Insurance Fee Prediction Based on k-Means Clustering and Regression Models: A Residual-Weighted Adjustment Method

機器學習_第5組
統計所 張伊萱

INTRODUCTION

- Challenges in Insurance Fee Prediction:Traditional regression models are typically used to predict continuous response variables, assuming that all data points have an equal impact on the model. This assumption does not always hold in real-world data, as the presence of outliers can significantly affect the prediction results, thereby reducing the model’ s accuracy.
- Limitations of Existing Methods:Traditional regression models do not address outliers, which means their impact on the model is not effectively adjusted. As a result, the prediction accuracy decreases, and the model fails to effectively differentiate the characteristics of different groups.

- Research Motivation:
  - How can clustering techniques be used to group the data, enabling the model to make predictions based on the distinct characteristics of different groups?
  - How can weighting adjustments be introduced to reduce the impact of outliers, thereby improving prediction accuracy?

METHODOLOGY

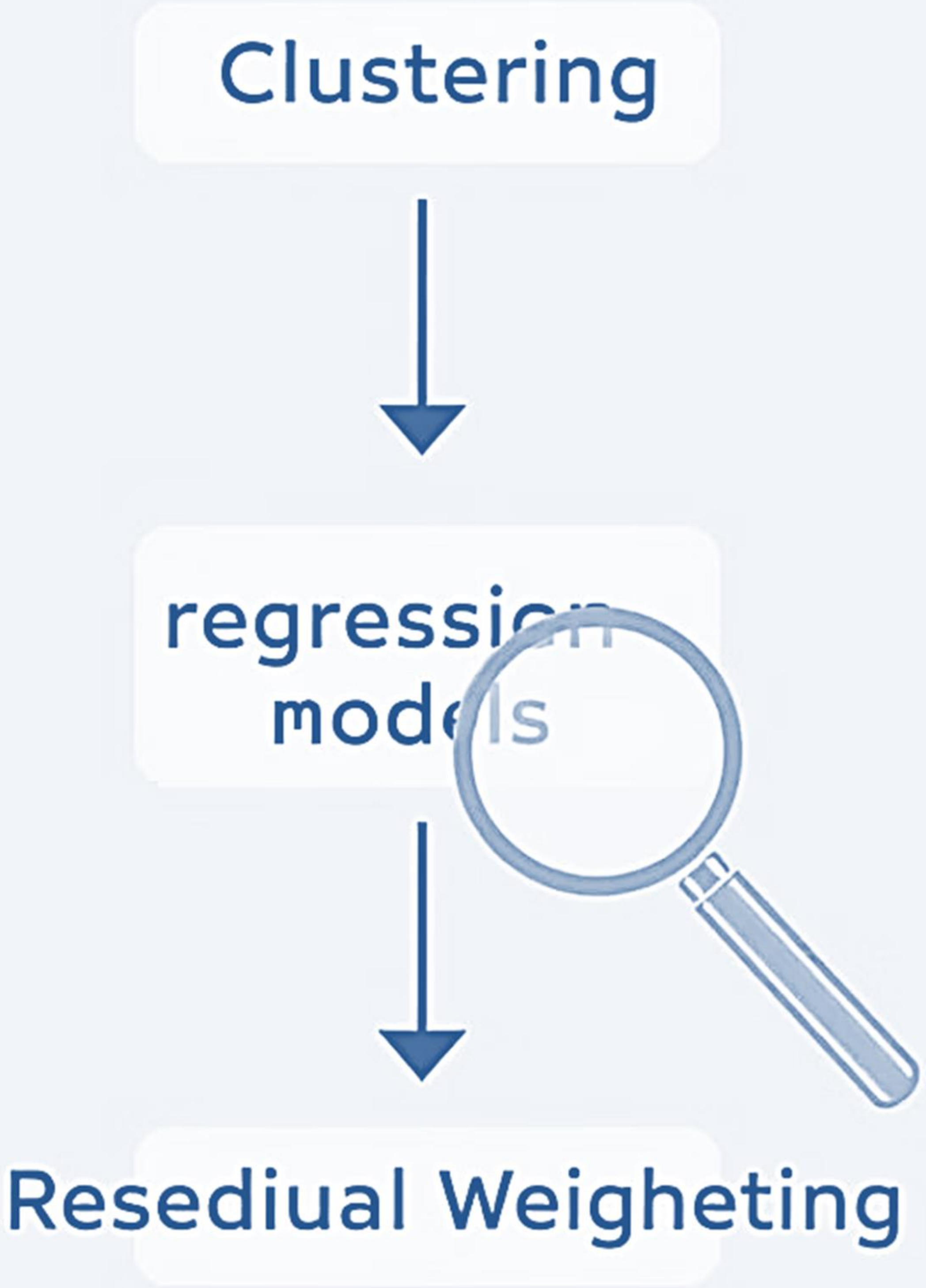
The proposed methodology integrates k-means clustering,regression models, and a residual weighting adjustment technique to enhance the accuracy and robustness of insurance premium predictions. By dividing the data into homogeneous groups through k-means clustering, distinct patterns within each group are captured by separate regression models.
Additionally, the influence of outliers is minimized using a residual-based weighting adjustment, ensuring a more reliable model performance.

