Grenze International Journal of Engineering and Technology, January Issue

Evaluating Predictive Models for Precision in Insurance Pricing

Rajul Bafna[1] and Devendra Joshi[2]
[1-2]Department of Computer Engineering and Technology School of Computer Science and Engineering, Dr. Vishwanath Karad MIT World Peace University Pune, India
Email: 1032222791@mitwpu.edu.in, devendra.joshi1@mitwpu.edu.in

***Abstract*— In the insurance and healthcare sectors, accurately forecasting medical insurance premiums is essential. This study compares a number of regression models, to estimate insurance costs based on demographic and lifestyle data. We applied feature scaling to improve model consistency.**
Grid search was used to adjust the Light GBM hyperparameters, and Light GBM was used as the meta-learner in the stacking ensemble. Each model was evaluated using the coefficient of determination The Stacking Regressor demonstrated the efficiency of model ensembling by outperforming all other models with the highest coefficient of determination score. This Comparative approach highlights ensemble and hybrid approaches for accurate cost forecasting in insurance applications and clarifies the predictive capabilities of different algorithms.

***Index Terms*—Medical insurance, Cost prediction, Machine learning, Light GBM, Stacking regressor, KMeans clustering, Health analytics.**

I. INTRODUCTION

Medical insurance expenses has grown in significance for both policyholders and service providers in the contemporary healthcare and insurance industries. Predictive models are used by insurance companies for risk assessment, premium structure design, and financial planning management. On the other hand, policyholders benefit from fair rates that reflect their health and lifestyle choices.
The advancement of data-driven technology and machine learning techniques now enables more sophisticated and accurate models for these behaviours. The rising costs of healthcare have made medical insurance an important component for security for individuals.
Estimating the insurance premium accurately has become important, for both the insurance providers and for the customers to get fair prices.
The Stacking model uses the strength of various models to produce more accurate predictions, while LightGBM is utilized for efficient way of handling large datasets and complex interactions. Many past studies have implemented traditional techniques like Linear and Ridge Regression, but few have explored the ensemble models like Light GBM and Stacked Regressors.
These models give an improved accuracy by combining the strength of multiple algorithms. The goal is to identify the most effective model in terms of accuracy.

Grenze ID: 01.GIJET.12.1.76_1

II. LITERATURE REVIEW

Several studies have applied Machine learning models in cost analysis to clarify stratergies and support effective healthcare management.

Kaushik et al. (2022) [1]introduced a regression framework that matches demographic characteristics such as age, bmi, gender and smoking behaviors to predict insurance behaviors.

Shojaee et al. (2022) [2]used Decision Trees and ensemble methods to estimate hospital stay lengths for patient with hip fracture, which provided an indirect analysis of insurance expenses. This predicts our approach of using patient focused data and supervised modeling techniques.

In Their study, Bhatia et al. (2022) [3] focused on health insurance prediction using regression.Their conclusion was that ensemble techniques exceed conventional models.

Influence of preprocessing on model accuracy was highlighted by Goel and Chaudhary (2022) [4] who emphasized the importance of feature preprocessing in enhancing the model. This helped us supporting our application for methods such as log transformation, standard scaling, GridSearchCV.

Kharwal (2021) [5]presented an approach of machine learning for predicting insurance premiums.The impact of smoking was highlighted and also the premium expenses, helping in our decision to add this factor in both model training and feature engineering.

Predictive modelling and clustering are used to forecast insurance risk and premiums, according to studies by Boodhun and Jayabalan [6] Sailaja et al [7] and Omar et al. [8]. Accuracy and segmentation were enhanced by methods such as REP Tree , stacking regressors, and K-means clustering, which supported data-driven decision-making in insurance analytics.

Sharma and Patel (2022) [9] compared tree-based classifiers like Decision Tree, Random Forest, and XGBoost. Our decision to use similar models to improve underwriting decisions was influenced by their research, which highlighted how effective XGBoost is at increasing classification accuracy and interpretability through SHAP values.

A medical insurance cost prediction model was proposed by Sailaja et al. (2021) [10] utilising a Stacking Regressor that combines XGBoost, CatBoost, and Voting Regressors. Their study showed that ensemble models greatly increase the accuracy of insurance cost prediction, with an explained variance score of 0.89.

To improve the accuracy and robustness of predictive modelling, Kumar et al. (2024) [11] suggested an ensemble stacking model that combines LightGBM, PLSRegression, and ElasticNet Regressors. For better insurance premium prediction, Mehta et al. (2023) [12] created an ensemble regression method that combined Random Forest and Gradient Boosting models. Both studies demonstrate how integrating several algorithms greatly improves model performance, establishing ensemble learning as a dependable method for precise and effective insurance risk assessment.

