**Project Report**

**On**

**Insurance Premium Prediction**

*Submitted*

*In partial fulfilment for the award of the Degree of*

# PG-Diploma in Big Data Analytics

**(Know-IT Pune)**

|  |  |
| --- | --- |
| **Guided By:** | **Submitted By:** |
| Mr. Milind Kapase | Abhishek Suryawanshi (240843025002) |
|  | Devesh Gautam (240843025012) |
|  | Shubham Jeware (240843025039) |
|  | Piyush Kharalkar (240843025025)  Kaustubh Patil (240843025020) |

**Centre for Development of Advanced Computing**

**(C-DAC), ACTS (Pune- 411008)**

**CERTIFICATE**

**TO WHOMSOEVER IT MAY CONCERN**

**This is to certify that**

|  |
| --- |
| **Abhishek Suryawanshi(240843025002)** |
| **Devesh Gautam(240843025012)** |
| **Shubham Jeware(240843025039)** |
| **Piyush Kharalkar(240843025025)** |
| **Kaustubh Patil(240843025020)** |

Have successfully completed their project on

**Insurance Premium Prediction**

Under Guidance of Mr. Milind Kapase Sir

**Acknowledgement**

We extend our heartfelt gratitude to Mr. Milind Kapase, our esteemed Project Guide for "Insurance Premium Prediction" undertaken as a part of the PG-DBDA curriculum at CDAC ACTS Pune. His invaluable guidance and unwavering support have been instrumental throughout this project, from its inception to its completion. His expertise and insightful suggestions have significantly enriched our understanding and implementation of advanced concepts in the fields of PySpark, SparkSQL, Tableau, AWS and Machine Learning.

We also wish to express our sincere appreciation to Ms. Dhanashree Rangole, our dedicated Course Coordinator, for her continuous encouragement and assistance. Her commitment to fostering a conducive learning environment has been a driving force behind our academic growth.

Furthermore, we acknowledge CDAC ACTS Pune for providing us with the opportunity to undertake the project "Insurance Premium Prediction" This platform has enabled us to apply theoretical knowledge to real-world scenarios and expand our proficiency in diverse technical domains.

In this journey of knowledge and skill enhancement, we are grateful to have received support from various quarters. We would like to extend our thanks to all those who played a significant role, directly or indirectly, in our project's success. Your contributions have been invaluable.

**ABSTRACT**

Predicting insurance premiums accurately is crucial for both insurers and policyholders. This project explores machine learning models to predict insurance premiums based on customer demographics, health indicators, and financial attributes. We implemented multiple models, including Linear Regression, Lasso Regression, Decision Trees, XGBoost, and CatBoost, comparing their effectiveness in capturing patterns in the data.

**Table of Contents**

| **S. No** | **Title** | **Page No.** |
| --- | --- | --- |
| 1 | Introduction | 1-2 |
| 1.1 | Introduction | 1 |
| 1.2 | Objective | 2 |
| 2 | Literature Review | 3-4 |
| 3 | Methodology and Techniques | 5-10 |
| 3.1 | Methodology | 5 |
| 3.1.1 | Data Extraction and Storage | 5 |
| 3.1.2 | Data Cleaning and Transformation | 6 |
| 3.1.3 | Exploratory Data Analysis (EDA) | 6 |
| 3.1.4 | PowerBI Visualization and Dashboard Creation | 7 |
| 3.1.5 | Machine Learning Model Application | 8 |
| 3.2 | Dataset | 8 |
| 3.3 | Model Description | 9-10 |
| 3.3.1 | Preprocessing | 9 |
| 3.3.2 | Feature Standardization and Encoding | 9 |
| 3.3.3 | Linear Regression | 9 |
| 3.3.4 | Lasso Regression | 9 |
| 3.3.5 | Decision Tree Regressor | 10 |
| 3.3.6 | XGBoost Regressor | 10 |
| 3.3.7 | CatBoost Regressor | 10 |
| 3.3.8 | Evaluation Metrics | 10 |
| 4 | Implementation | 11-13 |
| 5 | Results | 14-19 |
| 5.1 | Analysis | 14 |
| 5.2 | PowerBI | 15 |
| 5.3 | Machine Learning | 16 |
| 5.4 | Deployment | 19 |
| 6 | Future Enhancements | 21 |
| 7 | Conclusion | 22 |
| 8 | References | 23 |

**Chapter 1: Introduction**

**1.1 Introduction**

Insurance premium prediction is a crucial aspect of the insurance industry, where companies assess risk and determine appropriate pricing for policyholders. Premiums are influenced by various factors, including customer demographics, health conditions, financial history, and lifestyle choices. Accurately predicting these premiums ensures fair pricing for customers while optimizing risk management for insurers.

The field of insurance generates vast amounts of data, encompassing variables such as income levels, health scores, previous claims, credit scores, and age. Analyzing this data is essential to identifying trends, understanding risk factors, and making data-driven pricing decisions. By uncovering patterns and correlations, data analysis helps insurers refine their models, enhance customer segmentation, and improve policy pricing strategies.

Machine learning, a branch of artificial intelligence, enables computers to learn from historical data and make predictions without explicit programming. In the context of premium prediction, machine learning models can be trained to estimate insurance costs based on past records. By leveraging algorithms such as Linear Regression, Lasso Regression, Decision Trees, XGBoost, and CatBoost, insurers can develop predictive models that improve accuracy, reduce biases, and streamline decision-making.

Visualization plays a key role in transforming complex insurance data into intuitive representations. Through graphs, charts, and interactive dashboards, stakeholders can explore trends, compare risk factors, and gain actionable insights into premium pricing. By combining machine learning and data visualization, this project aims to develop an efficient and interpretable premium prediction model that aids insurers in making fair, data-driven pricing decisions.