Related Research

- **k-means clustering:** To determine the number of clusters k, the Elbow Methodis used. This method analyzes the trend of the Within-Cluster Sum of Squares (WCSS) and selects k such that the reduction inWCSS becomes relatively flat (Humaira, H. and Rasyidah, R. , 2020).

- Weighted Least Squares (WLS): WLS, discussed by Carroll and Ruppert (2017), adjusts for heteroscedasticity by introducing a weight matrix to correct for variance differences among data points, unlike OLS, which assumes equal variance.

- Robust Regression: Rousseeuw and Leroy (2003) highlight robust regression methods, which assign lower weights to outliers or high-leverage points to reduce their influence, ensuring model stability when handling real-world data with errors or extreme values.



Clustering

regression models

Resediual Weigheting

Workflow

Step 1

K-means clustering is applied using continuous variables to group the data and minimize intra-group heterogeneity, with the optimal number of clusters determined by the elbow method.

Step 2

Separate regression models are built for each group to better capture the behavioral characteristics of each group in predictions.

Model: linear regression, random forest regression, and CatBoost regression

Step 3

Weighted adjustments are applied to data points with large residuals to reduce the impact of outliers on the results.

$$weight = \frac{1}{1 + \text{weight\_factor} \times |residual|}$$

ILLUSTRATIVE EXAMPLE

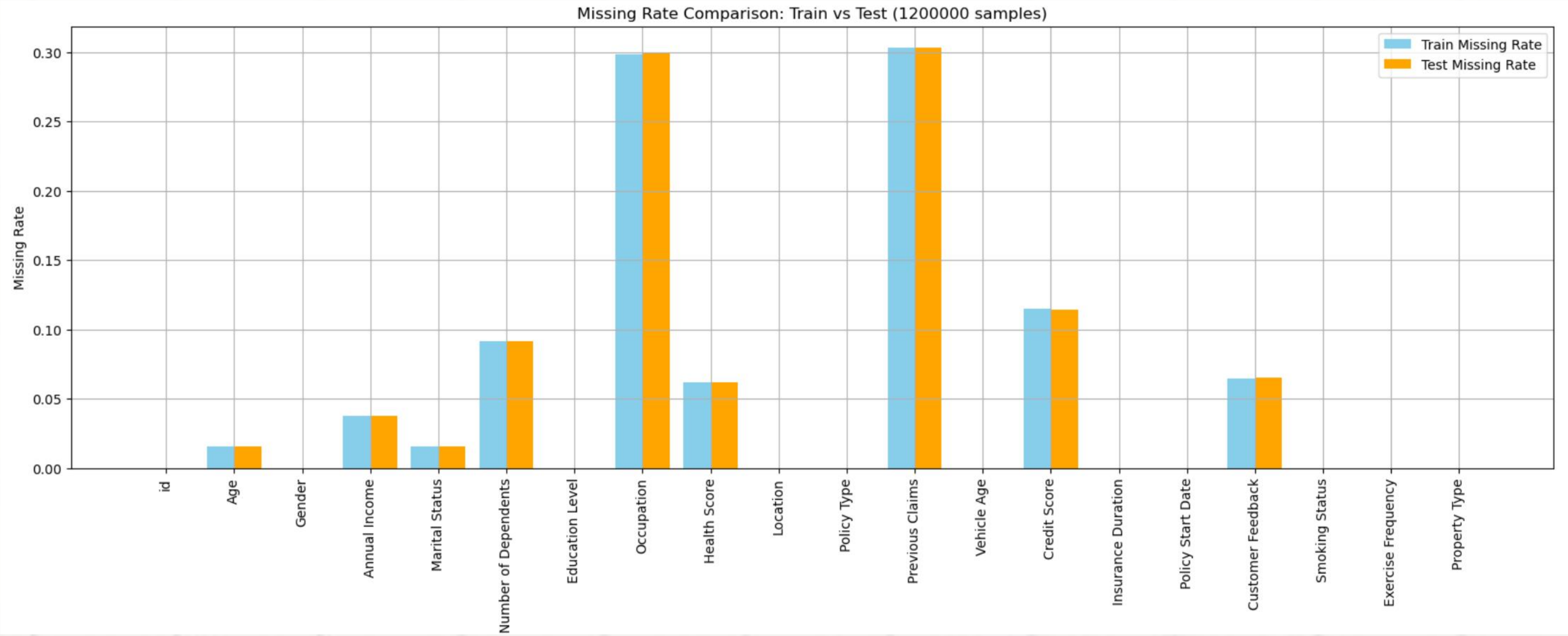
To demonstrate the methods described, we utilize the Regression with an Insurance Dataset available on Kaggle.
This dataset contains a variety of features related to insurance policies, including both continuous and discrete variables. Our objective is to preprocess the data, handle missing values, and predict the premium amount. The
dataset is split into two main sets: the training set, consisting of 1,200,000 records with 21 variables (including the target variable), and the testing set, which contains 800,000 records with 20 variables. Additionally, 10% of the training set is reserved as a validation set to evaluate the performance of the models during training.