III. DATASET DESCRIPTION

The study uses a dataset named insurance, which includes demographic and health related data about individuals and also contains the medical insurance fees they were charged.

The dataset comprises 1,338 records (before filtering), with each row representing an individual's insurance profile

TABLE I. DATASET DESCRIPTION

Attribute	Table Column Head		
	Table column subhead	Data Type	Notes
Age	Individual Age	Int	-
Sex	Individual Gender	Object	-
Bmi	Body Mass Index	Float	-
Children	Number of children	Int	-
Smoker	Smoker/non-Smoker	Object	Yes/No
Region	Geographic Region	Object	Categorical
Charges	Individual cost	Float	Target Variable

A. Summary Statistics

No significant missing value was detected

	age	bmi	children	charges
count	1338.000000	1338.000000	1338.000000	1338.000000
mean	39.207025	30.663397	1.094918	13270.422265
std	14.049960	6.098187	1.205493	12110.011237
min	18.000000	15.960000	0.000000	1121.873900
25%	27.000000	26.296250	0.000000	4740.287150
50%	39.000000	30.400000	1.000000	9382.033000
75%	51.000000	34.693750	2.000000	16639.912515
max	64.000000	53.130000	5.000000	63770.428010

Figure 1. Dataset Description

IV. PROPOSED METHODOLOGY

The proposed methodology predicts medical insurance costs by comparing regression models using the insurance.csv dataset. To normalise the target variable's (charges') distribution), the dataset is pre-processed by encoding categorical variables and using a logarithmic transformation. Standard Scaler is used for feature scaling. A Stacking Regressor, tuned Light GBM, Ridge Regression, and Linear Regression are among the models that are trained and assessed using R2 Score, MAE, and MSE. GridSearchCV is used to apply hyperparameter tweaking to Light GBM. The most accurate model for practical implementation is then highlighted by visualising each model's performance and using it to forecast insurance costs.

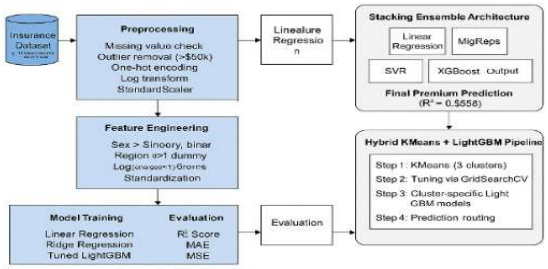

Figure 2. System Architecture

A. Exploratory Data Analysis and Pre-Processing

No significant missing data was detected

```
[ ]  # getting some informations about the dataset
     insurance_dataset.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1338 entries, 0 to 1337
Data columns (total 7 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   age       1338 non-null   int64
 1   sex       1338 non-null   object
 2   bmi       1338 non-null   float64
 3   children  1338 non-null   int64
 4   smoker    1338 non-null   object
 5   region    1338 non-null   object
 6   charges   1338 non-null   float64
dtypes: float64(2), int64(2), object(3)
memory usage: 73.3+ KB
```

Figure 3. Dataset Instance

B. Model Approach

- Linear Regression The most basic model is linear regression.It is assumed that the input features and the target variable have a linear relationship. Linear regresson acts as a baseline model to compare against more complex algorithms.
- Ridge Regression: This technique enhances Linear Regression by introducing L2 regularisation and preventing overfitting in the presence of multicollinearity or high-dimensional data.
- Light GBM (Light Gradient Boosting Machine): It works well on structured data and it performs well for outliers and nonlinear relationships, which makes it weel suited for identifying interactions in the dataset.
- Stacking Regressor: An ensemble technique called stacking combines several regression models to take advantage of each one's unique advantages. This is especially effective when models produce various kinds

of errors. In this project, the stacking model uses Linear Regression, Ridge Regression, SVR, XGBoost as base learners and Light GBM as the meta-model (final estimator).

- Kmeans Clustering + Tuned Light GBM: To improve prediction accuracy by first grouping similar data points (segmentation) using Kmeans clustering, and then applying a LightGBM regressor tuned with hyperparameters to each cluster independently. This hybrid approach accounts for heterogeneity in the dataset that a single global model might miss.

