

Insurance Premium Estimation: Data-Driven Modeling and Real-Time Predictions

¹Aditya A. Desai,

²Saurabh S. Patil,

³Shreyas P. Mandavkar,

B. Tech. (Elect), RIT Islampur,

B. Tech. (IT), RIT Islampur,

B. Tech. (CSE), RIT Islampur,

Abstract - This project's abstract centers on the task of predicting individual insurance premiums by leveraging personal health data and assessing seven regression models, including Linear Regression, Decision Tree Regression, Random Forest Regression, Gradient Boosting Regression, and KNN. Following model training on a dedicated dataset and the subsequent prediction generation, rigorous accuracy testing against real-world data highlighted the superior performance of Gradient Boosting and Random Forest algorithms. The project's progression toward Heroku deployment via MLOps technologies involved establishing automated data pipelines, model versioning, CI/CD pipelines for testing and deployment, containerization, Heroku configuration, scalability, security, and user interface development. Continuous monitoring, optimization, and documentation completion ensured a successful transition from development to a production-ready insurance premium prediction service on Heroku..

Key Words: Linear Regression, Decision Tree Regression, Random Forest Regression, Gradient Boosting Regression, KNN, MLOPS, CI/CD, Heroku.

1.INTRODUCTION

The insurance industry is currently experiencing a significant transformation in its approach to premium estimation, thanks to the integration of machine learning and predictive modeling techniques. This research paper is a culmination of the collective efforts of researchers who have explored various machine learning algorithms and statistical methods to enhance the accuracy and fairness of insurance premium predictions. These advancements in premium estimation are paramount as they directly impact the accessibility and affordability of insurance policies for individuals and businesses alike. Over the years, numerous studies have contributed significantly to this evolving field.

In their 2013 study titled "Predicting Insurance Premiums Using Machine Learning Techniques," Dedek and Fashoyin delve into the application of machine learning models such as regression, decision trees, and neural networks for insurance premium prediction. Their work not only evaluates the comparative performance of these models but also sheds light on their practical implications, offering valuable insights to the industry.

Pradhan and Guru (2016) focus on health insurance premium prediction using machine learning algorithms in

their paper titled "Health Insurance Premium Prediction Using Machine Learning Algorithms." Their research not only explores data preprocessing techniques but also rigorously evaluates the performance of models like Random Forest and Support Vector Machines, striving to improve the precision of premium predictions in the context of health insurance.

In the paper titled "A Study of Predictive Modeling Techniques for Insurance Premiums" by Kadam and Patil (2013), the emphasis is on examining predictive modeling techniques, including linear regression and decision trees, for the estimation of insurance premiums. This study goes beyond model selection to address critical aspects such as feature selection, data preprocessing, and comprehensive model evaluation, all essential components in the context of insurance premium prediction.

Chidambaram and Balasubramanian (2013) conduct a comparative study in their research paper titled "Predicting Motor Insurance Premiums: A Comparative Study." Their research rigorously assesses the accuracy and efficiency of various machine learning and statistical models, including Random Forest and Gradient Boosting, for predicting motor insurance premiums, providing valuable insights for the motor insurance sector.

Finally, Samanta and Maity (2017) explore the application of ensemble learning techniques in "Health Insurance Premium Prediction Using Ensemble Learning." Their research investigates the synergistic effects of combining multiple models to enhance predictive accuracy in the context of health insurance premiums.

Together, these studies have significantly advanced the field of insurance premium estimation through the application of machine learning techniques. This research paper builds upon these findings, striving to provide a comprehensive understanding of the state-of-the-art methodologies and their practical implications in the dynamic landscape of the insurance industry. Beyond model performance, this paper delves into the practical considerations of deploying such models on platforms like Heroku using MLOps technologies, emphasizing data preprocessing, scalability, and ethical implications. This research aims to contribute to the ongoing discourse on achieving accurate and equitable insurance premium estimation, offering valuable insights for industry professionals and data scientists seeking to navigate this evolving landscape.

