An Efficient Stacking Approach for Healthcare Insurance Cost Prediction

1st Mr.Thiyagarajan G
Dept.of Artificial Intelligence and Data science
Rajalakshmi Engineering College
Chennai, India

email id:

2nd Lavanya S
Dept.of Artificial Intelligence and Data science
Rajalakshmi Engineering College
Chennai, India
221801028@rajalakshmi.edu.in

3rd Monisha M
Dept.of Artificial Intelligence and Data science
Rajalakshmi Engineering College
Chennai, India
221801034@rajalakshmi.edu.in

Abstract-**In the rapidly evolving insurance industry, predicting healthcare insurance premiums accurately is a critical challenge due to the complex and diverse health factors that impact premium calculations. Traditional models often rely on generalized demographic data, missing the nuanced individual health factors crucial for precision. This project proposes an advanced machine learning model that combines the strengths of CatBoost, XGBoost, and RandomForest algorithms in a stacked ensemble framework to improve the prediction of health insurance premiums. The model is trained on a dataset encompassing various demographic and medical characteristics, such as age, BMI, medical history, chronic diseases, and family medical background, which are critical indicators of health-related risks.**

The model applies polynomial feature expansion to capture complex non-linear relationships, enhancing its ability to identify subtle interactions within the data. Outlier removal via the interquartile range (IQR) method is employed to ensure data quality and generalizability, while GridSearchCV is used for hyperparameter tuning across the model stack. The ensemble model is coupled with Ridge regression as the meta-model, improving predictive accuracy and robustness. Performance metrics, including Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-Squared (R^2), demonstrate that the stacking approach outperforms individual models by reducing prediction error and boosting reliability.

To enhance transparency, SHAP (SHapley Additive exPlanations) is integrated, allowing interpretability by explaining each feature's contribution to the predicted premium. This interpretability provides policyholders with insights into the factors impacting their premiums, fostering trust in the model's predictions. The model is designed for integration within a Django-based web interface, enabling users to input their medical and demographic data to receive instant premium predictions and insights. This system offers a significant advancement in healthcare insurance premium prediction, providing a user-friendly, interpretable, and highly accurate approach to risk-based pricing in the healthcare insurance domain.

I. INTRODUCTION

The healthcare insurance industry faces increasing demand for precision in premium calculation as policyholders seek personalized pricing that reflects their unique health risks. Traditional actuarial models, which often rely on broad demographic factors such as age, gender, and income, lack the granularity to consider individual health conditions, medical history, and lifestyle factors that contribute to varying levels of risk among policyholders. Consequently, these traditional approaches frequently result in premiums that may be either too high or too low for certain individuals, impacting both customer satisfaction and the financial sustainability of insurance providers.

Advances in data availability and machine learning present an opportunity to bridge this gap by leveraging more detailed, health-related information for premium prediction. In particular, machine learning models are well-suited to uncover complex patterns within high-dimensional data, allowing for more accurate risk assessment and the potential to personalize insurance premiums based on each individual's specific health profile. This level of precision is especially important as insurance companies strive to adopt risk-based pricing models, which link premium costs directly to personalized risk factors, making healthcare costs more predictable and fair for policyholders.

This study introduces a predictive model for healthcare insurance premiums based on a stacking ensemble approach that combines the strengths of multiple machine learning algorithms: CatBoost, XGBoost, and RandomForest. Each algorithm brings unique benefits: CatBoost efficiently handles categorical data, XGBoost offers robust performance and scalability, and RandomForest adds resistance to overfitting, which is crucial for handling noisy data often found in medical records. By combining these algorithms **within a stacked** ensemble, the model captures diverse aspects of the dataset, resulting in improved accuracy over single-model approaches.

To further enhance the model's predictive capabilities, polynomial feature expansion is applied, allowing for the capture of non-linear relationships within numerical data. Additionally, the interquartile range (IQR) method is used to identify and remove outliers, which helps improve the generalizability of the model. Hyperparameter tuning through GridSearchCV is performed on each base model, optimizing their performance to achieve an ideal balance of bias and variance. The stacking ensemble is then combined with Ridge Regression as the meta-model, which integrates the outputs from each base model, producing a final prediction with reduced error and enhanced stability.