C. Key Insights from EDA

Distribution of age value

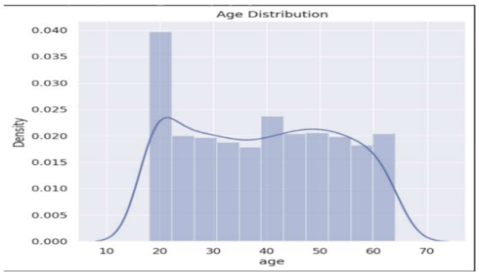

Figure 4. Distribution of age value

Gender Count

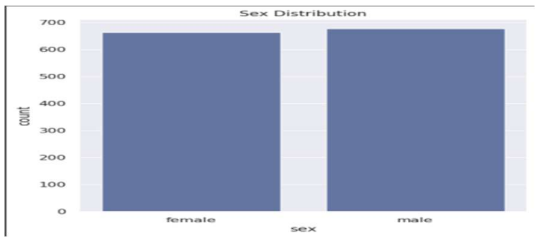

Figure 5. Gender Count

bmi vs age

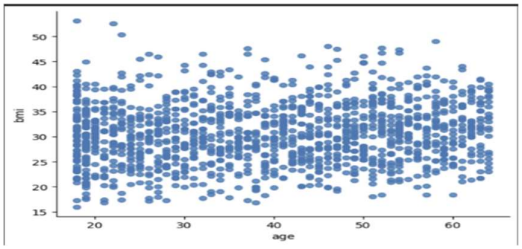

Figure 6. bmi vs age

Smoker Column

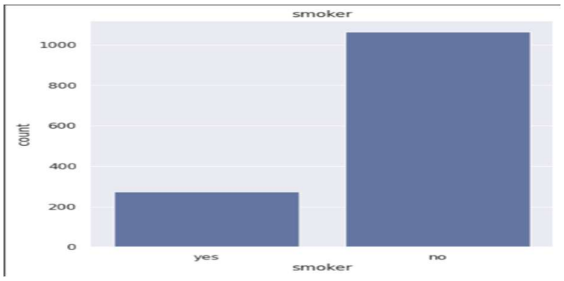

Figure 7. Smokeing count

D. Data preprocessing and Feature Engineering

The original dataset contained both numerical and categorical variables. Missing values were confirmed in the dataset, entries with values more than $50,000 were removed because their extreme nature could incline the model's learning so that to resolve outliers in the target variable charge. One-hot encoding was used to transform categorical numerical representation. This procedure removed the possibility of ordinal interpretation and produced binary columns that represented each category. Variance were stabilized using log transformation (np.log1p) This enhances model performance, particularly for linear models.
StandardScaler was used to standardize all numerical features using feature scaling.

V. Model Training and Evaluation

This section discusses the training methodologies, hyperparameter optimization techniques, and cross validation strategies to ensure a good model performance.

A. Hyperparameter Tuning Techniques

GridSearchCV for LightGBM:

- Parameters Tuned: num_leaves, learning_rate, n_estimators
- Cross-validation folds: 3
- Objective: Maximizing R^2 Score
- Resulting best model significantly improved accuracy over the default LightGBM model.

Optuna (Experimental Hybrid):

- Applied to tune LightGBM models individually for each KMeans cluster.
- Allowed exploration of high-dimensional hyperparameter space efficiently

B. GridSearchCV for LightGBM:

During hyperparameter optimisation, k-fold cross-validation (k=3) was used to validate all tuned models.

C. Inverse Transformation in Regression

Normalize the distribution of a skewed variable (like insurance charges).Model performance is improved especially for models assumed normally After model training and prediction, the predicted values are in the log-transformed space. To interpret the predictions in the original units (USD), you need to apply the inverse of log1p

VI. Results and Evaluation

Table II. Comparison of Model Performance (R^2 Score)

Model	R2 Score
Lasso Regression	0.7413
Linear Regression	0.7447
Ridge Regression	0.7835
Hybrid KMeans + LGBM	0.8782
Tuned LightGBM	0.8808
Stacking Regressor	0.8858

A. Interpretation

The baseline models, Lasso and Linear Regression, are limited to capturing linear relationships. They perform poorly on complex and nonlinear data patterns, as indicated by their moderate R2 values (≈0.74).
Ridge Regression: Multicollinearity is successfully reduced by the addition of L2 regularisation, which leads to a slight increase in prediction accuracy (R2 = 0.7835).
By addressing intra-cluster variance, the hybrid KMeans + LightGBM model improves performance to R2 = 0.8782 by clustering heterogeneous data before applying specialised LightGBM learners.
Tuned LightGBM: By increasing the R2 score to 0.8808 and enhancing the model's ability to handle nonlinear dependencies, hyperparameter optimisation further improves model performance.
Stacking Regressor: The best predictive ability is obtained by combining several base learners in an ensemble (R2 = 0.8858). It achieves the lowest overall prediction error and robust generalisation by combining complementary model strengths.