2. METHODOLOGY

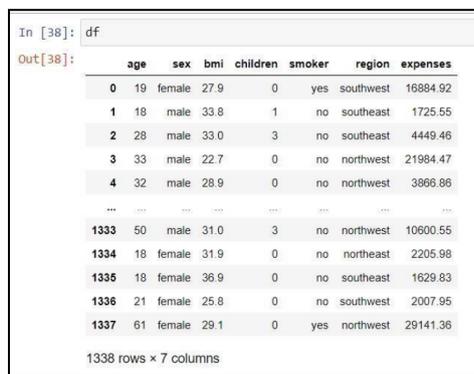
2.1 Data Preprocessing and Cleaning

The primary source of data for this project was Kaggle, provided by user Dmarco. The dataset comprises 1,338 records and includes six attributes: 'age,' 'gender,' 'bmi' (Body Mass Index), 'children,' 'smoker,' and 'charges.' These attributes collectively form the basis for predicting insurance charges, a key task in the insurance domain.

To prepare the dataset for regression analysis, several critical data preprocessing steps were undertaken:

2.1.1 Structured Format and CSV File:

The data was originally structured and stored in a CSV file, facilitating easy access and manipulation for analysis.



```
In [38]: df
Out[38]:
```

	age	sex	bmi	children	smoker	region	expenses
0	19	female	27.9	0	yes	southwest	16884.92
1	18	male	33.8	1	no	southeast	1725.55
2	28	male	33.0	3	no	southeast	4449.46
3	33	male	22.7	0	no	northwest	21984.47
4	32	male	28.9	0	no	northwest	3866.86
...
1333	50	male	31.0	3	no	northwest	10600.55
1334	18	female	31.9	0	no	northeast	2205.98
1335	18	female	36.9	0	no	southeast	1629.83
1336	21	female	25.8	0	no	southwest	2007.95
1337	61	female	29.1	0	yes	northwest	29141.36

1338 rows x 7 columns

Fig. 1. Sample of Dataset

2.1.2 Attribute Selection:

Not all attributes in the dataset are equally relevant for predicting insurance charges. Some attributes may even negatively impact prediction accuracy. Therefore, a thoughtful attribute selection process was implemented to retain only the most influential attributes. In this dataset, 'age' and 'smoker' were found to have the most significant impact on charge prediction, while 'children' had minimal effect. Consequently, 'children' was removed from the input to the regression model, contributing to improved computational efficiency and reduced processing time.

2.1.3 Impactful Factors:

It's important to note that in health insurance, numerous factors can influence insurance charges, including pre-existing health conditions, family medical history, marital status, location, and past insurance history. However, based on the dataset and analysis, 'age' and 'smoking status' emerged as the attributes with the most substantial impact on charge prediction, with 'smoker' being the most influential.

By meticulously cleaning and preprocessing the dataset, we aimed to enhance the suitability of the data for regression analysis. This process not only improves prediction accuracy but also enhances the overall performance and speed of the regression models employed.

In summary, this project utilized a dataset from Kaggle, meticulously cleaned and preprocessed it, and selected the most impactful attributes ('age' and 'smoker') for predicting insurance charges. These steps were crucial in optimizing the dataset for regression analysis and improving the accuracy of the predictions, aligning with the objectives of the project in the context of health insurance charge estimation.

2.2 Data Validation And EDA:

2.2.1 Data validation

Data validation is an essential component of the machine learning pipeline, focusing on assessing the quality and integrity of source data before model training. This involves checking for statistical properties like feature distributions and addressing issues such as zero standard deviation and complete missing values, as these can hinder model effectiveness. Data cleaning and transformation are performed as needed to prepare the data for modeling. Regular monitoring and documentation of data quality ensure ongoing reliability, contributing to accurate and dependable machine learning models.