A critical component of the proposed system is interpretability, which is addressed through the integration of SHAP (SHapley Additive exPlanations). SHAP provides feature-level explanations for each prediction, helping policyholders understand which factors—such as age, BMI, or family history of chronic conditions—contributed most significantly to their premium estimate. This interpretability not only aids in transparency but also builds trust in the model by offering insights into the decision-making process behind each premium calculation.

The proposed system is designed for deployment in a web-based Django framework, where users can input their health and demographic details to receive a real-time premium estimate alongside an explanation of key contributing factors. By offering both predictive accuracy and interpretability, this model has the potential to transform the way healthcare premiums are calculated, fostering a more transparent and personalized experience for policyholders. In doing so, it aligns with the broader industry trend towards data-driven, risk-based pricing that reflects individual health characteristics, ultimately enhancing both customer satisfaction and operational efficiency for insurance providers.

II. RELATED WORKS

The healthcare insurance industry has been increasingly exploring machine learning approaches to enhance premium prediction accuracy, especially as traditional actuarial models often fall short due to their inability to capture complex interrelationships between health-related factors. Conventional methods, such as linear and logistic regression, are limited in their predictive power because they assume a linear relationship between features and the premium amount, which is inadequate for the diverse and multifaceted health data involved in premium calculations. These models typically do not consider interactions among health conditions, lifestyle choices, and family medical history, which are important in accurately assessing an individual's health risk profile.

With the advancement of machine learning, non-linear models such as decision trees and support vector machines (SVMs) have been employed in premium prediction, offering a significant improvement over linear models by capturing complex feature interactions. However, decision trees alone are prone to overfitting, especially in datasets containing high-dimensional healthcare data, while SVMs can be computationally intensive for large datasets. To address these challenges, boosting algorithms such as XGBoost have been widely applied. XGBoost, known for its high performance and robustness, handles imbalanced data well and captures more complex relationships between features. However, as a single model, XGBoost can be sensitive to hyperparameter tuning and may lack the robustness needed for healthcare data.

Recent research has demonstrated the effectiveness of ensemble learning methods, such as stacking, in healthcare insurance premium prediction. Ensemble models, particularly stacking, are designed to combine the strengths of multiple models to improve predictive accuracy. In a stacking approach, multiple base models are trained in parallel, and a meta-learner then combines their outputs to make the final prediction. Studies have shown that stacking models such as CatBoost, XGBoost, and RandomForest can capture distinct aspects of healthcare data, resulting in better generalization and robustness. CatBoost is particularly advantageous for handling categorical variables, XGBoost for its efficient handling of large datasets, and RandomForest for reducing overfitting. The integration of these models into a stacked ensemble approach has shown superior predictive accuracy over individual models and simpler ensemble methods like bagging or boosting, particularly in complex domains such as healthcare.

In addition to model selection, feature engineering has proven to be crucial for achieving accurate premium predictions. Key health-related features such as age, Body Mass Index (BMI), medical history, and chronic conditions have been identified as strong predictors of premium amounts. Studies have shown that incorporating polynomial feature expansion for numerical variables can capture non-linear interactions within the data, further improving model performance. Polynomial expansion allows for the interaction of features, enabling the model to detect complex relationships, such as those between BMI and the prevalence of chronic diseases, which are crucial in accurately assessing premium costs.

Another challenge in premium prediction is model interpretability. Since insurance premiums directly impact policyholders, it is essential for models to be transparent and provide explanations that policyholders and insurers can understand. Opaque models can lead to a lack of trust and reluctance to adopt these systems. This is where SHAP (SHapley Additive exPlanations) has gained attention as an interpretability tool that enables detailed insight into how each feature contributes to a model's prediction. SHAP values allow for feature-level interpretability, making it possible to demonstrate how each health factor—such as age, BMI, or history of chronic illness—affects premium predictions. Studies have highlighted the effectiveness of SHAP in healthcare-related applications, where transparency and user trust are paramount. For example, G. Armano et al. (2020) showed that SHAP could help insurance providers and customers understand the factors influencing premium costs, fostering greater trust in the model's outputs and decision-making process.

