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

張伊萱

Department of Statistics, College of Management,
National Cheng Kung University, Taiwan
Email: r26121023@gs.ncku.edu.tw

ABSTRACT

This study proposes a method for insurance premium prediction that integrates k-means clustering, regression models, and a residual-based weighting adjustment. The approach first divides data into groups using k-means clustering, then applies separate regression models for each group, and adjusts the influence of data points with large residuals to reduce the impact of outliers. Experimental results in real insurance data sets indicate that the proposed method may offer improved prediction accuracy and stability for some categories compared to traditional regression models, providing a useful approach for handling heterogeneous data in insurance. Future research could explore alternative clustering techniques and incorporate more advanced machine learning models to further enhance performance.

Keywords: Insurance prediction, k-means clustering, regression models, residual weighting.

1 INTRODUCTION

Insurance premium prediction is a critical task in actuarial science and risk management, where the accuracy of predictions plays a key role in assessing financial risks. Traditional regression models often assume that the data points are homogeneous, and each observation contributes equally to the model. However, real-world data frequently contains heterogeneity and anomalies that can distort the results. In such cases, traditional models may fail to deliver reliable predictions.

This study aims to address these challenges by proposing a new approach that combines k-means clustering, regression models, and a residual-based weighting adjustment method. By leveraging k-means clustering, the data is divided into groups with minimized internal heterogeneity, allowing for more tailored regression models. In addition, a residual weighting mechanism is introduced to reduce the influence of outliers, thus improving the robustness of the model.

The proposed method provides a promising solution for handling complex, heterogeneous data in insurance premium prediction and offers a more stable and accurate alternative to traditional regression models.

2 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.

2.1 K-means Clustering

In insurance fee prediction, data heterogeneity often leads to a single model being insufficient to capture all the relationships within the samples. To address this issue, this study utilizes the unsupervised learning k-means clustering method, which divides the data into several groups, where each group contains more similar data.

2.1.1 Clustering Objective

The objective of k-means clustering is to minimize the squared Euclidean distance between samples and their respective cluster centroids, formulated as:

$$\min \sum_{i=1}^{n} \sum_{j=1}^{k} r_{ij} \|x_i - \mu_j\|^2$$

where r_{ij} indicates whether sample x_i belongs to cluster j, and μ_j is the centroid of cluster j. This method has been widely applied in clustering heterogeneous data.

2.1.2 Cluster Number Selection

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

By clustering, the study can effectively capture the characteristics of different insurance fee patterns, providing more targeted data for subsequent regression modeling.

2.2 Weighted Regression

Weighted regression is a method used to handle outliers or samples with differing variabilities in the data. When different observations in the data have varying levels of reliability, it is possible to assign different weights to each data point. These weights reflect the contribution of each observation to the regression model. A common issue in regression analysis is heteroscedasticity, which refers to the varying variability (variance of error terms) of different sample points. Weighted regression allows for the assignment of different weights to correct for these differences

.

2.2.1 Core Concepts

The core concepts of weighted regression are centered around the following key components:

Weighted Least Squares (WLS) One of the primary techniques in weighted regression is the Weighted Least Squares (WLS) method. This method is discussed in detail in *Transformation and Weighting in Regression* (Carroll Ruppert, 2017), which explains the mathematical foundation of weighted regression, particularly in the context of heteroscedasticity. Traditional Ordinary Least Squares (OLS) assumes equal variance of error terms (homoscedasticity), which does not always hold in real-world data. The WLS approach introduces a weight matrix to adjust for differences in variance among data points. By assigning weights to each observation, WLS can correct for the heteroscedasticity that is commonly found in practical regression analysis.

Robust Regression

Another approach discussed in *Robust Regression and Outlier Detection* (Rousseeuw Leroy, 2003) focuses on the use of weighted regression in the context of outlier detection and robustness. This method not only accounts for varying error variances but also reduces the influence of outliers in the regression model. By assigning lower weights to observations that are considered outliers or have high leverage, robust regression methods ensure that these points do not unduly influence the model. This is particularly useful when dealing with real-world data that may contain errors or extreme values.

Incorporating these techniques into our insurance fee prediction model helps account for variability in data reliability, improving model accuracy and robustness.

2.3 Workflow Overview

The proposed methodology, as illustrated in Figure 1, follows these steps:

1. **Clustering:** Using k-means clustering, the training dataset is divided into several homogeneous clusters,

allowing for more focused regression modeling tailored to each cluster's characteristics. The optimal number of clusters is determined using the elbow method.