VII. VISUALIZATION SUMMARY

A. Model Performance Comparison (R2 score)

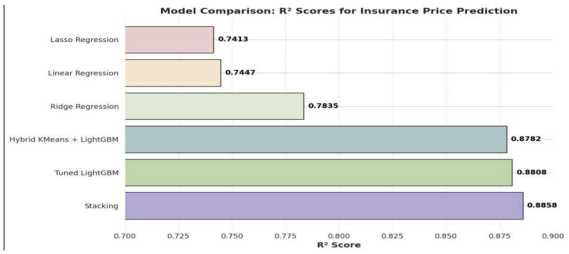

Figure 8. Model Performance Comparision

B. MAE Comparison Across Models

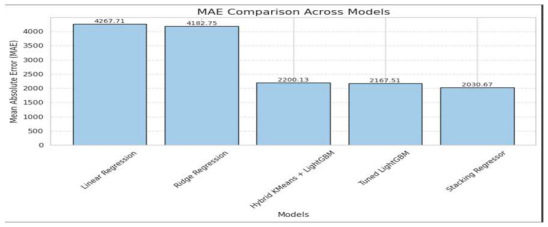

Figure 9. MAE Comparison across Models

C. Predicted vs actual plot

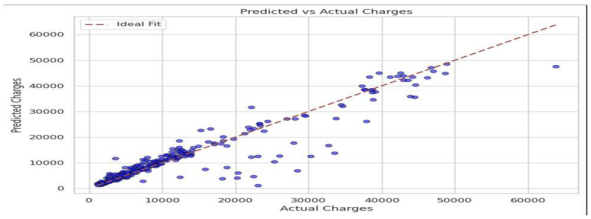

Figure 10. Predicted vs actual plot

D. KMeans Clustering Visualization

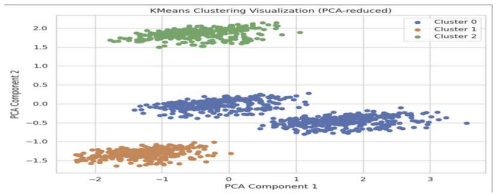

Figure 11. KMeans Clustering visualization

E. LightGBM Feature Importance

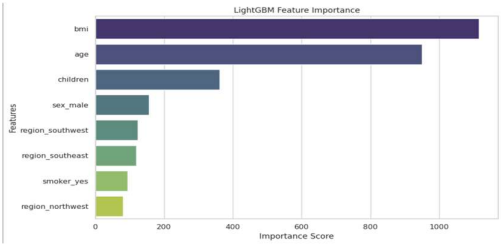

Figure 12. LightGBM Feature Importance

F. Distribution of Insurance Charges before and after Log Transform

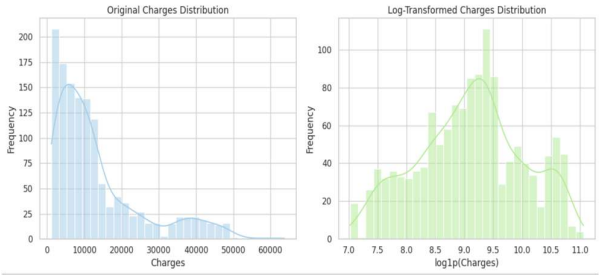

Figure 13. Distribution of Insurance Charges

G. Cross Validation Score Plot

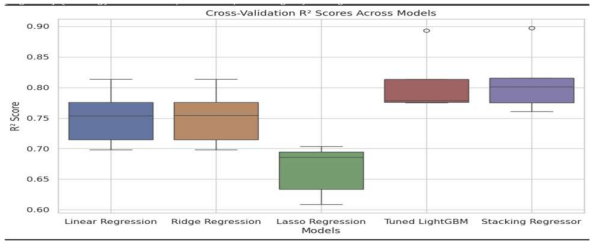

Figure 14. Cross Validation Plot

IX. CONCLUSIONS

The Stacking Regressor outperformed all other models (R2 = 0.8858, Fig. 9), closely followed by Tuned Light GBM and Hybrid KMeans + Light GBM. The MAE comparison (Fig. 10), the Predicted vs. Actual plot (Fig. 11), and KMeans clustering (Fig. 12) assisted in capturing subgroup-specific patterns.