**1.2 Objective**

The objectives of the project work are as following-

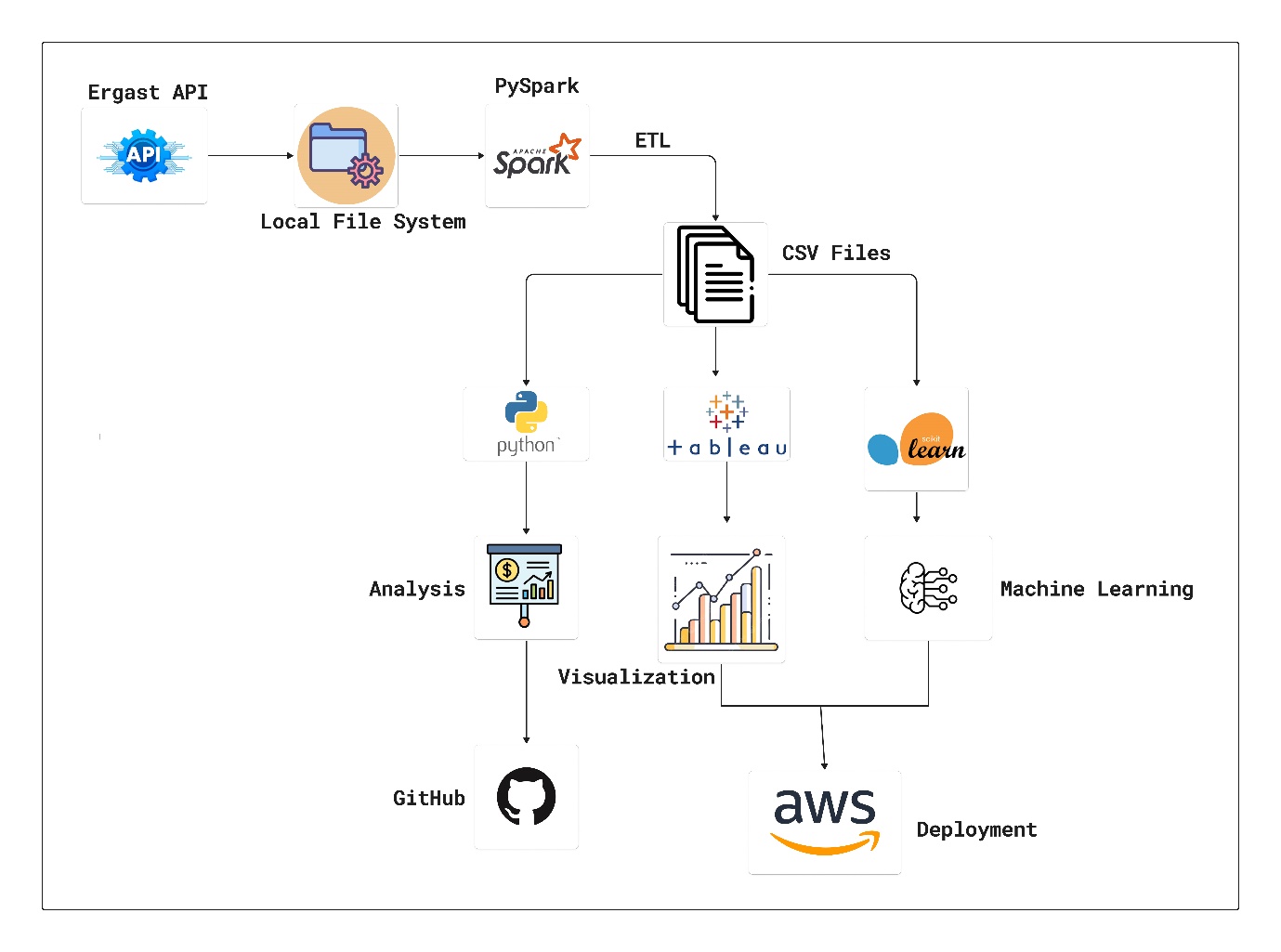
* Predict insurance Premium using Machine Learning Algorithms.
* Compare results between different Models
* Explore impact of regularization on our models.
* Provide insights for real world implementation.
* Handle data challenges and ensure fair prediction.

**Chapter 3: Methodology and Techniques**

**3.1 Methodology:**

The methodology employed for the insurance premium prediction project follows a structured approach encompassing data preprocessing, feature engineering, exploratory data analysis, model training, evaluation, and visualization. By leveraging various machine learning techniques, the project aimed to develop an accurate and interpretable model for predicting insurance premiums.

Each phase of the methodology is carefully designed to ensure data quality, model effectiveness, and real-world applicability. The following sections elaborate on these phases, providing a comprehensive understanding of how the project progressed and successfully identified key factors influencing insurance premiums.



**Fig.1 Proposed Work**

**3.1.1 Data Extraction and Storage:**

The project began by collecting raw insurance premium data, which included key attributes such as customer demographics, health scores, previous claims, credit scores, and income levels. This data was stored in the local file system for further processing and analysis. The dataset provided a comprehensive foundation for building predictive models, ensuring that all relevant factors influencing insurance premiums were considered.

**3.1.2 Data Cleaning and Transformation :**

The data cleaning process began by creating a SparkSession named "Data Cleaning" and enabling Hive support for handling SQL tables. The dataset was then processed using the following steps:

* Categorizing Features: Identified and classified columns as numerical or categorical.
* Handling Missing Values:
  + Counted unique values in categorical columns (excluding "Policy Start Date").
  + Converted "Policy Start Date" into an actual date format.
  + Counted missing values per column and removed rows with missing values in key fields.
  + Replaced missing categorical values with the most frequent value (mode).
  + Replaced missing numerical values with the median to maintain consistency.
* Data Formatting & Cleaning:
  + Standardized column names by removing spaces and special characters.
* Saving the Cleaned Data: Stored the preprocessed dataset for further analysis and model training.

This structured approach ensured data consistency, completeness, and accuracy, making it suitable for predictive modeling.

**3.1.3 Exploratory Data Analysis (EDA) with Python Libraries:**

The transformed and cleaned dataset underwent Exploratory Data Analysis (EDA) using NumPy, pandas, seaborn, and matplotlib to uncover patterns and insights. Normality testing (Shapiro-Wilk Test) revealed that most numerical columns were not normally distributed, necessitating non-parametric statistical tests. Spearman correlation analysis found weak or no significant relationships between Annual Income, Age, Number of Dependents, Location, and Premium Amount. Mann-Whitney U Test showed no significant difference in premiums between genders, while Kruskal-Wallis tests identified significant variations in Premium Amount based on Marital Status and Policy Type, further explored using Dunn’s Posthoc Test. Visualizations, including boxplots, heatmaps, and histograms, provided deeper insights into premium distributions and correlations. These analyses offered a comprehensive understanding of insurance premium determinants and their statistical relationships.

**3.1.4 PowerBI Visualization and Dashboard Creation:**

With valuable insights derived from EDA, the next step involved employing Tableau for visualization. Using Tableau's intuitive interface, interactive visualizations were crafted to represent complex race data comprehensibly. Dashboards were designed to provide interactive insights of drivers and constructors, enabling one to monitor driver and team performance effectively.

**3.1.5 Machine Learning Model Application:**

Building on the preprocessed and enriched dataset, the project applied various machine learning algorithms to predict insurance premiums. The models implemented included Linear Regression, Lasso Regression, Decision Trees, XGBoost, and CatBoost, each offering unique strengths in capturing relationships between features and premium amounts.

