Structured Data Enabled Predictive Modelling for Property Insurance

by Ninad Patil

Introduction

With the growth of the real estate sector, there's a surge in demand for property insurance. Crucial for insurance companies to accurately predict insurance costs to ensure fair pricing and financial stability.

Traditionally, experts relied on mathematical methods and historical data for pricing property insurance. However, with evolving risks and data complexities, there's a need for more advanced tools.

Considering the multitude of variables at play, including property type and client demographics, our aim is to achieve a high level of precision in predicting the cost of insurance. Developing a system for accurately forecasting property insurance premiums.

Machine learning models offers enhanced accuracy in predicting insurance prices. Enabling a predictive model that closely approximates the actual insurance premiums.

Dataset Overview

1 It comprises of 256,136 rows and 42 columns

2 Having 20 Numerical Numerical and 22 Categorical Features Features

3 Covering details related to policy, personal information, property specifics

Data Preprocessing & EDA

Handling Missing Values

Imputed the values with appropriate techniques like techniques like median, forward fill

Data Visualization

Created various graphs to comprehend the distribution of the features and the spread.

Additon of New Feature

Enhancing the dataset with new features.

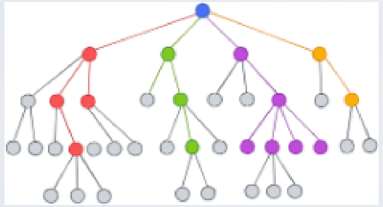
Standardization

Ensuring uniformity and consistency in dataset dataset attributes.

Handling Categorical Features

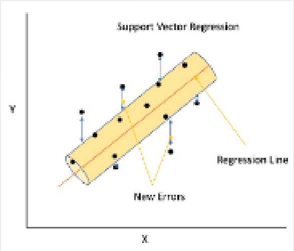
Used One-hot encoding technique, mapping of binary categorical features

Model Selection



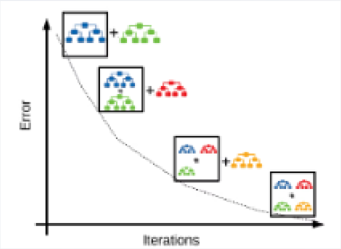
Decision Tree Regression

Utilizes a tree structure to model relationships between between features and and target variables. variables.



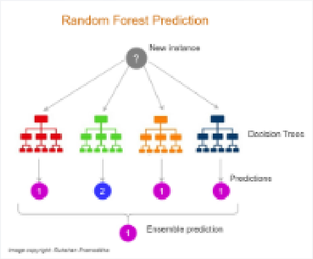
Support Vector Regression

Utilizes support vectors to approximate a continuous function, function, suitable for for both linear and non-linear data.



Gradient Boosting Regression

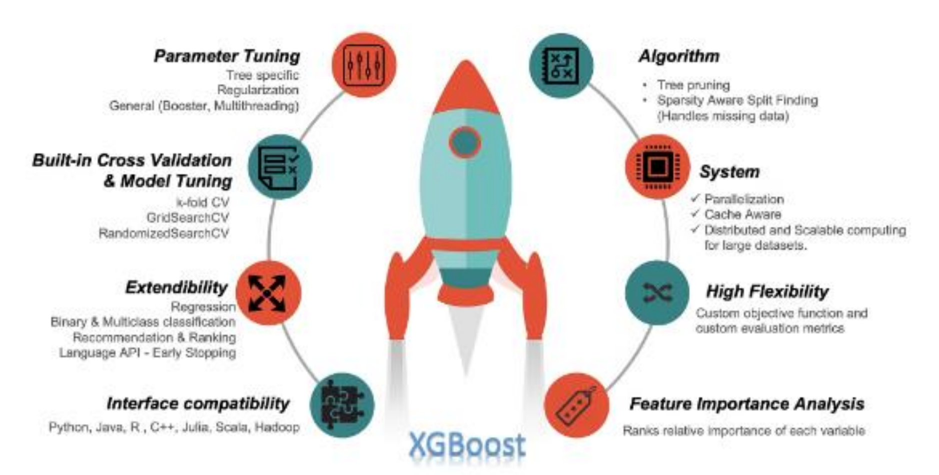
Builds multiple decision trees sequentially, each correcting errors of the previous one, resulting in powerful powerful predictive performance.



