High Level Design

Insurance Premium Prediction

Author: Rohit Shrimangle

Documents Version Control:

Date Issued	Version	Description	Author
25-07-2024	**1**	**Initial HLD**	**Rohit Shrimangle**
28-07-2024	**2**	**Final HLD**	**Rohit Shrimangle**

Contents:

Abstract

In recent years, the prediction of insurance premiums has become a critical area of focus for both insurance companies and consumers. This project aims to develop a machine learning model to accurately predict insurance premiums based on individual health profiles. The model leverages a variety of health-related features, including age, medical history, lifestyle factors, and biometric data, to provide personalized premium estimates. Our approach not only enhances the accuracy of premium predictions but also offers a scalable solution that can be integrated into insurance platforms to assist consumers in understanding and managing their insurance needs. This project demonstrates the potential of machine learning to transform the insurance industry by making premium calculations more personalized and data-driven.

1. Introduction

1.1. Why this High-Level Design Document

A High-Level Design (HLD) document in a machine learning project outlines the overall architecture and structure of the project. It serves as a bridge between the requirements specification and the detailed design, providing a comprehensive view of the system.

1.2. Why this High-Level Design Document

The scope section of an HLD document outlines the boundaries and limitations of the project. It clarifies what will be included in the project and what will be excluded, ensuring all stakeholders have a clear understanding of the project's extent.

2. General Description

Insurance price prediction involves using data analytics and machine learning techniques to estimate the cost of insurance premiums for individuals or entities based on various risk factors. The goal is to develop models that accurately forecast the potential costs associated with providing insurance coverage, thereby enabling insurers to set appropriate premium prices.

2.1 Problem Statement:

The goal of this project is 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. This can assist a person in concentrating on the health side of an insurance policy rather than the ineffective part.

2.3. Proposed Solution

With The classical machine learning tasks like Data Exploration, Data Cleaning, Feature Engineering, Model Building and Model Testing. Try out different machine learning algorithms that's best fit for the above case. Some Famous Algorithms: - Multiple Linear Regression, Decision tree Regression and Gradient Boosting, Decision tree, Regression using the algorithm which gives us the good accuracy for the prediction.

2.4. Further Improvements

Monitoring and Maintenance

- **Performance Monitoring**: Continuously monitor model performance in production using dashboards and automated alerts to detect drifts in accuracy or other performance metrics.
- **Retraining and Updates**: Establish a schedule for regular retraining of models with new data to keep them current and accurate.
- **Error Analysis**: Conduct regular error analysis to understand where the model is making mistakes and why, and use these insights to make improvements.

Advanced Techniques and Innovations

- **Deep Learning**: Explore deep learning models for complex, non-linear relationships in the data.
- **Transfer Learning**: Use transfer learning to leverage pre-trained models and adapt them to your specific problem.
- **Real-Time Data Integration**: Incorporate real-time data sources such as telematics for auto insurance or IoT devices for home insurance to continuously update risk assessments.

Business Alignment and Collaboration

- **Stakeholder Engagement**: Regularly engage with business stakeholders to ensure the model aligns with business objectives and provides actionable insights.
- **Domain Expertise**: Collaborate with domain experts to ensure that the model incorporates relevant business knowledge and assumptions.
- **Regulatory Compliance**: Ensure that the model complies with industry regulations and standards to avoid legal and financial repercussions.

Ethical Considerations and Fairness

- **Bias Detection and Mitigation**: Implement techniques to detect and mitigate bias in the model to ensure fair pricing for all policyholders.
- **Transparency**: Ensure transparency in how the model makes predictions, providing explanations to policyholders and regulators.

User Interface (UI) Development

- **Data Visualization**: Incorporate data visualization tools to help users understand the data and model predictions better. Use charts, graphs, and other visual aids.
- **Real-Time Interactivity**: Enable real-time interactivity in the UI to provide instant feedback on input changes, showing how different factors affect the insurance price predictions.
- **Feedback Mechanism**: Implement a feedback mechanism where users can provide input on model predictions, which can be used to further refine and improve the model.
- **Security and Privacy**: Ensure the UI follows best practices for security and privacy, protecting sensitive user data.
- **Mobile Compatibility**: Design the UI to be compatible with mobile devices, ensuring accessibility for users on different platforms

2.5. Data Requirements

1. Quality of Data

- **Accurate Policyholder Information**: Ensure that the personal details of policyholders (age, gender, Smoker) are correct and up-to-date.
- **Claims History**: Accurate records of past claims, including dates, amounts, and types of claims.

2. Relevance of Data

- **Target Variable**: Insurance premium amount.
- **Predictor Variables**:
    - Age of the policyholder.
    - Vehicle make and model for auto insurance.
    - Health conditions and history for health insurance.
    - Property value and location for home insurance.

3. Volume of Data

- **Sufficient Historical Data**: Collect historical data covering several years to capture different market conditions and risk factors.
- **Balanced Data**: Ensure that the data includes a wide range of policyholders with different risk profiles.

4. Data Types

- **Numerical**: Age, annual mileage (for auto insurance), property value.
- **Categorical**: Gender, occupation, vehicle type, property type.
- **Date/Time**: Date of birth, policy start date, claim dates.

5. Data Preparation

- **Missing Values**: Use mean/mode imputation or predictive modeling for missing values.
- **Outliers**: Use statistical methods to detect and treat outliers in claims data.
- **Scaling**: Normalize or standardize numerical features.

6. Data Distribution

- **Transformations**: Apply transformations if the target variable (premium amount) is skewed.
- **Multicollinearity**: Check for and address multicollinearity among predictor variables.