These models were evaluated based on accuracy, interpretability, and robustness, helping to analyze the impact of different customer attributes such as income, credit score, health score, and claims history. By leveraging machine learning, the project aimed to enhance risk assessment, pricing strategies, and overall decision-making in the insurance domain.

**3.2 Dataset**

The dataset used for this project contains insurance premium-related data, capturing key attributes such as customer demographics, health indicators, financial history, and past claims. This dataset serves as the foundation for building predictive models that estimate insurance premiums based on risk factors.

The data includes structured fields such as age, annual income, credit score, health score, previous claims, and location, which are critical in determining premium amounts. Various preprocessing techniques were applied, including handling missing values, feature transformations, and scaling, to ensure data quality and model readiness.

By leveraging this dataset, the project aims to analyze patterns, correlations, and risk factors, ultimately improving insurance pricing strategies and enhancing predictive accuracy through machine learning models.

**3.3 Model Description**

**3.3.1 Preprocessing**

**Data Preprocessing**

Data preprocessing is a critical step in ensuring the accuracy and reliability of our predictive model. The process begins with creating a SparkSession named "Data Cleaning", which enables efficient handling of large datasets and Hive support for SQL-based operations.

Handling Data Types & Missing Values

We first identify and categorize columns into numerical and categorical types to streamline further transformations. Unique values are counted for each categorical column, excluding "Policy Start Date", which is then converted into an actual date format for proper temporal analysis. Missing values are systematically handled—rows with missing values in key columns are dropped, while missing categorical values are replaced with the most frequent value (mode), and numerical values are imputed using the median to maintain data integrity.

**3.3.2 Feature Standardization and Encoding:**

To ensure a fair comparison across different machine learning models, features were standardized to bring them onto a common scale. Feature standardization was applied, transforming numerical features to have a mean of zero and a standard deviation of one.

For categorical variables such as location, smoking status, and policy type, encoding techniques like one-hot encoding were used to convert them into numerical representations. This process ensures that the dataset is properly structured for integration into machine learning models, improving model performance and interpretability.

**Business Assumptions for Scaling**

The dataset initially lacked variance, making traditional transformations like log transformation and power transformation ineffective. To address this, we introduced business-driven scaling assumptions to create a structured relationship between features and premium amounts.

Higher Income → Lower Risk → Lower Premium

Higher Health Score → Lower Risk → Lower Premium

Lower Credit Score → Higher Risk → Higher Premium

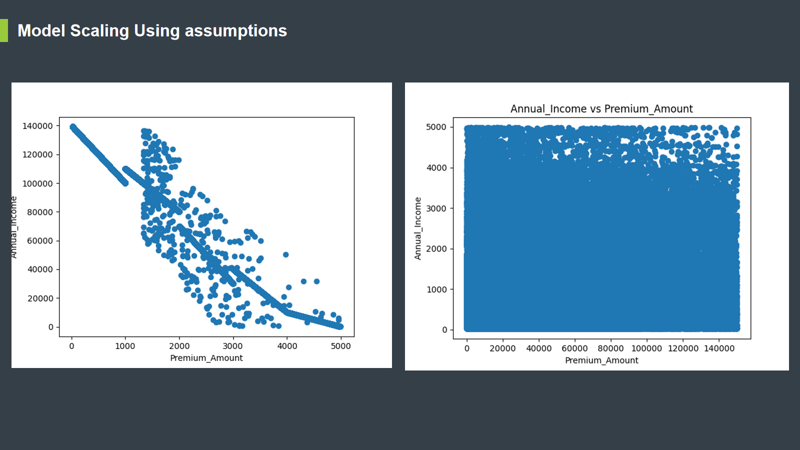
Age-Based Risk & Premium Structure:

Older Individuals → Higher Risk → Higher Premium

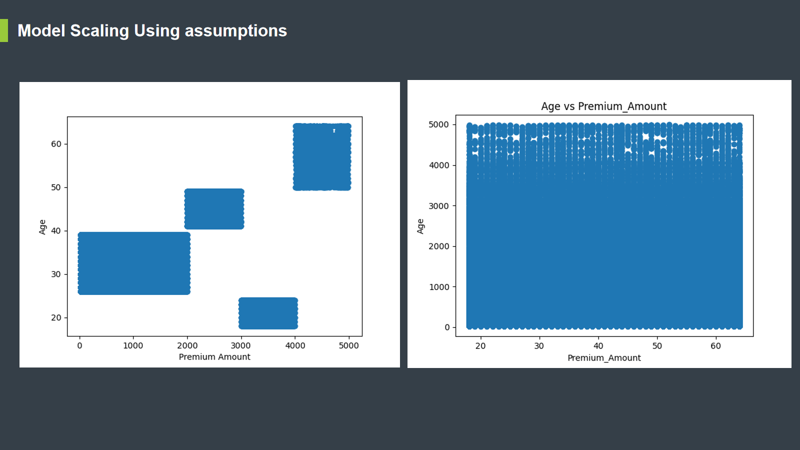
Younger Individuals → Not Earning Enough → Higher Risk → Higher Premium

Middle-Aged Individuals → Stable Earnings & Good Health → Lower Risk → Lower Premium

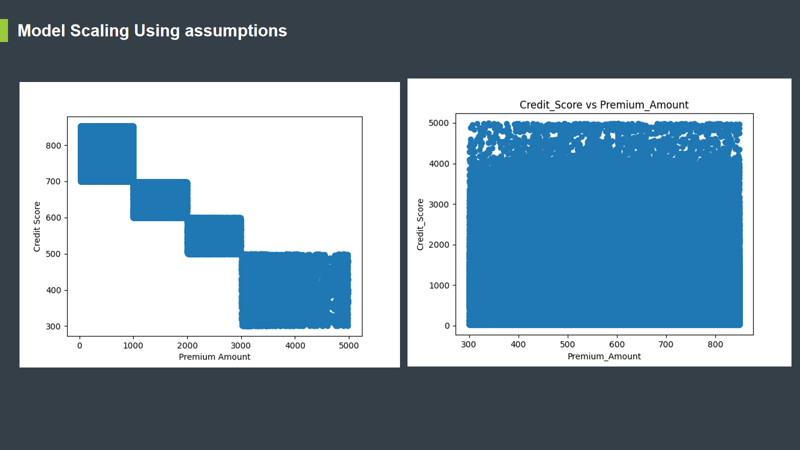
This approach ensured that the data reflected real-world insurance risk assessment, improving the model’s ability to learn meaningful patterns.



