Property insurance risk assessment and pricing using machine learning predictive models

Millicent Auma Omondi [1]

[1]African Masters for Machine Intelligence (AMMI)

Introduction

Machine Learning (ML) offers diverse applications that can significantly enhance insurance company operations. The use cases includes fraud detection, loss prevention, customer engagement, customer churn prevention, and premium pricing, among others. By leveraging ML techniques, insurance companies have the potential to save costs and enhance the overall customer experience. Our case focuses on predicting home insurance building and contents cover premiums, i.e, the money you pay in exchange for an insurance cover.

Background

In our daily lives, we are constantly exposed to various risks that pose a threat to individuals, businesses, buildings, and property. These risks can take different forms, including illness, death, and property loss. While it is not always possible to completely avoid these risks, the insurance industry operates on the premise of their existence and aims to mitigate or limit the potential damages they may cause. To achieve this, the insurance sector has developed a range of products that utilize financial resources to provide individuals and organizations with protection against these risks. Premium pricing serves as a mechanism employed by insurance firms to determine the cost of insurance policies. Machine learning (ML) can be leveraged by insurance companies to gain deeper insights into these risk factors and develop more accurate premium pricing models.

Problem Statement

Insurance companies face the challenge of striking a balance between attracting customers and maintaining their own financial viability. Setting appropriate premiums is crucial for this balance. If premiums are set too low, the company may struggle to cover potential losses and sustain operations. On the other hand, excessively high premiums can discourage customers from purchasing insurance altogether.

Objectives

- Identify the machine learning models that demonstrate the highest predictive performance for the home insurance premium.
- Find out what are the most influential features that significantly impact home insurance premiums.

Risk Assessment

Before an insurance company accepts to insure a property, it first conducts risk assessment. It considers several property related risk factors as follows:

- **Activity conducted in the building.**
- **Financial status of the prospective client.**
- **The building's geographic location.**
- **Building description,**e.g size,construction type, roof type, number of rooms,etc.
- **Strengths and weaknesses of the building.**
- **Past loss experience.**

Dataset

The dataset used was obtained from kaggle spanning 5 years from 2007 - 2012. It consists of $256,136$ observations, with 66 features that describe the home insurance policy characteristics.

Data Pre-processing

Before training the ML predictive models, the dataset went through some pre-processing steps. These steps include: **Checking for duplicate values** and **missing values**. Columns with substantial missing values were dropped. We used mean imputation to fill in the missing values of columns with fewer missing values. **Data transformation** into the right formats, **standardization**, **creation of new variables** and **handling outliers** using the winsorizer method to cap the extreme values.

Methodology

- We used supervised ML models like linear regression, lasso regression, ridge regression, decision-tree based regression models such as decision trees, random forests, gradient boosting, and extreme gradient boosting (XGBoost), and support vector regression (SVR).
- The dataset was split into two, 80% for training and 20% for testing. k-fold cross validation (CV) was applied with $k = 5$.
- We applied performance metrics such as R square score (R^2), mean absolute error (MAE) and mean square error (MSE). The lower MSE and MAE values, the better the model. R^2 ranges from 0-1, with values closer to 1 suggesting a good fitting model.

Exploratory Data Analysis

We performed some exploratory data analysis to look at the relationships within the data. From figure 1, we observe that homes that have a record of claims in the past 3 years tend to pay more premiums. Houses with more bedrooms also tend to pay more as seen in figure 2. This shows that the size of the house matters.

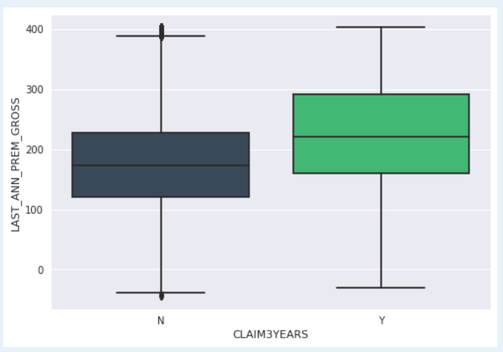

Figure 1. Premium vs Claim3years.

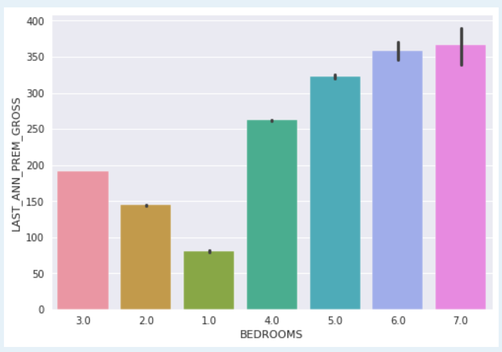

Figure 2. Premium vs Number of bedrooms.