2. **Initial Regression Models:** For each cluster, an initial regression model (e.g., linear regression, random forest regression, CatBoost regression) is built based on the training data. These models capture the underlying relationships within each cluster.

3. **Residual Weighting Adjustment:** Based on the residuals from the initial regression models, a weighting adjustment is applied to refine the models. The weight for each data point is calculated using the formula:

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

. This adjustment reduces the influence of large residuals, ensuring that the final model is less affected by outliers.

4. **Validation and Performance Evaluation:** The adjusted models are applied to the validation dataset to compute metrics such as Mean Absolute Percentage Error (MAPE), enabling a comparison of the accuracy and robustness of the proposed models.

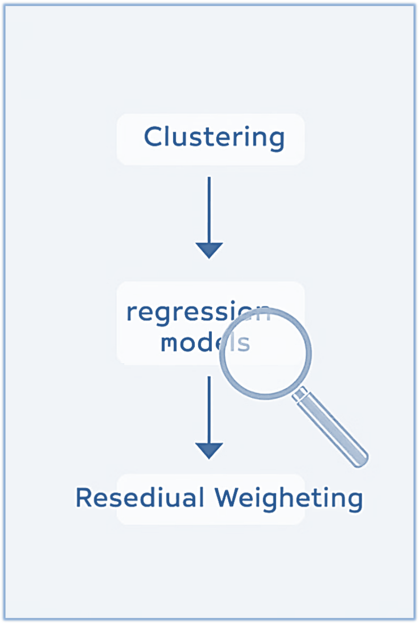

Figure 1: Flowchart of the Proposed Methodology

By integrating k-means clustering, regression modeling, and residual weighting adjustments, this workflow effectively addresses data heterogeneity and enhances the accuracy and robustness of insurance fee predictions.

3 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.

3.1 Data Preprocessing

Handling Missing Values In the accompanying figure, the proportion of missing values for each variable is displayed.

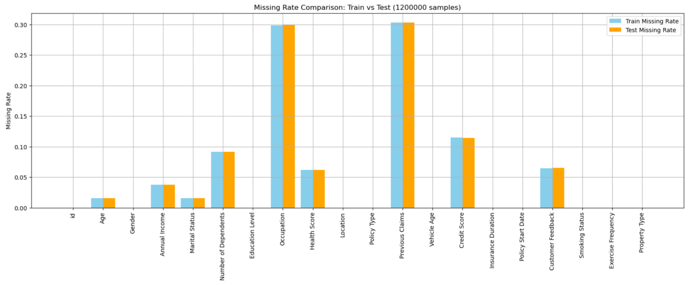

Figure 2: Enter Caption

- **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 To better understand the response variable, we first visualize its distribution. The histogram and boxplot below illustrate the distribution of the insurance charges `Premium Amount`), highlighting any potential skewness or outliers.

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 transformation brings the distribution closer to normal, enhancing the model's fit. After applying the log transformation, we proceed with regression modeling for improved predictions.

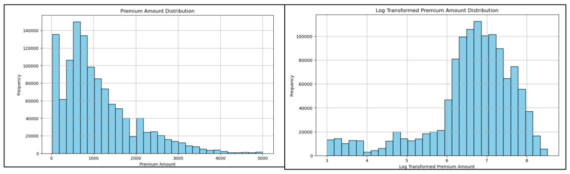

Figure 3: Distribution of `Premium Amount`

3.2 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:

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

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

3.3 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. The following table presents the resulting distribution of samples across the clusters.

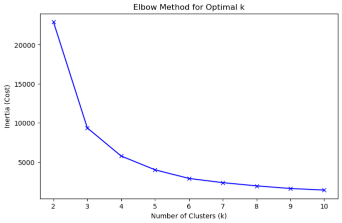

Figure 4: Elbow Method for Optimal k

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

Subsequently, 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.

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. The following table shows the private and public scores for the submissions:

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.

4 CONCLUSION

Improved Accuracy and Stability: This method enhances insurance fee prediction by combining clustering regression and residual weighting adjustment. Clustering refines modeling based on data heterogeneity, while residual weighting reduces the impact of outliers, improving stability. It significantly boosts accuracy, especially in handling extreme values and anomalies.

Potential of the Method: Although there is room for improvement, this approach offers a new solution to outlier issues that traditional regression models cannot address, demonstrating its potential in cases with strong data fluctuations and heterogeneity.

5 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`