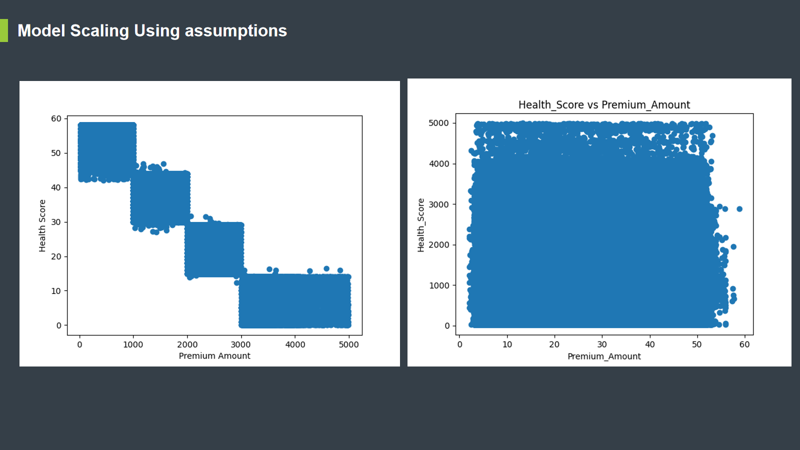
**Fig: Old(right) vs New(left) data for Annual income vs Premium amount**



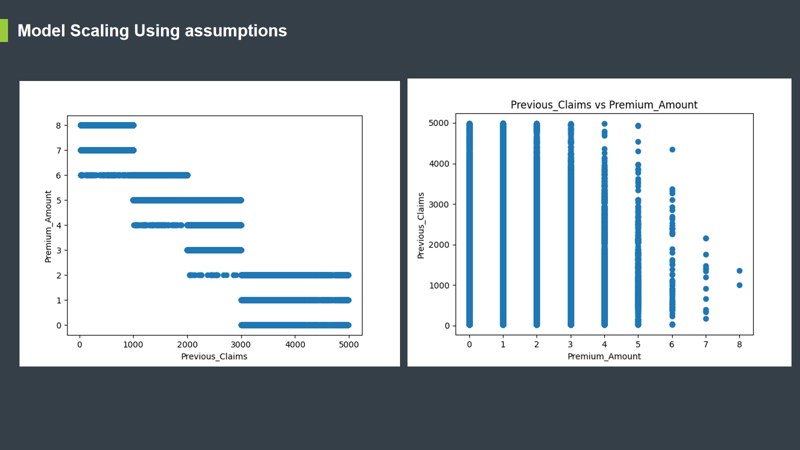
**Fig: Old(right) vs New(left) data for Age vs Premium amount**



**Fig: Old(right) vs New(left) data for Credit score vs Premium amount**



**Fig: Old(right) vs New(left) data for Health score vs Premium amount**



**Fig: Old(right) vs New(left) data for Previous claims vs Premium amount**

**3.3.6 Linear Regression**

Linear Regression was implemented as a baseline model to predict insurance premiums by assuming a linear relationship between input features and the target variable. It calculates the best-fit line that minimizes the difference between actual and predicted premiums. While it provides interpretability and simplicity, its inability to capture nonlinear relationships limits its predictive power. The model was evaluated using R² score, Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). Although it established an initial benchmark, it was outperformed by more advanced models that better handled complex interactions between features.

**3.3.7 Lasso Regression**

Lasso Regression, an extension of Linear Regression, incorporates L1 regularization, which penalizes large coefficients and forces some to shrink to zero. This aids in feature selection, making it particularly useful when dealing with high-dimensional data. By reducing the impact of irrelevant features, Lasso helps mitigate overfitting while maintaining model simplicity. However, since it still assumes a linear relationship, it struggles with capturing nonlinear dependencies in the data. The model was evaluated using R² score, MSE, RMSE, and feature importance analysis, showing an improvement over standard Linear Regression but still lagging behind tree-based models.

**3.3.8 Decision Tree Regressor**

Decision Trees were employed to model nonlinear relationships between features and premium amounts. The algorithm recursively splits data based on feature values, creating a tree-like structure where each branch represents a decision rule. Decision Trees are highly interpretable and can automatically capture feature interactions. However, they are prone to overfitting, especially when grown too deep. The model's performance was evaluated using MSE, RMSE, and feature importance scores. While it provided better accuracy than regression models, it was still outperformed by ensemble techniques such as XGBoost and CatBoost.

**3.3.9 XGBoost Regressor**

XGBoost, an advanced gradient boosting algorithm, was implemented to improve predictive performance. It works by training multiple weak learners (decision trees) sequentially, correcting errors from previous iterations. XGBoost is known for its handling of missing values, regularization capabilities, and scalability, making it a powerful choice for structured data. The model was fine-tuned using hyperparameters such as learning rate, max depth, and the number of estimators to enhance performance. Evaluated using MSE, RMSE, and feature importance, XGBoost demonstrated superior accuracy and generalization ability, making it one of the top-performing models in the project.

**3.3.10 CatBoost Regressor**

CatBoost, another gradient boosting model, was specifically designed to handle categorical data efficiently without requiring extensive preprocessing. It applies an ordered boosting approach, reducing overfitting while improving stability. Since the dataset contained categorical features like smoking status, location, and policy type, CatBoost offered a competitive edge over other models. The algorithm was tuned for optimal performance, and evaluation metrics such as MSE, RMSE, and feature importance analysis were used to assess its effectiveness. CatBoost performed comparably to XGBoost, particularly excelling in handling categorical-heavy data.

**3.3.9 Evaluation Metrics and Confusion Matrices:**

The project evaluates model performance beyond just accuracy, ensuring a more comprehensive assessment. Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R² score are used to measure how well each model predicts insurance premiums. MSE and RMSE quantify the error magnitude, while R² score determines how much variance in premium amounts is explained by the model.

For tree-based models like XGBoost, CatBoost, and Decision Trees, feature importance analysis is used to identify key factors influencing predictions. This helps in understanding which attributes—such as income, health score, and credit score—most significantly impact insurance pricing.

By incorporating multiple evaluation metrics, the project ensures that the selected model is accurate, interpretable, and generalizable, allowing for fair and data-driven premium predictions.

**Chapter 4: Implementation**

1. Use of Python Platform for writing the code with PySpark

2. Hardware and Software Configuration:

**Hardware Configuration:**

● CPU: intel i5 or similar, 8 GB RAM

**Software Required:**

1. **Local File System**: Local File System provides a scalable and reliable method for storing and managing data on a single machine or networked storage devices.