2.2.2 EDA (Exploratory Data Analysis):

Visualized the relationship between the dependent and independent features. Also checked relationship between independent features to get more insights about the data.

2.3 Feature Engineering:

Following data preprocessing, the next step involves standard scaling to normalize all numeric features, ensuring they are on a consistent scale. Additionally, we employ one-hot encoding to transform categorical features into numerical representations. To streamline these operations and maintain data integrity, a pipeline is established, incorporating the standard scaling of numerical features and the encoding of categorical features in a systematic and automated manner. This pipeline is designed to enhance the efficiency and reproducibility of the data preparation process, facilitating the subsequent stages of our project.

2.4 Model Training & Saving:

2.4.1 Model Building

Following the comprehensive pre-processing procedures detailed earlier, including scaling and encoding, the prepared dataset is seamlessly channeled through a pipeline to evaluate its performance across various machine learning

models. Specifically, we assess the efficacy of Linear Regression, Decision Trees, Random Forest, and Gradient Boosting.

Our experimentation reveals that Gradient Boosting stands out as the top-performing model when applied to the test data. Building upon this success, we further enhance the model's performance through a process known as Gradient SearchCV. This involves an exhaustive search for the optimal hyperparameters within the Gradient Boosting algorithm. Through the application of Gradient SearchCV on the Gradient Boosting model, we fine-tune its parameters, leading to a notable improvement in its predictive accuracy and overall performance. This iterative process allows us to extract the maximum predictive power from our model, ensuring that it is finely tuned to deliver the best results for our specific dataset and problem domain.

2.4.2 Model Saving

Model is saved using dill library in .pkl format.

2.5 Model Evaluation

After constructing and saving our machine learning model, we rigorously evaluate its performance using multiple key metrics. These metrics include the Root Mean Squared Error (RMSE) to gauge predictive accuracy and the R-squared (R2) score to measure the variance explained by the model. Additionally, we set a classification accuracy threshold of 80 percent, relevant for tasks involving policy categorization based on premiums. This comprehensive evaluation approach ensures that our model not only excels in regression accuracy but also meets the specified classification threshold, signifying its effectiveness in estimating insurance premiums and making accurate policy categorizations.



Run ID	Model	RMSE	R2	Classification Accuracy
7	2e4f2c08-34050eed-c57	0.7578	0.78567	True
8	560d4368-59e3-498e-b4b0-4c9662031c9a	NaN	NaN	has been started
9	560d4368-59e3-498e-b4b0-4c9662031c9a	0.7578	0.78567	True

Fig. 2. Evaluation Matrix

2.6 Flask Setup for Web Application.

Subsequent to the model's successful development and validation, we embarked on the crucial phase of building an API using Flask. This step marked the transformation of our machine learning model into a practical and accessible tool for end-users. Within the Flask framework, we crafted a web application to facilitate testing and interaction with the model. The web application we created serves as a user-friendly interface where individuals can input their data. Once a user submits their data through the web interface, our Flask application takes charge of extracting this data. This extracted information is then seamlessly passed to our trained machine learning model.

The primary purpose of this stage is to harness the predictive capabilities of our model in real-time. Specifically, the model utilizes the input data to estimate insurance premiums. Users can now obtain personalized premium estimates quickly and conveniently through this user-friendly web application.

In essence, our Flask-based web application bridges the gap between the complex machine learning model and end-users, making insurance premium estimation a user-driven and accessible process.

2.7 Deployment

After thoroughly evaluating and fine-tuning our project, the next crucial step in the deployment process is to make it accessible to users. We do this by leveraging version control and cloud hosting platforms.

First, we utilize GitHub as our version control system to maintain a well-organized and collaborative development environment. GitHub allows us to securely store and track changes to our project's source code, making it easy for multiple team members to collaborate, contribute, and maintain a history of the project's development.