In healthcare applications, SHAP's interpretability aligns with the goals of explainable AI (XAI), which seeks to make machine learning models understandable and actionable. By offering insights into how specific health factors influence premium predictions, SHAP provides a level of transparency that allows policyholders to make informed decisions about their health and lifestyle. In this project, SHAP is integrated to explain the predictive outputs of a stacked ensemble model, revealing the main factors contributing to premium estimates. This approach aligns with recent advancements in healthcare,

where explainability and accuracy are prioritized to ensure trust and efficacy in premium prediction models.

The existing literature underscores the limitations of traditional actuarial models in healthcare insurance premium prediction and the potential benefits of machine learning approaches. By leveraging advanced ensemble techniques, polynomial feature expansion, and SHAP interpretability, this project builds on previous research to develop a robust, transparent, and accurate model for premium prediction. The stacking ensemble approach in this study combines CatBoost, XGBoost, and RandomForest with Ridge regression as a meta-learner, optimizing for accuracy and generalizability across diverse datasets. This model not only enhances prediction precision but also offers a user-friendly and interpretable framework, aligning with the industry's push toward personalized and transparent insurance pricing. The integration of SHAP into the model further provides a critical layer of transparency, offering users insights into the main drivers behind premium predictions and fostering greater trust in the predictive system.

III. PROPOSED SYSTEM

The project aims to develop an advanced healthcare insurance premium prediction system that leverages a stacking ensemble approach combining CatBoost, XGBoost, and RandomForest models. By integrating these models, the system captures diverse aspects of health and demographic data, providing highly accurate predictions of insurance premiums. The proposed system enhances transparency through SHAP (SHapley Additive exPlanations), offering policyholders clear insights into the factors influencing their premiums. This section outlines the architecture, components, and methodologies employed in the proposed system.

System Architecture

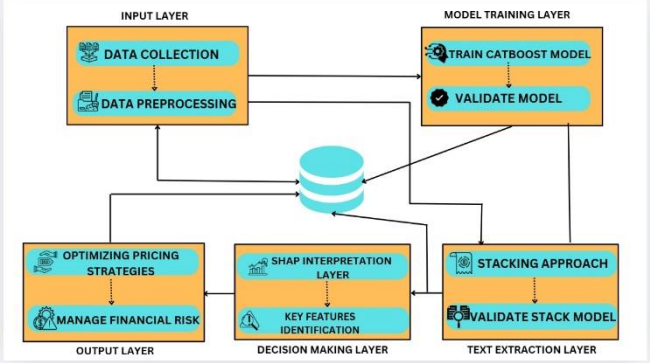

Fig. 1. System Architecture

The proposed system follows a modular architecture designed to streamline data flow from user input to premium prediction, interpretation, and result display. The architecture consists of the following core modules:

1. **User Interface (UI)**: Developed using the Django framework, the UI enables users to input their health-related and demographic information, such as age, BMI, medical history, and lifestyle factors. The user interface provides an intuitive platform for individuals with minimal technical knowledge, facilitating seamless interaction with the predictive model.