2. **PySpark**: The Python library for Apache Spark, used for big data processing and analytics. It provides APIs for ETL tasks, machine learning, and data analysis on distributed clusters.

3. **Python**: The programming language used to write PySpark scripts, interact with various components of the project, and create Streamlit apps.

4. **Google Colab:** Google Colab is a cloud-based interactive environment that allows users to write and execute Python code, making it ideal for data science and machine learning workflows. Google Colab supports seamless integration with Google Drive, allowing easy storage and sharing of notebooks. Its collaborative features, automatic resource management, and pre-installed libraries make it a powerful and convenient platform for coding, data analysis, and visualization without the need for local setup.

1. **Visualization** Tools: Visualizing ETL and analysis results with the help of tools like Matplotlib, and Seaborn.
2. **GitHub**: A version control platform that helped to manage and collaborate on the project’s source code. It's a central place for storing and tracking changes to codebase.
3. **Koyeb**: Koyeb is a serverless platform for deploying web applications and APIs.

It provides automatic scaling, global deployment, and built-in networking.

It also helps in enabling seamless CI/CD workflows.

1. **Netlify** : Netlify is a cloud-based platform for deploying web applications and static websites. It also enables fast and scalable front-end deployments.

**Chapter 5: Results**

**5.1 Analysis**

To evaluate the performance of different regression models, we used Mean Squared Error (MSE), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared (R²) as key performance metrics. The comparison of Linear Regression, Lasso Regression, XGBoost, Decision Tree Regression, and CatBoost is summarized in the table below:

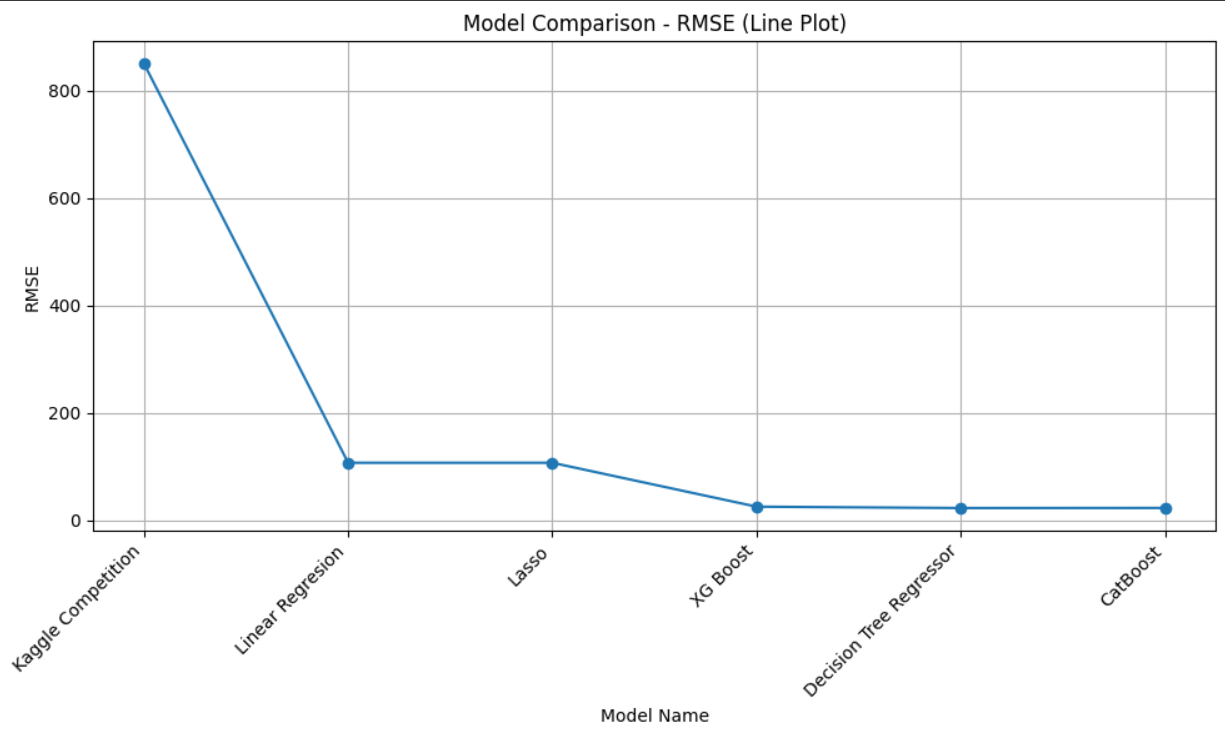
Linear Regression & Lasso Regression: Both models exhibited similar performance, with high MSE (~11364) and RMSE (~106.60), indicating significant error margins. Their R² score (0.9848) suggests they fit the data well but are outperformed by more advanced models.

XGBoost: This model significantly reduced error values, achieving an MSE of 618.29, MAE of 5.39, and RMSE of 24.86, indicating better predictive power. The R² score (0.9991) highlights its superior ability to capture data variance.

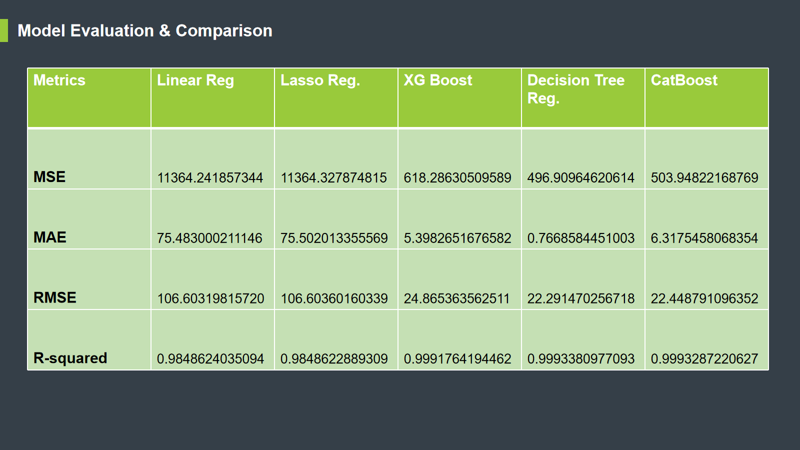
Decision Tree Regression: This model outperformed XGBoost in terms of error reduction, with MSE (496.90), RMSE (22.29), and an R² score of 0.9993, making it one of the best-performing models.

CatBoost: CatBoost provided competitive performance, with MSE (503.94), RMSE (22.44), and an R² score of 0.9993, closely aligning with Decision Tree Regression.

From these results, Decision Tree Regression and CatBoost proved to be the most effective models for predicting the target variable, demonstrating minimal error and maximum explained variance.