Age, BMI were found to be important predictors by feature importance analysis (Fig. 13), while log transformation greatly enhanced the charge distribution (Fig. 14). According to cross-validation scores, the models also showed good generalisation (Fig. 15).

REFERENCES

[1] K. Kaushik, A. Bhardwaj, A. D. Dwivedi, and R. Singh, "Machine Learning-Based Regression Framework to Predict Health Insurance Premiums," International Journal of Environmental Research and Public Health, vol. 19, no. 13, Art. no. 7898, Jun. 2022. Do

[2] M. D. Shojaee, F. Asghari, M. Jannati, and N. Jannati, "Application of machine learning models in predicting length of stay among patients with hip fracture," International Journal of Environmental Research and Public Health, vol. 19, no. 13672, pp. 1–14, 2022, doi: 10.3390/ijerph191013672.

[3] K. Bhatia, S. S. Gill, N. Kamboj, M. Kumar, and R. K. Bhatia, "Health Insurance Cost Prediction using Machine Learning," in Proc. 2022 3rd Int. Conf. Emerging Technology (INCET), Belgaum, India, May 2022, pp. 1–6. doi: 10.1109/INCET54531.2022.9824201.

[4] S. Goel and A. Chaudhary, "Prediction of Health Insurance Price using Machine Learning Algorithms," in Proc. 2022 Int. Conf. on Recent Advances in Engineering and Computational Sciences (RAECS), IEEE, 2022. doi: 10.1109/RAECS56867.2022.10097485.

[5] A. Kharwal, "Health insurance premium prediction with machine learning," Thecleverprogrammer.com, Oct. 26, 2021. [Online]. Available: https://thecleverprogrammer.com/2021/10/26/health-insurance-premium-prediction-with-machine-learning/

[6] N. Boodhun and M. Jayabalan, "Risk prediction in life insurance industry using supervised learning algorithms," Complex Intell. Syst., vol. 4, pp. 145–154, 2018. doi: 10.1007/s40747-018-0072-1.

[7] N. V. Sailaja, M. Karakavalasa, M. Katkam, D. M., S. M., and D. N. Vasundhara, "Hybrid regression model for medical insurance cost prediction and recommendation," in Proc. 2021 IEEE Int. Conf. on Intelligent Systems, Smart and Green Technologies (ICISSGT), Visakhapatnam, India, Nov. 2021, pp. 1–6. doi: 10.1109/ICISSGT52025.2021.00029.

[8] T. Omar, M. Zohdy, and J. Rrushi, "Clustering application for data-driven prediction of health insurance premiums for people of different ages," in Proc. 2021 IEEE Int. Conf. on Consumer Electronics (ICCE), Las Vegas, NV, USA, Jan. 2021, pp. 1–6. doi: 10.1109/ICCE50685.2021.9427598.

[9] R. Sahai, A. Al-Ataby, S. Assi, M. Jayabalan, P. Liatsis, C. K. Loy, A. Al-Hamid, S. Al-Sudani, M. Alamran, and H. Kolivand, “Machine Learning for Insurance Risk Classification and Interpretability,” in Lecture Notes on Data Engineering and Communications Technologies, vol. 165, Proc. International Conference on Data Science and Emerging Technologies (DSET), Springer, 2023, pp. 1–12.
[10] N. V. Sailaja, M. Karakavalasa, M. Katkam, D. M., S. M., and D. N. Vasundhara, “Hybrid Regression Model for Medical Insurance Cost Prediction and Recommendation,” in Proc. 2021 IEEE International Conference on Intelligent Systems, Smart and Green Technologies (ICISSGT), Visakhapatnam, India, Nov. 13–14, 2021, pp. 1–6, doi: 10.1109/ICISSGT52025.2021.00029.
[11] S. Kumar, M. Srivastava, V. Prakash, and S. M. N. Islam, “A Fusion of LightGBM, PLSRegression, and ElasticNet Regressors in Advanced Ensemble Stacking for Empowering Predictive Modeling,” in Proc. 2024 4th International Conference on Technological Advancements in Computational Sciences (ICTACS), Tashkent, Uzbekistan, Nov. 13–15, 2024, pp. 1–7, doi: 10.1109/ICTACS62700.2024.10840828.
[12] M. S. Patil, S. Kulkarni, and S. Khurpe, “Medical insurance premium prediction with machine learning,” Int. J. Innov. Eng. Res. Technol. (IJIERT), vol. 11, no. 5, pp. 5–12, May 2024, doi: 10.26662/ijiert.v11i5.pp5-12.