Dataset Overview

The dataset includes key attributes such as:

- **Continuous Features**: Age, annual income, Health Score, etc.
- **Discrete Features**: Gender, Marital Status, Location, etc.
- **Target Variable**: Premium Amount.

Handling Missing Values



For Continuous Features:
Missing values are filled using the SimpleImputer with the median strategy. This approach minimizes the impact of missing data while preserving the overall distribution of the features.

• For Discrete Features:
Missing values are replaced with a new category labeled “Missing.” This method retains the information about missing data without discarding any records.

Feature Engineering

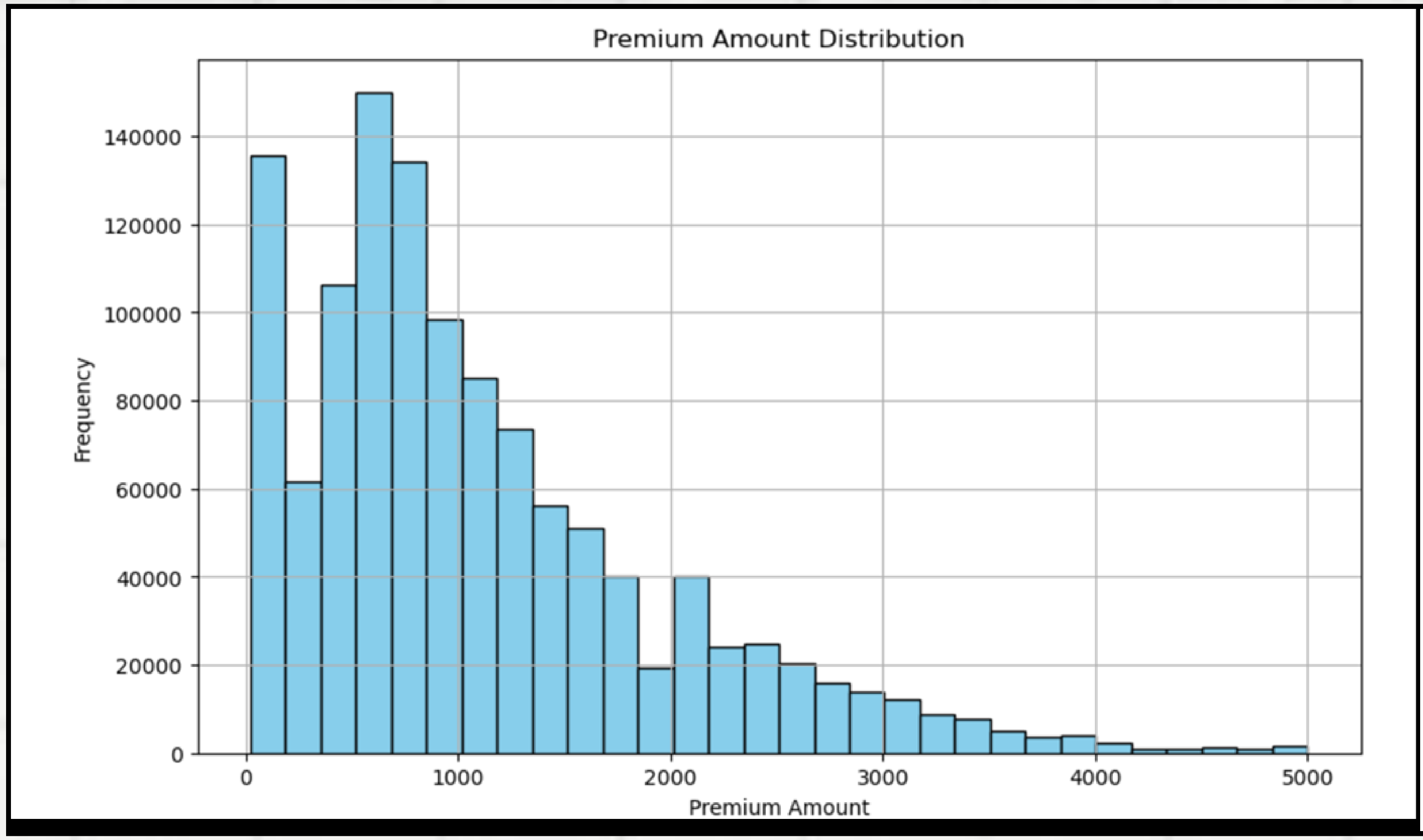
- Extracting Policy Start Year: The Policy Start Date column is converted to a datetime format, and the year is extracted to create a new feature called Policy Start Year, representing the start year of the policy. This feature is treated as a continuous variable.