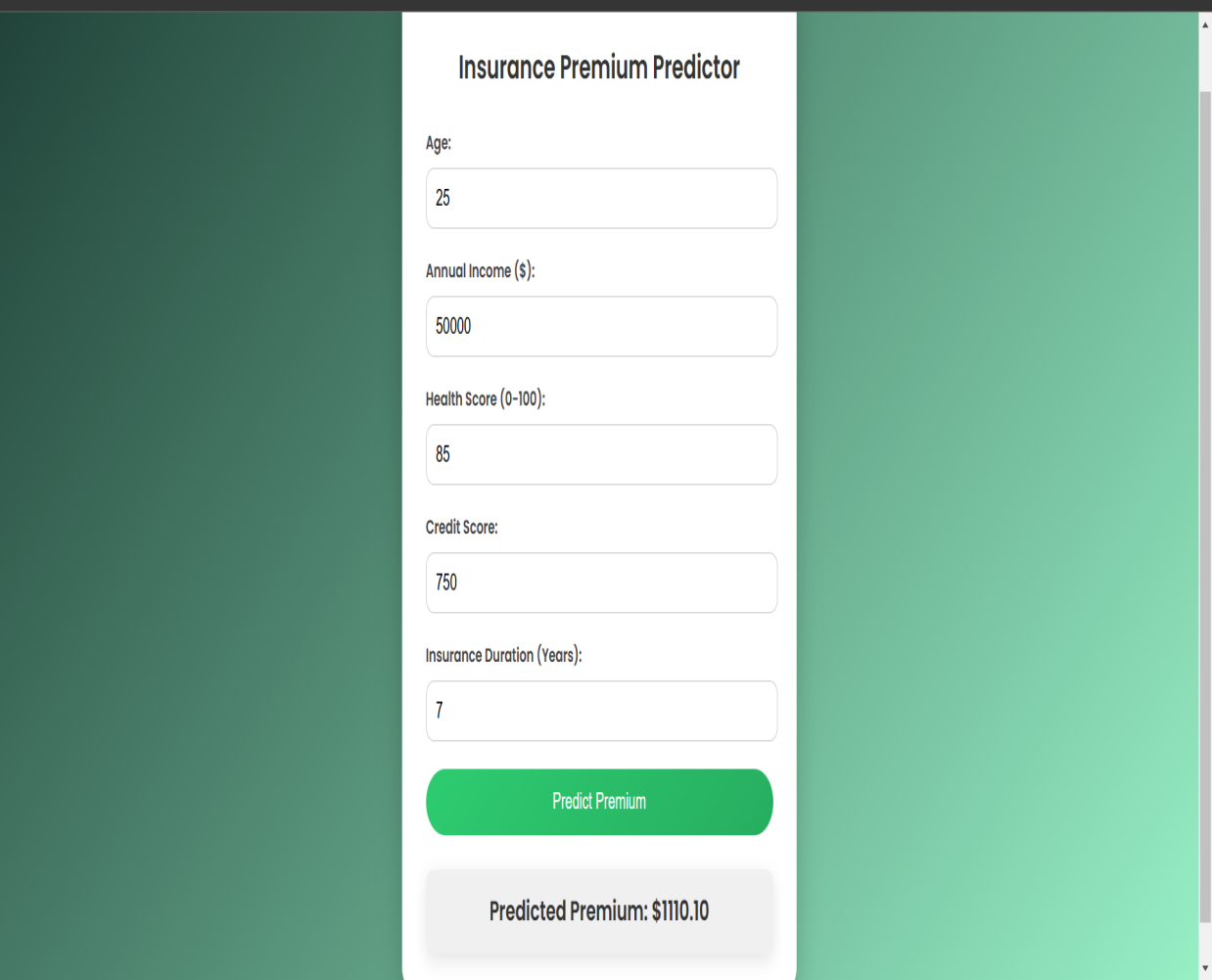


**Fig: Comparison between the different models used**



**Fig: Comparison using different metrics**

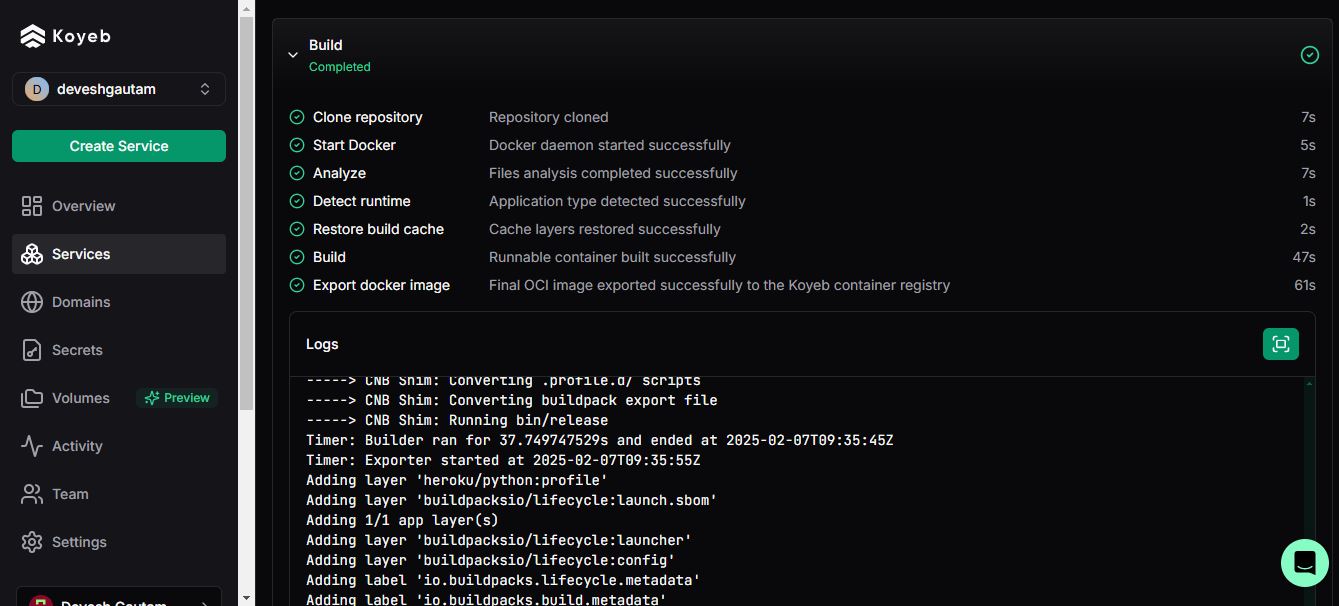
**5.4 Deployment**



**Fig 13 Home page**

# Frontend2

# Fig 14 Netlify log

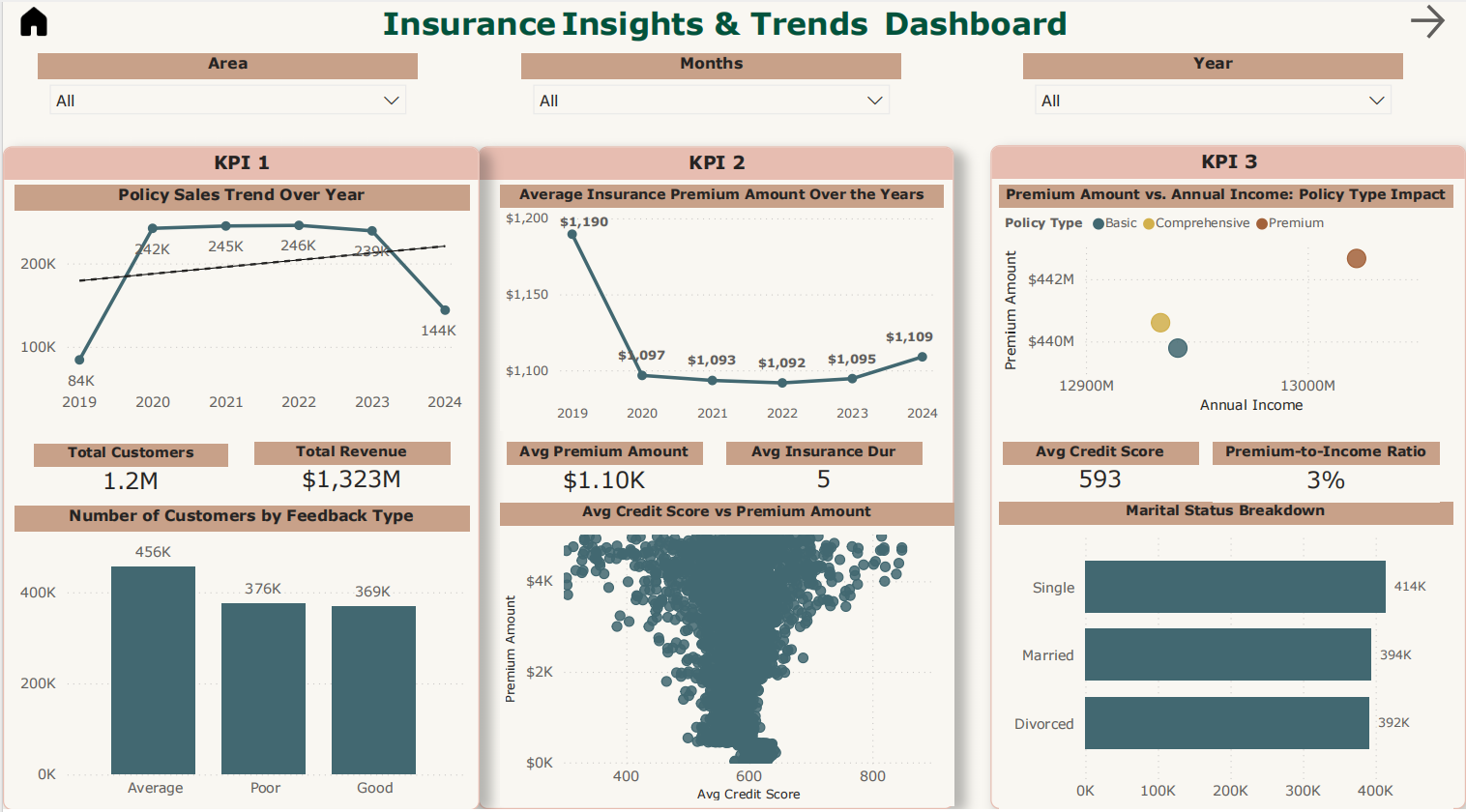


# Fig 15 Koyeb log

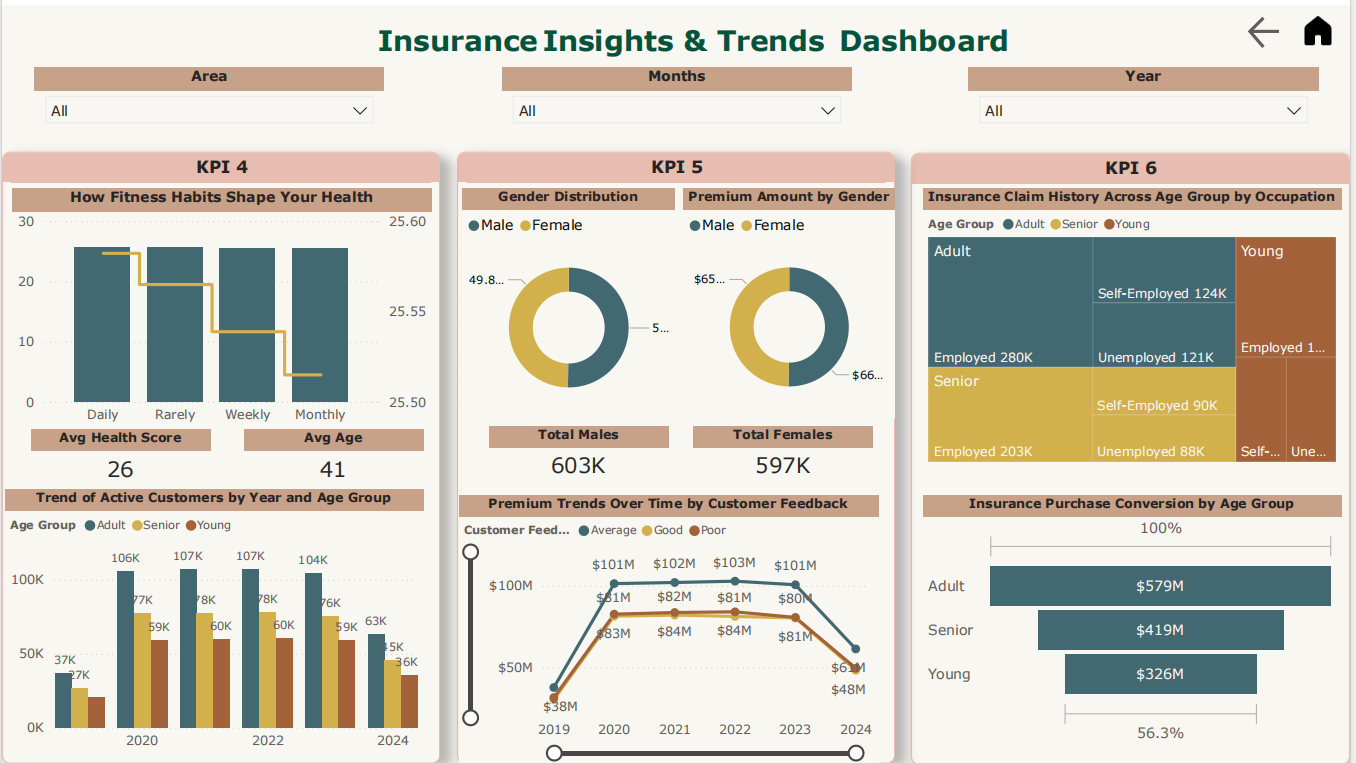
**5.4 Dashboard**

# Screenshot 2025-02-13 134542

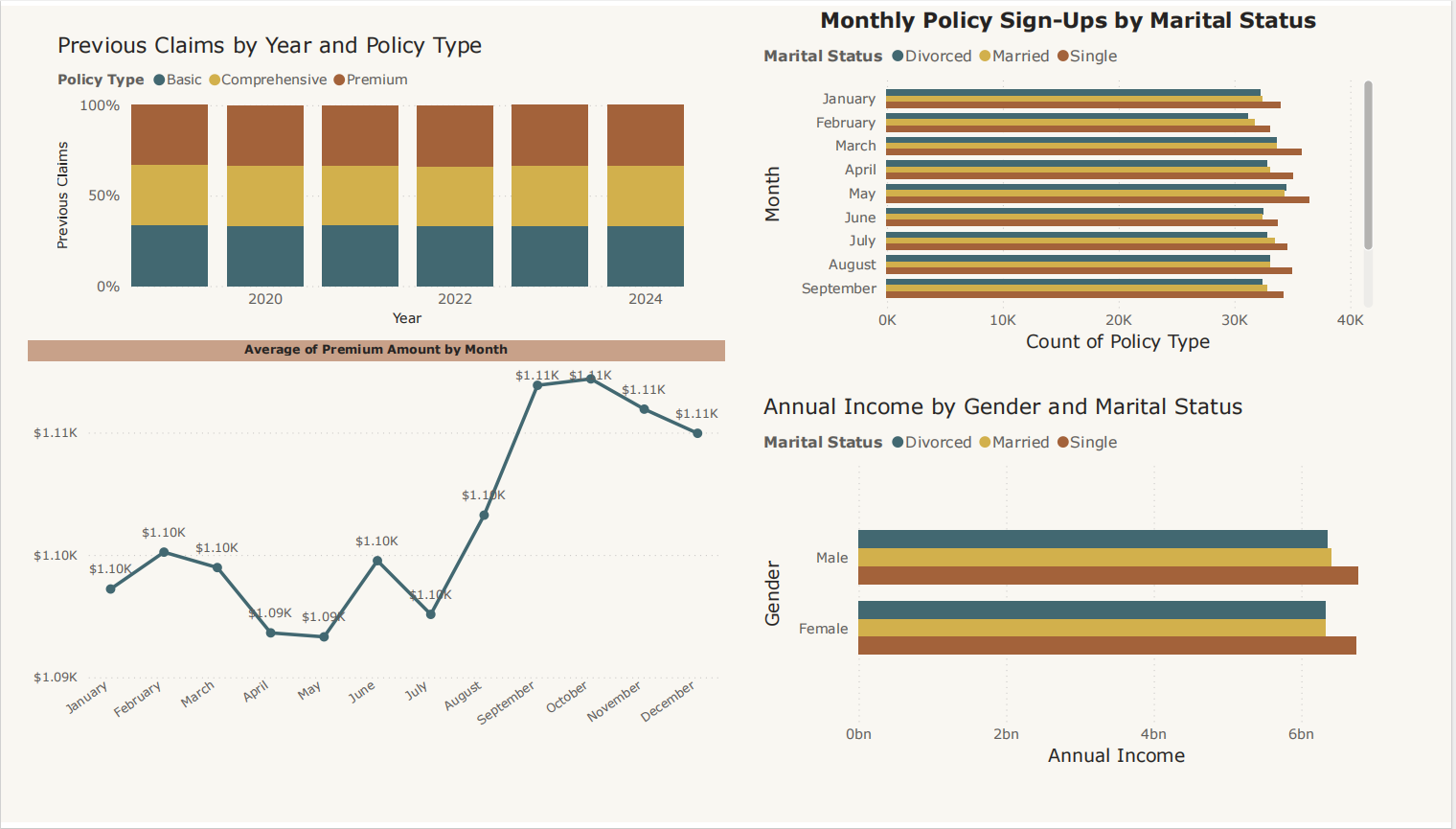
# Fig 16 Dashboard 1



# Fig 17 Dashboard 2



# Fig 18 Dashboard 3



# Fig 19 Dashboard 4

# Screenshot 2025-02-13 134724

# Fig 20 Dashboard 5

# Chapter 6: Future Enhancement

1. Improve Model Accuracy: Fine-tune hyperparameters to get better predictions.
2. Add More Features: Include real-world factors like claim history, policyholder behavior, or economic trends.