Random Forest Regression

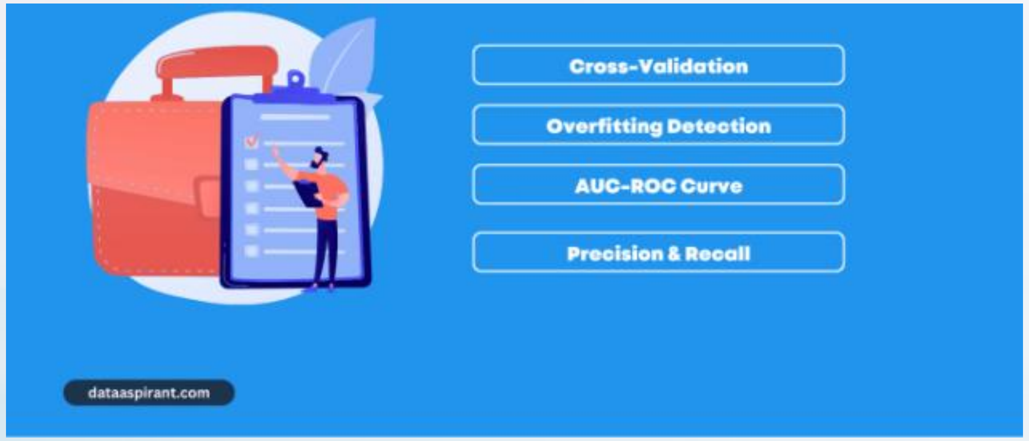
Utilizes an ensemble ensemble of decision decision trees to improve accuracy and and reduce overfitting.

Model Selection


Parameter Tuning
Tree specific
Regularization
General (Booster, Multithreading)
Built-in Cross Validation & Model Tuning
k-fold CV
GridSearchCV
RandomizedSearchCV
Extendibility
Regression
Binary & Multiclass classification
Recommendation & Ranking
Language API - Early Stopping
Interface compatibility
Python, Java, R , C++, Julia, Scala, Hadoop
Algorithm
Tree pruning
Sparsity Aware Split Finding (Handles missing data)
System
Parallelization
Cache Aware
Distributed and Scalable computing for large datasets.
High Flexibility
Custom objective function and custom evaluation metrics
Feature Importance Analysis
Ranks relative importance of each variable
XGBoost

XGBoost Regression

Implements an optimized gradient boosting algorithm, known for its efficiency and high performance on structured data.



LightGBM Regression

Utilizes gradient boosting with a focus on faster faster training speed and lower memory usage, ideal usage, ideal for large datasets.

Model Performance

Metrics/ /Model	XGBoost	LightGBM	Random Forest	Gradient Boosting	Decision Tree	SVR
MAE	30.64	31.00	32.66	33.82	45.99	35.67
MSE	2807.87	2859.47	3096.66	3202.57	5616.86	3942.38
RMSE	52.98	53.47	55.64	56.59	74.94	62.78
R^2 Score	0.72	0.71	0.69	0.68	0.44	0.61

Conclusion

1 Through rigorous model testing, XGBoost and LightGBM have emerged as our top-performing models. Achieved an R2 score of around 0.72, enhancing prediction accuracy for insurance pricing.

2 Leveraging structured data significantly aids in predicting property insurance prices. Enables smoother and more accurate predictions through organized information.

3 Obtaining appropriate datasets posed a significant challenge. Despite the abundance of datasets, not all encompassed all necessary aspects for comprehensive analysis.

4 To tackle dataset challenges, future improvements involve integrating property images property images uploaded by policyholders. We'll extract essential features from these from these images to enrich our structured data. Additionally, we'll introduce questionnaires or forms to capture missing data points.

5 Ultimately, we aim to seamlessly integrate image data and questionnaire responses to create a more comprehensive dataset.

6 These strategies aim to collectively refine and enhance the accuracy of property insurance price predictions.

Thank you..!!