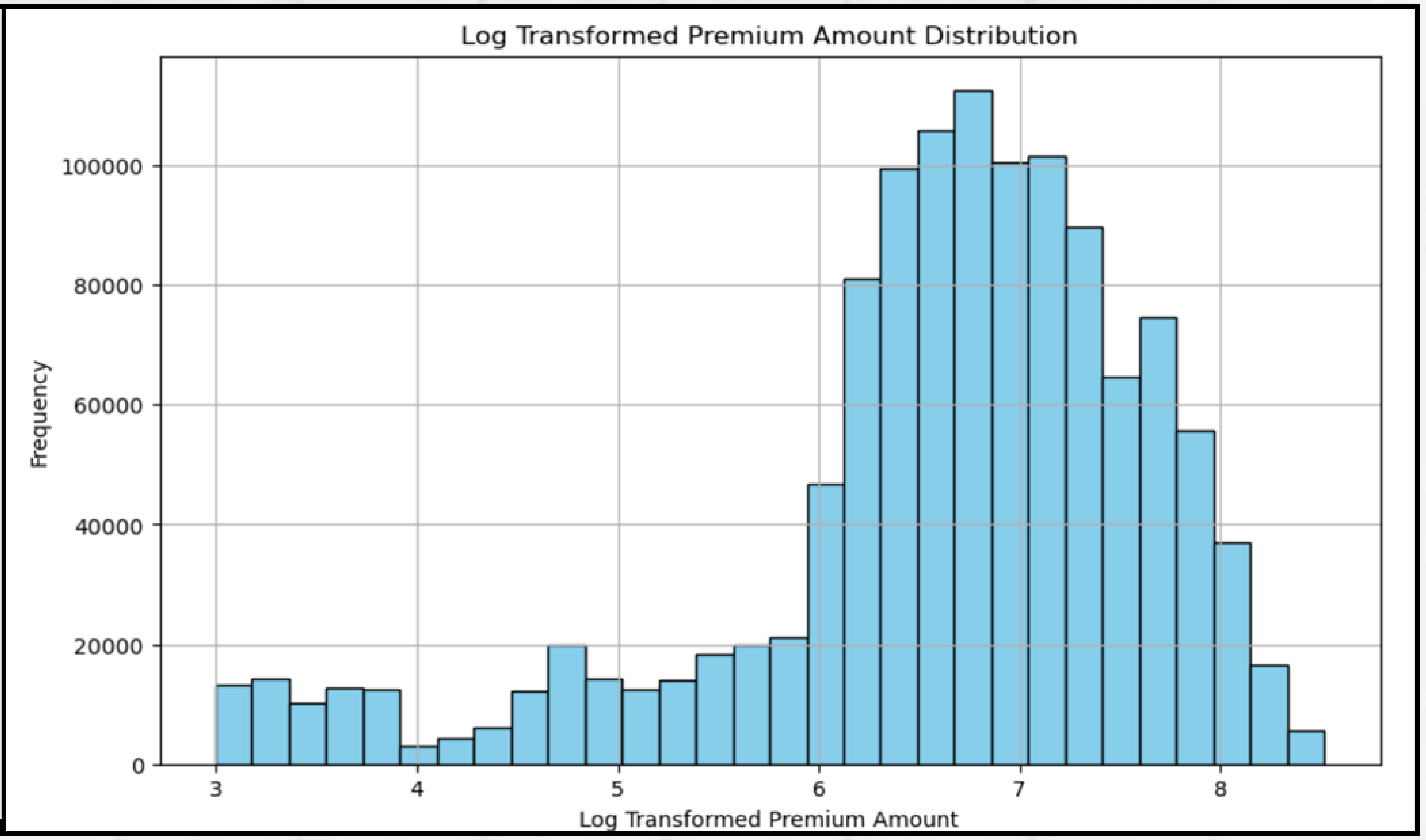
- Extracting Policy Start Month: Similarly, the Policy Start Date column is converted to a datetime format, and the month is extracted to create a new feature called Policy Start Month, representing the start month of the policy. This feature is treated as a continuous variable.

- Categorizing Dates into Early, Mid, or Late Month: Based on the day in the Policy Start Date, dates are categorized into three groups:

– If the day is between 1 and 10, it is categorized as "Early Month."
– If the day is between 11 and 20, it is categorized as "Mid Month."
– If the day is between 21 and the end of the month, it is categorized as "Late Month.
This new feature, Policy Start Period, is added and treated as a discrete feature

Response Variable Analysis





The histogram reveals that the Premium Amount distribution is right-skewed. To stabilize variance, reduce sensitivity to extreme values, and improve model performance, a logarithmic transformation is applied. This transformationbrings the distribution closer to normal, enhancing the model’ s fit. After applying the log transformation, we proceed with regression modeling for improved predictions.

Model

We developed a traditional regression model to predict the log-transformed Premium Amount and compared its performance with cluster-based weighted regression models (k-means + Linear Regression, k-means + Random Forest, k-means+ CatBoost). The evaluation metric used for comparison is the Mean Absolute Percentage Error (MAPE), calculated as:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100\%$$