Fitting and Evaluating the Models

Model	R Squared	R Squared CV	MSE	MAE
Linear Regression	0.7084	0.7131	2282.05	36.37
Ridge Regression	0.7084	0.7135	2281.98	36.37
Lasso Regression	0.7084	0.7027	2359.32	36.98
Decision Tree	0.6383	0.6332	2830.42	36.34
Random Forest	0.8140	0.8165	1455.40	26.32
KNN	0.6556	0.6571	1732.59	36.54
Gradient Boosting	0.7935	0.7984	1616.31	28.16
XGBoost	**0.8273**	**0.8307**	**1351.61**	**25.44**
SVR	0.006344	0.005015	7776.04	69.06

Table 1. Performance: Best scores are in bold.

Hyperparameter Tuning

To enhance the prediction performance of the top three models (random forests, gradient boosting and XGBoost), we conducted hyperparameter tuning using the GridSearchcv method. These models are all supervised ensemble machine learning techniques known for their effectiveness in regression tasks.

Model	R Squared	R Squared CV	MSE	MAE
Random Forest	0.8077	0.8117	1504.49	27.15
Gradient Boosting	0.8241	0.8281	1375.77	26.83
XGBoost	**0.8324**	**0.8351**	**1311.20**	**24.89**

Table 2. Model's performance after hyperparameter tuning: Best scores are in bold.

Assessing Feature Importance

By applying the Shapley additive explanation (SHAP) feature importance method, we were able to identify the most significant variables in our model as shown in Figure 3.

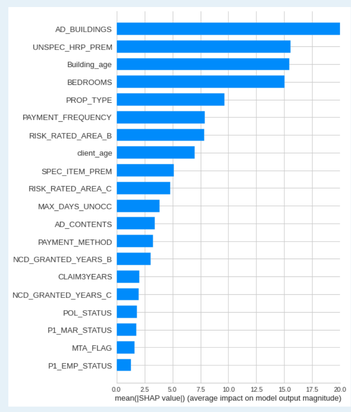

Figure 3. Important Features.

Fitting the Final Model: XGBoost

Instead of 66 features we only used the 20 features that were highlighted as important in training the final model. This ensured focusing on the most influential aspects.

	R Squared	MSE	MAE
Train	0.99907	7.2035	1.89412
Test	0.99912	6.9211	1.88174
CV	0.99908	7.13389	1.89526

Table 3. XGBoost Performance during training, testing and using cross validation

Visual Representation of XGBoost Performance

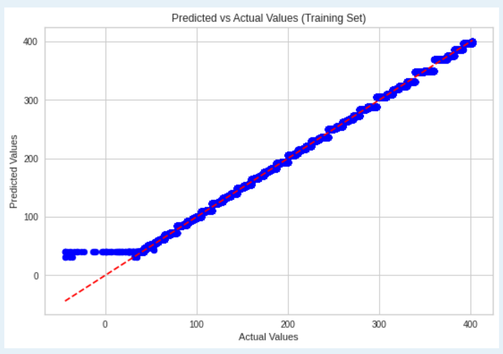

Figure 4. Train:Predicted vs Actual values.

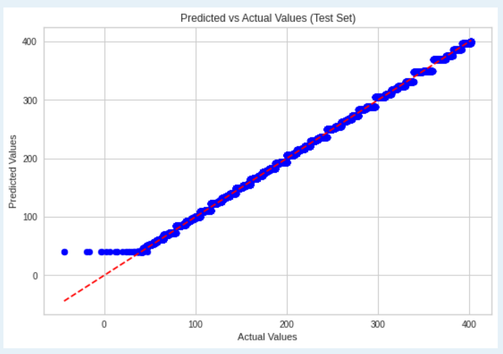

Figure 5. Test:Predicted vs Actual values.

The proximity of the predicted points (blue) to the red dashed line suggests that the model is making reliable predictions, demonstrating its ability to generalize well to unseen data.

Conclusion

In conclusion, we found that the XGBoost model outperformed other models and effectively fit the dataset. We were also able to identify the important features used for prediction.However, it is important to acknowledge a limitation of our study: the absence of access to local dataset. XGBoost could be used to forecast property insurance premium and compare the premiums obtained with actuarial formulas.

References

[1] Jyothsna, Chaparala, et al. "Health Insurance Premium Prediction using XGboost Regressor." 2022 International Conference on Applied Artificial Intelligence and Computing (ICAAIC). IEEE, 2022.

[2] Kulkarni, Mukund, et al. "Medical insurance cost prediction using machine learning." 2022 International Journal for Research in Applied Science Engineering Technology 10 (2022).