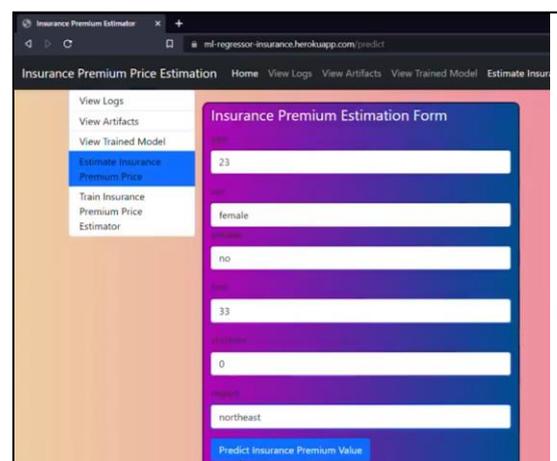


Fig. 3. Screenshot of Deployed Project

Once our project is version-controlled on GitHub, the next step is deployment. For this, we turn to Heroku, a popular cloud platform that simplifies the deployment and hosting of web applications. Heroku enables us to deploy our machine learning model, web application, and associated files seamlessly. The process involves configuring the application's environment, dependencies, and settings to ensure it runs smoothly in a production environment.

Heroku offers scalability, load balancing, and automatic server management, making it a robust choice for deploying applications. Moreover, Heroku's seamless integration with GitHub allows us to set up automatic deployments, ensuring that any changes pushed to our GitHub repository trigger automatic updates to our deployed application on Heroku.

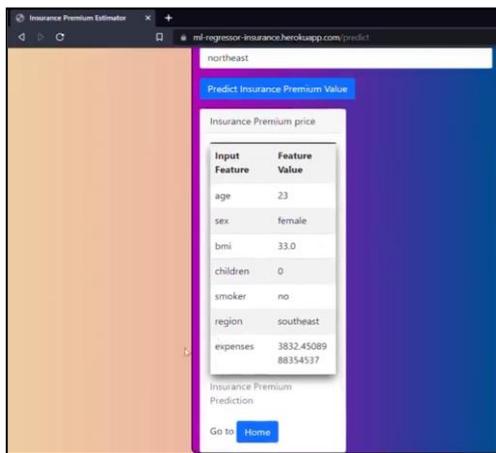


Fig. 4. Prediction

This deployment process not only makes our project accessible to users but also facilitates continuous integration and delivery (CI/CD), ensuring that our application remains up to date and reliable. Users can now interact with our insurance premium estimation tool via the web, providing a user-friendly and accessible solution to their needs.

3. CONCLUSIONS

In this project, we used a Kaggle dataset with 1,338 records and six attributes to predict insurance charges. We performed rigorous data preprocessing, including attribute selection and feature engineering. Our top-performing model was Gradient Boosting, fine-tuned using Gradient SearchCV.

We saved the model and evaluated it using RMSE, R2 score, and an 80% accuracy threshold for classification. We then built a user-friendly web application in Flask for real-time premium estimates.

The project was deployed on Heroku, integrating GitHub for version control. This comprehensive workflow demonstrates the value of data-driven decision-making in insurance premium estimation, with practical applications in the

industry for accurate pricing and enhanced customer experience.

REFERENCES

- [1] S. Samanta, M. Maity, "Health Insurance Premium Prediction Using Ensemble Learning," Published in: 2017 International Conference on Recent Advances in Computer Systems (2017).
- [2] V. Pradhan, S. R. Guru, "Health Insurance Premium Prediction Using Machine Learning Algorithms," International Journal of Innovative Research in Computer and Communication Engineering (2016).
- [3] S. Kadam, D. Patil, "A Study of Predictive Modeling Techniques for Insurance Premiums," International Journal of Computer Applications (2013).
- [4] J. Chidambaram, R. Balasubramanian, "Predicting Motor Insurance Premiums: A Comparative Study," International Journal of Computer Applications (2013).
- [5] T.A. Dedeke, T. Fashoyin, "Predicting Insurance Premiums Using Machine Learning Techniques," International Journal of Computer Applications (2013).