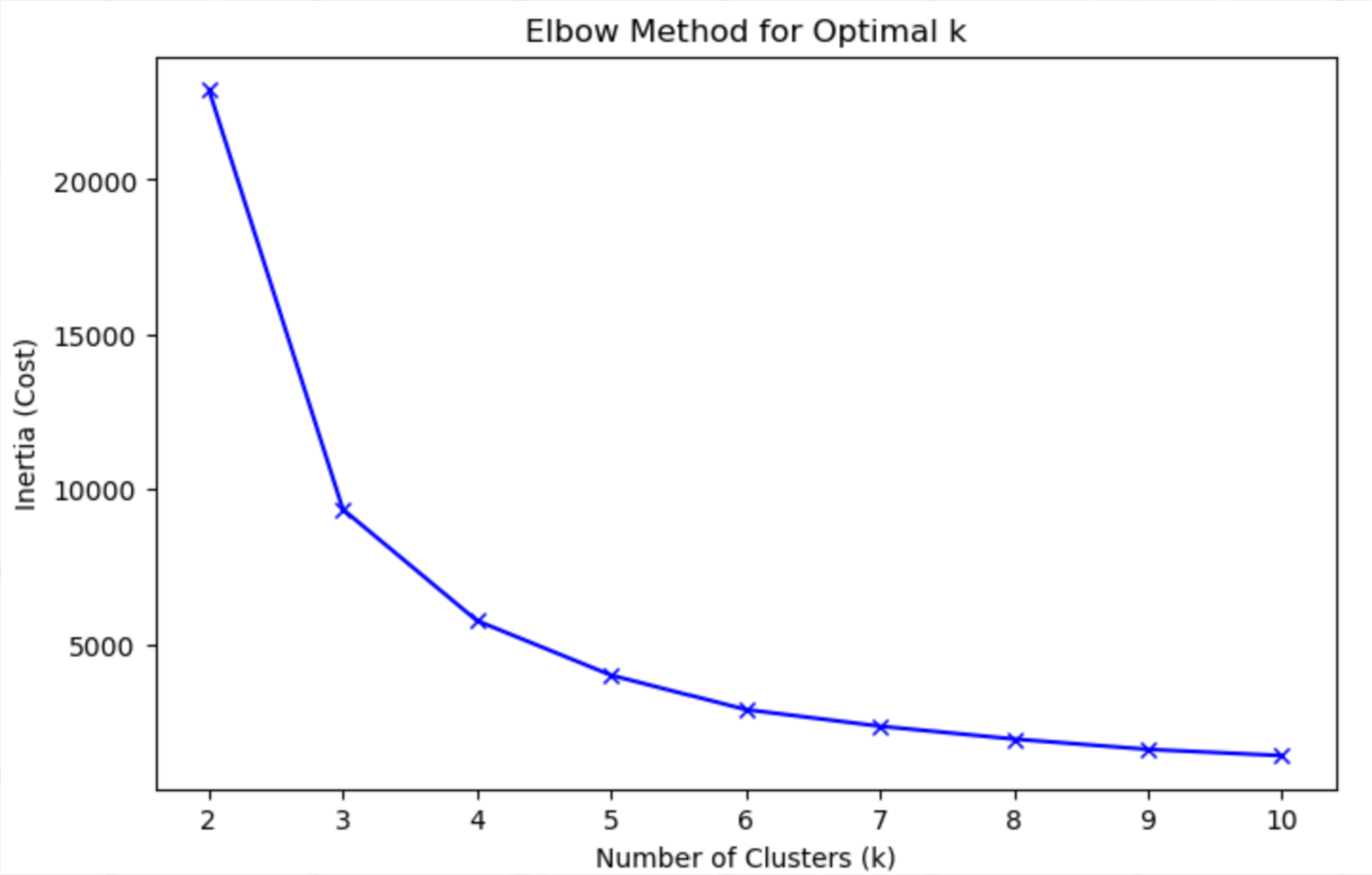
where y_i is the actual value, $\hat{y}_i$ is the predicted value, and n is the total number of observations. This metric quantifies prediction accuracy by expressing the errors as a percentage of the actual values. To ensure meaningful comparisons, the results were transformed back to the original scale of the Premium Amount after log-transformation.

Setting the weight factor to 0.6 strikes a balance by reducing the influence of outliers while still accounting for their potential value in the model.

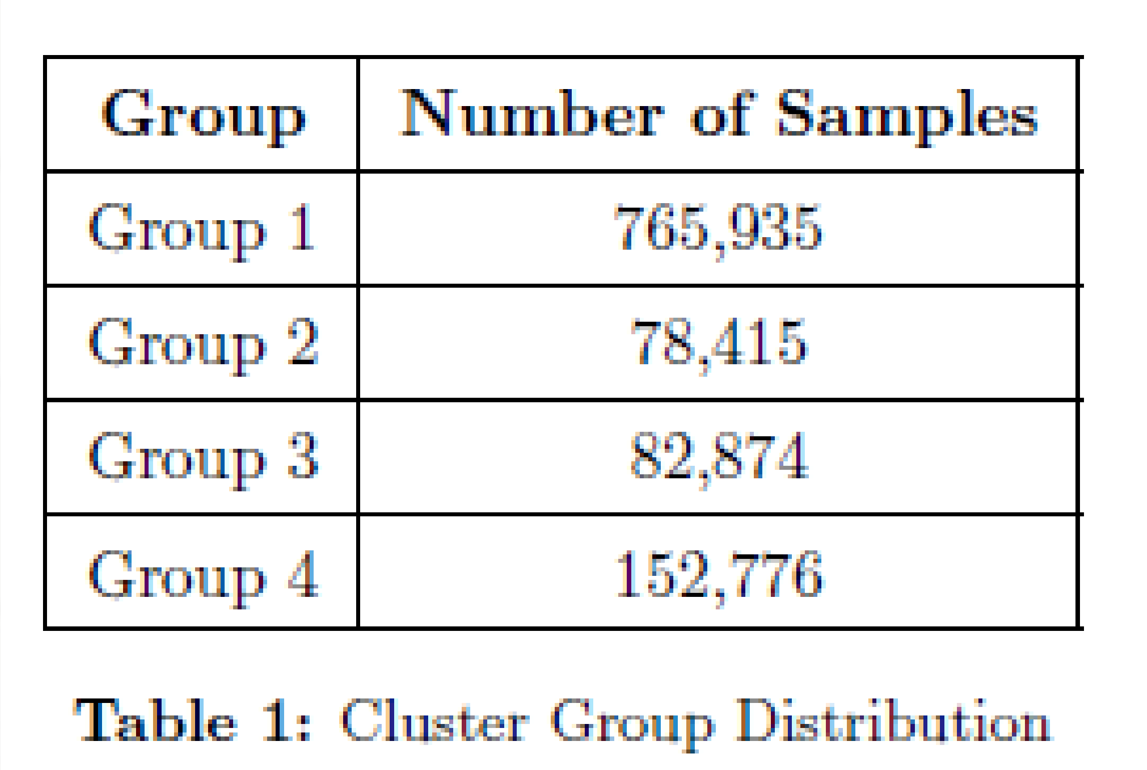
Model hyperparameter settings:
Random Forest: n estimators = 100, max depth = 10
CatBoost: iterations = 500, learning rate = 0.1