7. Data Segmentation

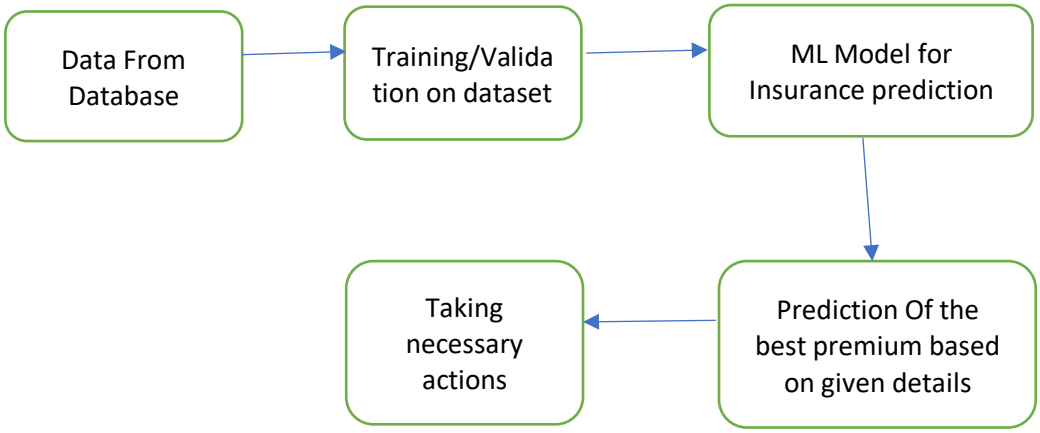
- **Training Set**: Use 70% of the data for training the regression model.
- **Validation Set**: Use 15% of the data for hyperparameter tuning.
- **Test Set**: Use the remaining 15% for final evaluation.

2.6. Tools Used

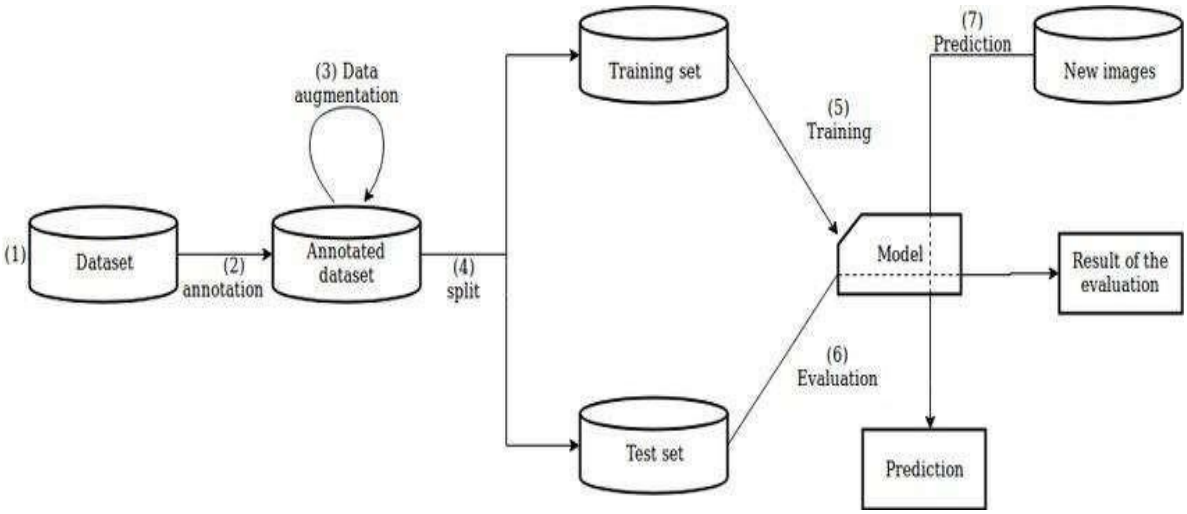
Python programming language, Pandas, numpy, scikit-Learn, SQL, Render, AWS, GitHub

3. Design Details

3.1. Process Flow



3.2. Model Training and Evaluation



4. Event Logs

Event logs are vital for ensuring the smooth operation, monitoring, and improvement of machine learning projects. By capturing detailed information about various events, from data ingestion to model deployment, they enable effective debugging, performance monitoring, compliance auditing, and continuous improvement. Implementing a comprehensive and well-structured logging system is essential for the success and reliability of ML projects.


```
{
 "timestamp": "2024-07-27T18:00:00Z",
 "event_id": "deployment_001",
 "event_type": "deployment",
 "description": "Deployed new model version to production",
 "severity_level": "info",
 "user": "deployment_pipeline",
 "additional_data": {
  "model_version": "v1.2.0",
  "deployment_status": "success"
 }
}
```


5. Performance

The Insurance price prediction model is performing well and giving the desired output. Here we are getting 88% Accuracy which is very good.

6. Reusability

The code written is reusable and the components used also have the ability to be reused without any problem.

7. Deployment

8. Conclusion

We have developed a ML model which will predict the insurance price required for a person on the basis of his/her physical, financial, gender and family which will help him/her to take the necessary steps towards his/her health and financial condition and helps to avoid the future crises.

9. References

1. https://www.kaggle.com/datasets/noordeen/insurance-premium-prediction

2. https://drive.google.com/file/d/1PUCqVKy21vtuKOYiBDhYJTUIa-7EP75g/view

3. https://google.com

4. https://chatgpt.com

5. https://www.youtube.com/

6. https://www.youtube.com/@campusx-official

7. https://www.youtube.com/@krishnaik06

8. https://www.linkedin.com