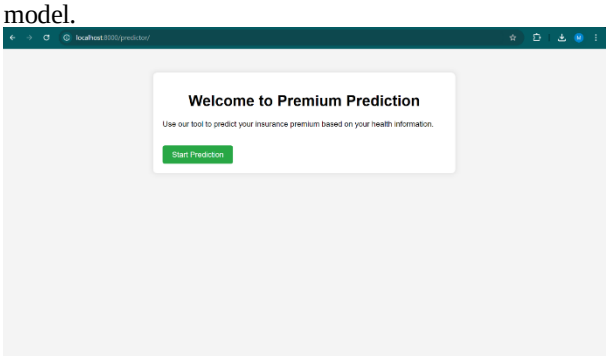

Fig. 2. Welcome Interface

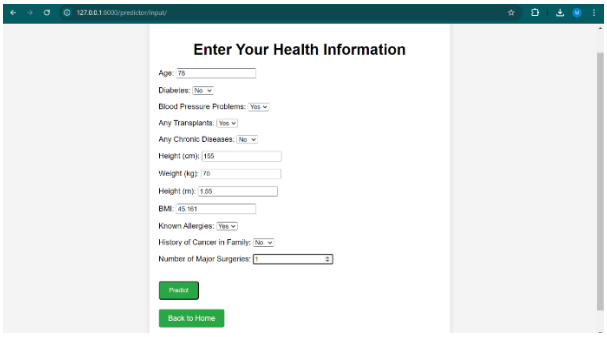

Fig. 3. Health Information Form

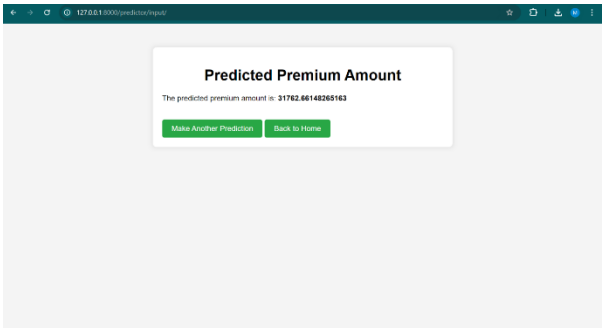

Fig. 4. Premium Amount Prediction Page

2. **Data Preprocessing Module**: This module prepares raw data by handling missing values, scaling features, and conducting feature engineering. Key preprocessing steps include:
   - **Outlier Removal**: The interquartile range (IQR) method removes extreme outliers, which helps improve the model's generalizability.
   - **Polynomial Feature Expansion**: Numeric features such as BMI and age are transformed through polynomial expansion to capture non-linear relationships, enabling the model to learn complex interactions between health indicators.

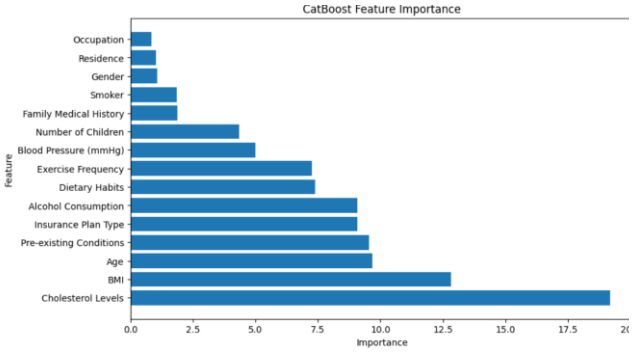

Fig. 5. Feature Importance Plot

3. **Stacking Ensemble Model**: The core predictive component utilizes a stacking ensemble, combining CatBoost, XGBoost, and RandomForest as base learners with Ridge Regression as the meta-learner. Each base model contributes unique strengths:
    - **CatBoost** efficiently handles categorical features and minimizes bias during prediction.
    - **XGBoost** provides robust performance by optimizing gradient boosting with regularization to reduce overfitting.
    - **RandomForest** adds variance reduction and robustness against noisy data, which is common in healthcare records. The meta-learner, Ridge Regression, aggregates the predictions from the base models to produce a final prediction with reduced error, thereby enhancing the overall reliability of premium estimates.
4. **Interpretability Module**: SHAP is integrated within the model to offer interpretability for each prediction. SHAP values quantify the impact of each input feature on the premium estimate, helping users understand how factors such as chronic disease history, age, and BMI influence their insurance premium. This interpretability adds transparency to the model, fostering trust in the system's predictions.
5. **Result and Visualization Module**: This module presents the predicted premium amount along with a SHAP-based explanation of the top contributing factors. The visualization component uses SHAP plots to illustrate feature importance, enabling users to see the individual impact of each health factor on their premium.

