicting their premium helps in better risk management.

3. Design Details

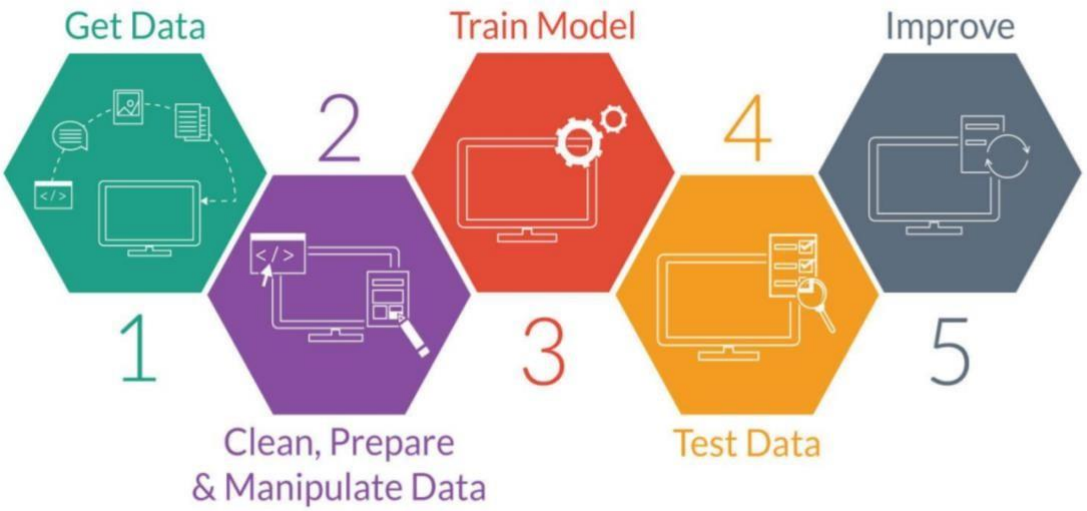
3.1 Process Flow

For predicting the Health Insurance Premium, we will use regression model. Below is the process flow diagram is as shown below.

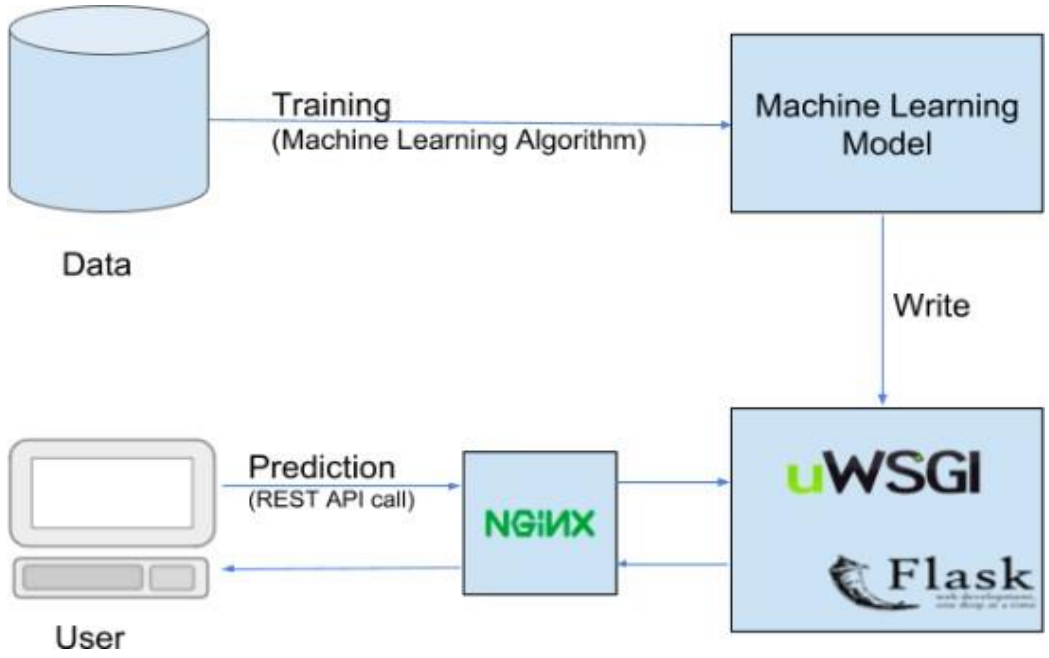
Proposed Methodology



3.1.1 Model Training and Evaluation



3.1.2 Deployment Process



3.2 Event Log

The system should log every event so that the user will know that process is running internally.

Initial Step-By-Step Description:

1. The system identifies at what step logging required
2. The system should be able to log each and every system flow
3. Developer can choose logging method. You can choose database logging / File logging as well.
4. System should not hang even after using loggings. Logging just because we can easily debug issues so logging is mandatory to do.

3.3 Error Handling

Should error be encountered, an explanation will be displayed as to what went wrong? An error will be defined as anything that falls outside the normal and intended usage.

4 Performance

We see that the accuracy of predicted amount was seen best i.e. 84% in gradient boosting decision tree regression. Other two regression models also gave good accuracies about 80% In their prediction.

4.1 Reusability

The code written and the components used should have the ability to be reused with no problems.

4.2 Application Compatibility

The different components for this project will be using as an interface between them. Each component will have its own task to perform, and it is the job of the python to ensure proper transfer of information.

4.3 Resource Utilization

When any task is performed, it will likely use all the processing power available until that function is finished.

5 Conclusion

Background In this project, three regression models are evaluated for individual health insurance data. The health insurance data was used to develop the three regression models, and the predicted premiums from these models were compared with actual premiums to compare the accuracies of these models. It has been found that Gradient Boosting Regression model which is built upon decision tree is the best performing model.

6 References

1. “Health Insurance Amount Prediction” Nidhi Bhardwaj, Rishabh Anand Delhi, India Dr. Akhilesh Das Gupta Institute of Technology & Management. Vol. 9 Issue 05, May-2020
2. Factors affecting health insurance premiums: Explore premiums: Explorative and predictive analysis Tarunpreet Kaur Iowa State University 2018.