Cluster Result

The optimal number of clusters is determined using the elbow method, as visualized in the plot, which indicates that 4 clusters is the optimal choice



Using k-means with k = 4, the clusters are formed.

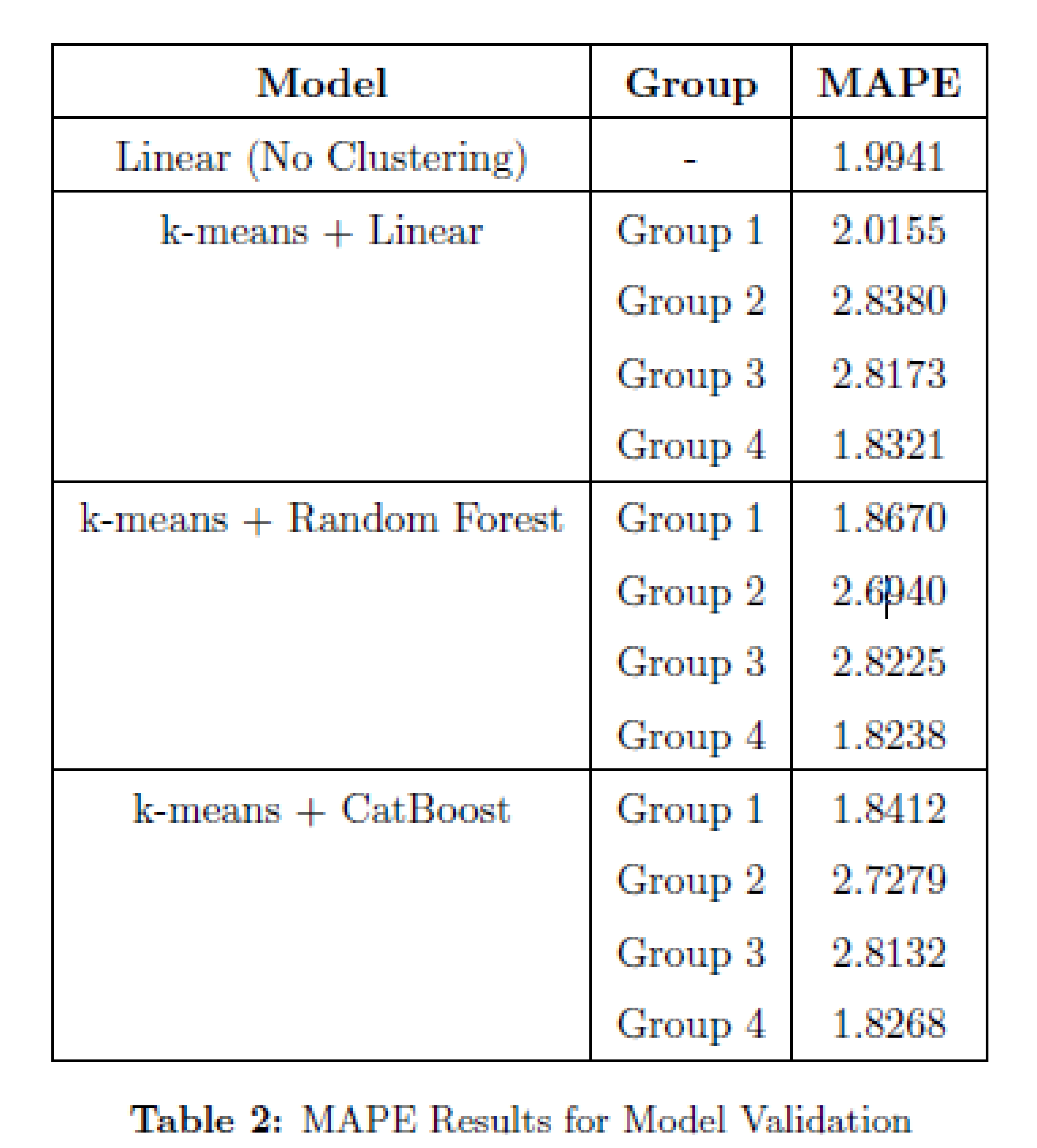


Group	Number of Samples
Group 1	765,935
Group 2	78,415
Group 3	82,874
Group 4	152,776

Table 1: Cluster Group Distribution

Result

models were built using Linear Regression, Random Forest, and CatBoost based on the clustering results, as well as a Linear Regression model built directly without clustering. The following table presents the MAPE values for model validation.