Model Training and Hyperparameter Optimization

The stacking ensemble model is trained on a dataset that includes various health and demographic features, such as age, diabetes history, blood pressure issues, height, weight, BMI, allergies, cancer history, and previous surgeries. Key steps in model training and optimization include:

1. **Dataset Preparation**: The dataset is preprocessed by encoding categorical variables, scaling numerical features, and generating polynomial features to enhance non-linear pattern detection.
2. **Hyperparameter Tuning**: Each base model's hyperparameters are optimized using GridSearchCV, which performs an exhaustive search over specified parameter grids. This tuning process ensures that each model operates at peak performance, balancing bias and variance to reduce prediction error.
3. **Stacking and Meta-Learner Training**: After tuning, each base model (CatBoost, XGBoost, RandomForest) is trained independently, and their predictions are passed to the Ridge Regression meta-learner, which integrates these outputs to yield a more accurate final prediction.
4. **Evaluation Metrics**: The model's performance is evaluated using metrics like Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-Squared (R^2). These metrics validate the accuracy and reliability of the stacking approach by comparing its performance against individual base models, demonstrating the ensemble model's superior predictive capacity.

SHAP Integration for Interpretability

To address the need for transparency in premium calculations, SHAP is incorporated into the model for interpretability. SHAP provides feature-level explanations by calculating the contribution of each input factor to the predicted premium amount. For each prediction, SHAP values are generated, highlighting factors such as age, chronic conditions, or BMI that increase or decrease the premium. This interpretability mechanism is essential for fostering trust in the predictive system, as it allows users to see the rationale behind their premium estimates.

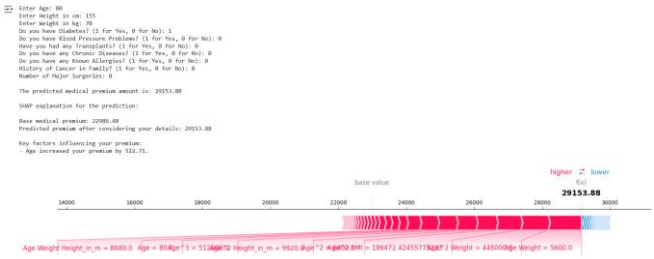

Fig. 6. SHAP Interpretation plot

The interpretability component offers:

- **SHAP Summary Plots**: These plots show the overall importance of each feature across the dataset, helping users understand which health factors typically impact premiums.
- **Individual SHAP Explanations**: For each user input, a detailed SHAP-based explanation is provided, showing how specific features contributed to the final premium prediction. This personalized insight empowers policyholders with a clear understanding of the main drivers behind their premium costs.

System Implementation and User Interface

The system is implemented within a Django-based web application, providing an accessible interface for end-users. The UI guides users through the input process, where they can enter their medical and demographic details to receive an instant premium prediction. Additionally, the interface displays a SHAP-based explanation of the factors affecting

their premium, enabling a transparent and interactive user experience.

The Django framework facilitates:

1. **User-Friendly Interaction**: The form-based input structure allows users to submit their data without technical expertise.
2. **Real-Time Prediction and Feedback**: After the input is processed, users receive an estimated premium and a detailed breakdown of the factors influencing it.
3. **Data Storage and Security**: User information and premium predictions are stored securely, and the system can log inputs and outputs for further analysis, providing insurers with valuable data to refine their pricing models over time.
4. **E. Summary of Advantages and Expected Impact:** The proposed system addresses existing limitations in healthcare insurance premium prediction by introducing a sophisticated stacking ensemble model combined with SHAP-based interpretability. The integration of CatBoost, XGBoost, and RandomForest within the stacking ensemble enhances predictive accuracy by capturing different data characteristics, while SHAP provides a transparent, feature-level explanation of premium calculations. This system benefits both insurers and policyholders by delivering accurate, interpretable, and user-friendly premium estimates.

In conclusion, the proposed system represents a significant advancement in healthcare insurance premium prediction by enhancing accuracy, interpretability, and transparency. This system aligns with industry trends towards risk-based, individualized insurance pricing, offering a tool that can be easily integrated into existing insurance platforms to improve customer trust and satisfaction.