3. Try Deep Learning: Use neural networks to capture more complex patterns in the data.
4. Make Real-Time Predictions: Deploy the model for instant premium calculations.
5. Improve Explainability: Use methods to understand how the model makes decisions.
6. Deploy in Production: Integrate the model into a cloud-based system for real-world use.

# Chapter 7: Conclusion

Our project provides a robust, data-driven framework for predicting insurance premium values with high accuracy. By leveraging historical policyholder data, risk factors, and various machine learning models, we have successfully built a system that offers valuable insights into premium estimation based on multiple influencing factors.

Moving forward, several key enhancements can further improve the accuracy and usability of our model. Future developments will focus on integrating more advanced machine learning techniques, such as deep learning and ensemble methods, to refine predictions. Additionally, incorporating real-time data feeds, such as economic trends, medical advancements, and policy changes, will enable a more dynamic and adaptive pricing model. Furthermore, enhancing the visualization and interpretability of predictions through interactive dashboards will make insights more accessible and actionable for insurers and policyholders.

By continuously evolving and integrating cutting-edge technologies, our framework aims to become an indispensable tool in the insurance sector, improving risk assessment, policy pricing, and customer experience.

# CHAPTER 8: References

* <https://archive.apache.org/dist/spark/docs/2.4.5/api/python/>
* <https://scikit-learn.org/dev/index.html>