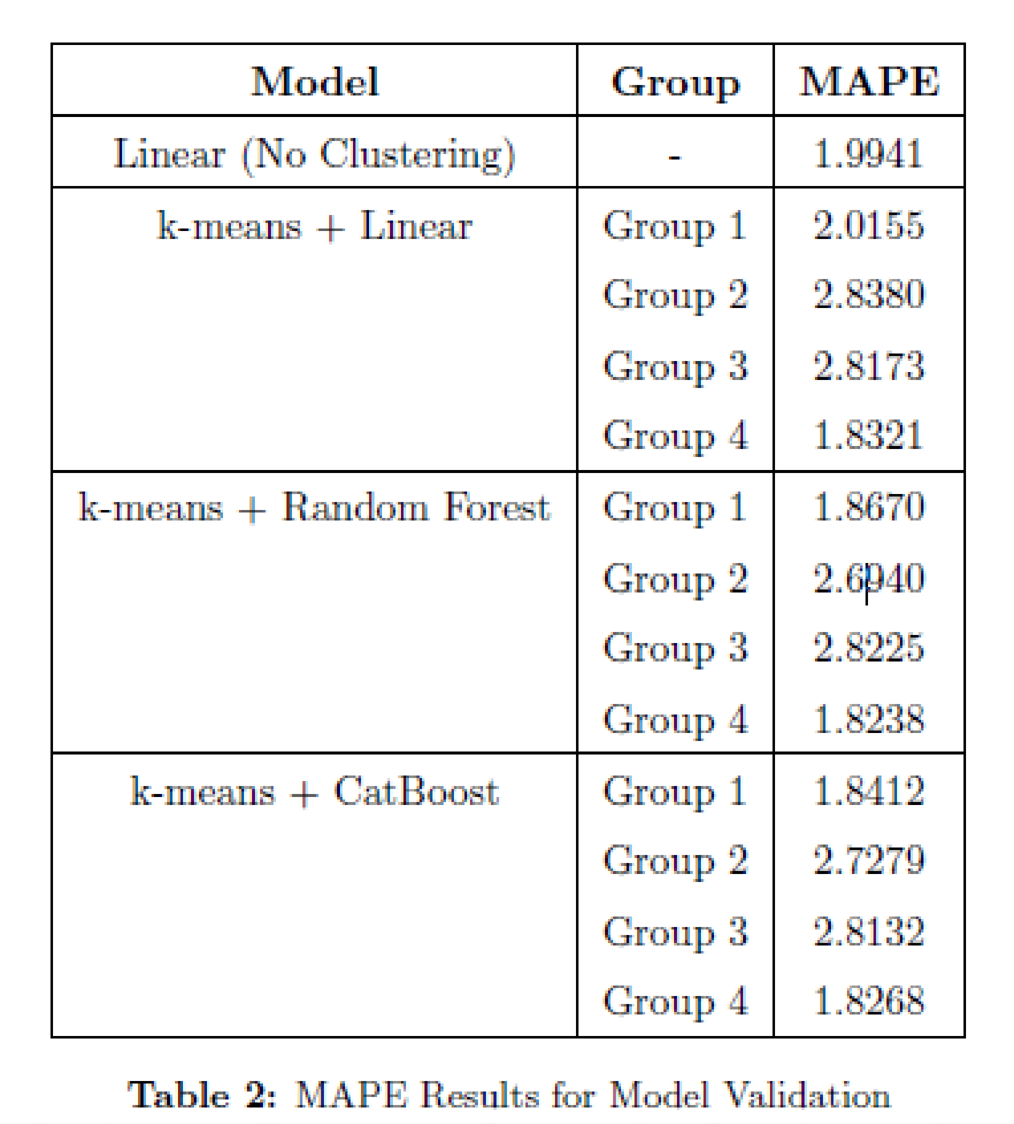


Model	Group	MAPE
Linear (No Clustering)	-	1.9941
k-means + Linear	Group 1	2.0155
	Group 2	2.8380
	Group 3	2.8173
	Group 4	1.8321
k-means + Random Forest	Group 1	1.8670
	Group 2	2.6940
	Group 3	2.8225
	Group 4	1.8238
k-means + CatBoost	Group 1	1.8412
	Group 2	2.7279
	Group 3	2.8132
	Group 4	1.8268

Table 2: MAPE Results for Model Validation

Result

models were built using Linear Regression, Random Forest, and CatBoost based on the clustering results, as well as a Linear Regression model built directly without clustering. The following table presents the MAPE values for model validation.