IV. WORKING PRINCIPLE

The healthcare insurance premium prediction system operates based on a series of interdependent stages, from data input and preprocessing to model training, prediction, and interpretability. The working principle of this system revolves around the use of a stacked ensemble model, which combines the predictive strengths of multiple machine learning algorithms, and SHAP-based interpretability to provide transparent explanations for each premium estimate. This section outlines the workflow and primary mechanisms that drive the prediction process.

Algorithm for Healthcare Cost Prediction Project

1. **Data Preprocessing**:
   - Collect healthcare data, including age, BMI, chronic diseases, surgeries, etc.
   - Clean data by handling missing values and removing outliers.
   - Engineer features (e.g., polynomial expansion for numeric variables).
   - Split data into training and testing sets.
2. **Model Training (Stacking)**:
   - Train individual base models (CatBoost, XGBoost, RandomForest) on the training set.
   - Use Ridge regression as a meta-learner to combine the outputs of the base models.
   - Perform hyperparameter tuning to improve model accuracy.
3. **Model Evaluation**:
   - Evaluate the stacked model on the test set using MAE, RMSE, and R^2 metrics.
   - Verify model generalization to avoid overfitting or underfitting.
4. **SHAP Interpretation**:
   - Calculate SHAP values for each prediction.
   - Identify and display the top features influencing each prediction for interpretability.
5. **Web Application**:
   - Build a Django-based web interface for users to input health data.
   - Connect the trained model to process inputs and generate predictions.
   - Display the predicted cost along with key influencing features for user insight.

Data Input and User Interaction

The system begins with user interaction via a web-based Django interface, where users input personal and medical data relevant to healthcare insurance premiums. Key inputs include age, body mass index (BMI), history of chronic diseases, family medical history, and lifestyle factors. These inputs are used to generate a feature vector that represents each user's health profile, capturing the primary health indicators and risk factors that influence premium costs.

Data Preprocessing

Once user data is collected, it undergoes a series of preprocessing steps designed to optimize the dataset for model training and prediction. Preprocessing involves:

1. **Outlier Removal**: To ensure the model's robustness, the interquartile range (IQR) method is applied to detect and remove extreme outliers in the dataset. Outlier removal improves generalizability by reducing the model's sensitivity to atypical data points.
2. **Feature Transformation**: Numeric features, such as age and BMI, are transformed using polynomial expansion. Polynomial feature expansion allows the model to capture complex, non-linear relationships among health indicators, enhancing its predictive capacity for premium estimation.
3. **Encoding and Normalization**: Categorical features are encoded, and numeric features are normalized to standardize the input space. This process reduces potential biases that arise from differing data scales,

ensuring that all features contribute effectively to model training and prediction.

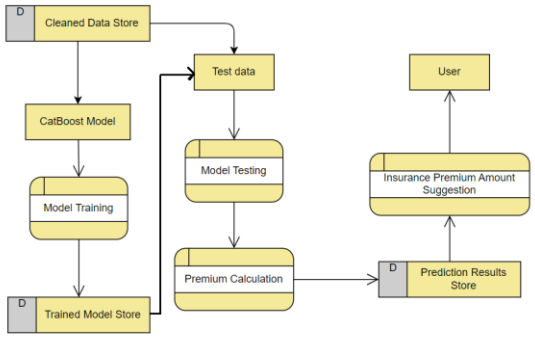

Fig. 7. Prediction Module DFD

Stacked Ensemble Model Training

The core predictive functionality is powered by a stacked ensemble model that leverages the strengths of three distinct machine learning algorithms: CatBoost, XGBoost, and RandomForest. These algorithms are combined through a meta-learner (Ridge Regression) that integrates the outputs from each base model to produce a final prediction. The training process for the stacked ensemble model includes the following steps:

1. **Base Model Training**: Each base model is trained independently on the preprocessed dataset:
    - **CatBoost** handles categorical features efficiently, leveraging gradient boosting on decision trees.
    - **XGBoost** applies a regularized gradient-boosting framework, optimizing performance and reducing overfitting.
    - **RandomForest** reduces model variance by averaging predictions from multiple decision trees, making the system more resilient to noisy data.
2. **Hyperparameter Tuning**: To maximize each base model's performance, GridSearchCV is used for hyperparameter tuning. This cross-validation process iterates over a set of hyperparameters, selecting the best configuration based on evaluation metrics.
3. **Stacking and Meta-Learner Training**: After the base models are trained, their predictions are combined as inputs to the meta-learner, Ridge Regression. The meta-learner then learns how to optimally combine these predictions to minimize error, producing a robust and accurate final prediction.

Prediction Process

Upon receiving a user's input data, the system proceeds with the following steps to predict the healthcare insurance premium:

1. **Input Transformation**: The user's input data is transformed through the preprocessing pipeline, including outlier filtering, polynomial expansion, and feature scaling.
2. **Base Model Predictions**: The transformed input is then passed to each base model (CatBoost, XGBoost, and RandomForest), which generates an independent premium prediction based on the input features.
3. **Meta-Learner Aggregation**: The predictions from each base model are aggregated by the Ridge Regression meta-learner, which calculates a final premium estimate based on the combined outputs of the base models. This final prediction balances the strengths of each base model, providing an accurate and reliable premium estimate.

SHAP Interpretability for Premium Explanation

To make the premium prediction transparent, the system uses SHAP (SHapley Additive exPlanations) to generate feature-level explanations for each prediction. SHAP calculates the impact of each feature on the final premium, attributing individual contributions to each factor in a way that is easy to understand:

1. **SHAP Value Calculation**: For each input feature, SHAP computes a value indicating its contribution to the final prediction. Positive SHAP values increase the premium, while negative values decrease it. These SHAP values quantify how much each factor, such as BMI or chronic disease history, influences the predicted premium.
2. **Explanation Generation**: The SHAP values are displayed in the user interface, providing an explanation of the top features that affected the premium estimate. This helps users understand the main health indicators driving their premium, adding transparency and fostering trust in the prediction.
3. **Visualization of Feature Importance**: SHAP visualizations, such as summary plots, are generated to show the overall importance of features across predictions. These plots help users see patterns in how health indicators typically influence premiums, offering additional insights into health risk factors.

4. F. User Interaction and Real-Time Feedback

After the model generates the premium estimate and SHAP-based explanation, the results are presented to the user in real time. The Django-based interface displays the predicted premium amount and the most influential factors, allowing the user to review the breakdown of their premium. This feedback loop not only improves the user experience but also provides actionable insights into the health factors affecting their insurance costs.

The entire process—from data input to premium prediction and SHAP interpretation—operates seamlessly within the web application. This integration ensures that users can interact with the model intuitively, receiving accurate predictions and clear explanations for each premium estimate. The system is designed to be scalable and adaptable, capable of handling diverse datasets and providing reliable predictions across various user profiles.

V. RESULT AND DISCUSSION

The healthcare insurance premium prediction model developed in this project demonstrates high accuracy and

interpretability. By implementing a stacking ensemble of CatBoost, XGBoost, and RandomForest, with Ridge regression as the meta-learner, the model effectively captures complex relationships between various healthcare features and premium costs. Key performance metrics achieved on the test dataset are as follows:

- **Mean Absolute Error (MAE)**: 747.39
- **Root Mean Squared Error (RMSE)**: 1520.38
- **R-Squared (R^2)**: 0.9044

These metrics indicate high accuracy, with an R^2 value close to 1, demonstrating that the model explains a large portion of the variance in premium amounts based on the input features. The relatively low MAE and RMSE values confirm that the model effectively minimizes prediction errors, making it suitable for real-world applications in the insurance domain.