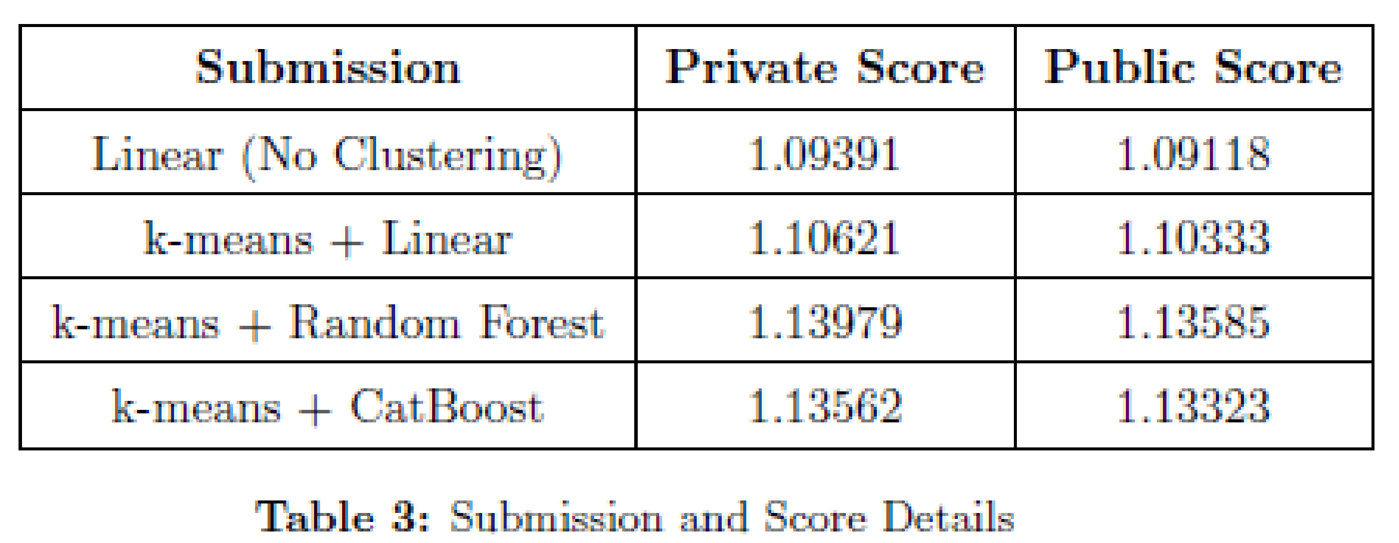
Model	Group	MAPE
Linear (No Clustering)	-	1.9941
k-means + Linear	Group 1	2.0155
	Group 2	2.8380
	Group 3	2.8173
	Group 4	1.8321
k-means + Random Forest	Group 1	1.8670
	Group 2	2.6940
	Group 3	2.8225
	Group 4	1.8238
k-means + CatBoost	Group 1	1.8412
	Group 2	2.7279
	Group 3	2.8132
	Group 4	1.8268

Table 2: MAPE Results for Model Validation

Clustering appears to improve prediction accuracy for certain groups, such as Group 4, but its benefits are not consistent across all clusters.

Result

Finally, the model is applied to the test data, and the predictions are submitted to Kaggle to calculate the score, with a lower score being better.



Submission	Private Score	Public Score
Linear (No Clustering)	1.09391	1.09118
k-means + Linear	1.10621	1.10333
k-means + Random Forest	1.13979	1.13585
k-means + CatBoost	1.13562	1.13323

Table 3: Submission and Score Details

The performance of the cluster-based weighted regression did not outperform the unclustered model.

Conclusion and Future Research

This method improves insurance fee prediction accuracy and stability by combining clustering regression and residual weighting adjustments. Clustering regression enables more refined modeling based on data heterogeneity, enhancing predictions, while residual weighting reduces the impact of outliers, stabilizing the model. Despite room for improvement, this approach provides a new solution for handling outliers, offering potential for application.

Future Directions

- **Explore Additional Clustering Methods**: Future work could explore alternative clustering methods like hierarchical clustering or DBSCAN to improve clustering results and compare their predictive performance.
- **Optimize Weighting Adjustments**: Further optimization of weighting adjustments, such as adaptive weight factors, could make the method more flexible and better at handling various outlier scenarios.
- **Cross-Domain Applications**: This method could extend beyond insurance fee prediction to other fields dealing with outliers and heterogeneous data, such as sales forecasting and healthcare cost estimation.

REFERENCES

Carroll, R. J., Ruppert, D. (2017). Transformation and weighting in regression. Chapman and Hall/CRC. Humaira, H., Rasyidah, R. (2020). Determining the appropriate cluster number using elbow method for k-means algorithm. In Proceedings of the 2nd Workshop on Multidisciplinary and Applications (WMA) 2018, 24-25 January 2018, Padang, Indonesia. Rousseeuw, P. J., Leroy, A. M. (2003). Robust regression and outlier detection. John Wiley & Sons.

GitHub link

https://github.com/Zxc15495qaw/ML_Final_Project

Thank you for listening