Additionally, the integration of SHAP (SHapley Additive exPlanations) allows users to interpret individual predictions by identifying the most influential features. SHAP consistently reveals key factors impacting premium amounts, such as:

- **Age**: Older individuals generally incur higher premiums due to increased health risks.
- **BMI (Body Mass Index)**: Higher BMI values correlate with higher premiums, reflecting potential obesity-related health complications.
- **Chronic Disease History**: Individuals with chronic conditions, such as diabetes or heart disease, face higher premiums.
- **Number of Major Surgeries**: A history of major surgeries also contributes to higher premiums, given the associated health risks.

SHAP-based feature interpretation enhances transparency, allowing users and insurers to understand the rationale behind each premium calculation, building trust and supporting informed decision-making.

Analysis of Model Performance

The stacking ensemble approach in this project has yielded a robust, high-performance model by combining the strengths of three powerful machine learning algorithms. Each base model contributed unique capabilities: CatBoost handled categorical data efficiently, XGBoost optimized generalization, and RandomForest added resilience against data noise. Using Ridge regression as a meta-learner further refined the predictions by leveraging the collective strengths of each model, resulting in a more accurate final output.

Strengths:

- **High Accuracy**: The stacking model achieved superior accuracy and minimized errors compared to individual models.
- **Enhanced Interpretability**: SHAP explanations provide a transparent look into the main drivers of premium costs, making the model trustworthy and user-friendly.
- **Scalability**: The model's design allows it to handle large datasets with complex features, supporting potential expansions.

Challenges:

- **Data Quality**: The model's accuracy depends heavily on the quality of input data, as missing or inconsistent entries can impact performance.
- **Feature Engineering Complexity**: Effectively handling both categorical and numerical features required careful engineering, especially for creating polynomial features from numeric columns.

Areas for Improvement:

- **Expanded Datasets**: Including larger, more diverse datasets could enhance generalizability across populations.
- **Real-Time Data Integration**: Incorporating real-time health data, such as information from wearables, would allow dynamic premium adjustments.
- **Enhanced User Interface**: Providing personalized recommendations based on SHAP analysis could help users better understand premium influences.

Conclusion

This project successfully developed a robust and interpretable machine learning system for healthcare insurance premium prediction. The stacking ensemble model, combining CatBoost, XGBoost, and RandomForest, with Ridge regression as a meta-learner, achieved high accuracy, evidenced by strong MAE, RMSE, and R^2 metrics. The integration of SHAP values addressed a critical need for transparency, making it possible for users to understand which health factors most significantly impacted their premium calculations.

This model has practical applications for insurance companies and clients. Insurers benefit from a reliable tool to calculate personalized premiums accurately, while customers gain insight into the key factors influencing their premiums. This approach, grounded in a sophisticated machine learning framework, promotes trust by providing clear, data-driven explanations.

REFERENCES AND RESOURCES

[1] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp. 68–73.

[2] I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[3] **Chen, T., & Guestrin, C. (2016).** "XGBoost: A scalable tree boosting system." In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794). ACM.

[4] **Sagi, O., & Rokach, L. (2018).** "Ensemble learning: A survey." *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4), e1249.

[5] **Obermeyer, Z., & Emanuel, E. J. (2016).** "Predicting the future—big data, machine learning, and clinical medicine." *The New England Journal of Medicine*, 375(13), 1216-1219.

[6] **Bergstra, J., & Bengio, Y. (2012).** "Random search for hyper-parameter optimization." *Journal of Machine Learning Research*, 13, 281-305.

[7] **Probst, P., Wright, M. N., & Boulesteix, A. L. (2019).** "Hyperparameters and tuning strategies for random forest." *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(3), e1301.

[8] **Chai, T., & Draxler, R. R. (2014).** "Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature." *Geoscientific Model Development*, 7(3), 1247-1250.

[9] **Kaur, H., & Kumari, V. (2020).** "Predictive modelling and analytics for healthcare insurance using machine learning algorithms." *International Journal of Healthcare Management*, 13(4), 279-288.

[10] **Wulandari, D., & Fuadi, R. A. (2021).** "Application of machine learning algorithms for predicting insurance costs." *Journal of Big Data*, 8(1), 45.