Springboard Capstone Report

Prediction of Annual Flood Insurance Premium for Houston Area

Cheng Cheng, Data Science Track

1. Introduction

Flood can happen anywhere, especially in the coastal cities. However, most homeowner's insurance does not cover flood damage. Flood insurance is a separate policy to cover the water damage to the building and/or contents caused by flooding, so whether to buy a separate flood insurance is of great interest to the homeowners if their property has flooding risk. Houston has a long history of extreme rainfall events, including Hurricane Harvey in 2017, thus flood insurance is an important concern for the homeowners in Houston area as they decide whether to add flood insurance to the regular homeowner's insurance.

Prediction of the flood insurance premium for the houses and buildings in Houston area is the problem to be solved in this study. This information is beneficial for the homeowners to make risk-managing decisions, as well as for the insurance companies. Furthermore, it's of great interest to other stakeholders such as state and local government agencies, public institutions, and non-profit groups because flooding is often a public emergency issue.

2. Data Acquisition and Cleaning

The National Flood Insurance Program (NFIP), managed by the Federal Emergency Management Agency (FEMA), provides flood insurance to the property owners, renters, and business owners. Note that the actual insurance policies are issued by private insurance companies, not by the NFIP or FEMA. The program has the insurance policy dataset of more than 23,000 participating NFIP communities across the U.S., including Houston area, which is updated every 40 to 60 days (https://www.fema.gov/openfema-data-page/fima-nfip-redacted-policies-v1). The dataset used in this study was downloaded in May 2022, which includes the flood insurance policy information from 2009 to 2019.

The size of the downloaded dataset is about 12 GB. After the preliminary exploration of the dataset by filtering the insurance policy dataset for Houston area in 2018 and 2019, more than 300,000 records were obtained before data wrangling. Additional analysis was conducted to and focus was placed on the 2019 dataset for Houston area, which has more than 100,000 records and is enough for machine learning model evaluation.

The dataset has 45 variables including the total insurance premium of the policy, which is the target feature: "totalinsurancepremiumofthepolicy" (Table 1). Based on the initial evaluation and speculation, some features may have noticeable impacts on the policy premium, such as flood zone designation, building elevation, insurance coverage, number of floors, etc.

Table 1: All variables in the original NFIP dataset.

Variable	Explanation	Data type
agriculturestructureindicator	Indicator of whether a building is reported as being an agricultural structure	category
basefloodelevation	Elevation at which there is a 1% chance per year of flooding from the elevation certificate	number
basementenclosurecrawlspacetype	Basement is defined for purposes of the NFIP as any level or story which has its floor subgrade on all sides	category
cancellationdateoffloodpolicy	The cancellation date of the flood policy (if any), thereby goes out of force	datetime
censustract	US Census Bureau defined census Tracts; statistical subdivisions of a county or equivalent entity that are updated prior to each decennial census	number
condominiumindicator	Indicator of what type of condominium property is being insured	category
construction	Indicator of whether the building is under construction	category
countycode	FIPS code uniquely identifying the primary county	number
crsdiscount	The Community rating system (CRS) flood insurance policy premium discount	category
deductibleamountinbuildingcoverage	The total deductible amount for buildings, both main and appurtenant, that can be applied against the loss	category
deductibleamountincontentscoverage	The total deductible amount for contents in both main and apartment structures that can be applied against the loss	category
elevatedbuildingindicator	Indicator of whether a building meets the NFIP definition of an elevated building	category
elevationcertificateindicator	Indicates if a policy has been rated with elevation certificate	category
elevationdifference	Difference between the elevation of the lowest floor used for rating or the floodproofed elevation and the base flood elevation (BFE), or base flood depth, as appropriate from the elevation certificate	number
federalpolicyfee	Dollar amount of the Federal Policy Fee of the policy	number
floodzone	Flood zone derived from the Flood Insurance Rate Map (FIRM) used to rate the insured property	category
hfiaasurcharge	Congressionally mandated annual surcharge	number
houseofworshipindicator	Indicator of whether a building is reported as being a house of worship	category
latitude	Approximate latitude of the insured building	number
locationofcontents	Indicator of where the contents are located within the structure	category
longitude	Approximate longitude of the insured building	number
lowestadjacentgrade	The difference of the lowest natural grade adjacent to the building from the reference level of the building	number
lowestfloorelevation	A building's lowest floor is the floor or level that is used as the point of reference when rating a building	number
nonprofitindicator	Indicator of whether a building is reported as being a non-profit	category
numberoffloorsininsuredbuilding	Code that indicates the number of floors in the insured structure	category
obstructiontype	Code that gives the type of obstruction in the enclosure	category

Variable	Explanation	Data type
occupancytype	Code indicating the use and occupancy type of the insured structure	category
originalconstructiondate	The original date of the construction of the building	datetime
originalnbdate	The original date of the flood policy	datetime
policycost	Calculated in dollars by adding together calculated premium, reserve fund assessment, federal policy fee, and HFIAA surcharge	number
policycount	Insured units in an active status	number
policyeffectivedate	The effective date of the flood policy	datetime
policyterminationdate	Date upon which the cancellation of a flood insurance policy becomes effective	datetime
policytermindicator	Indicates length of time for which policy is in effect	category
postfirmconstructionindicator	Indicator of whether construction was started before or after publication of the FIRM	category
primaryresidenceindicator	Indicator of whether the insured building/condominium unit is the primary residence of the insured	category
propertystate	The state in which the insured property is located	string
reportedzipcode	5-digit Postal Zip Code for the insured property	number
ratemethod	Indicates policy rating method	category
regularemergencyprogramindicator	Identifies the phase of the NFIP in which a community is currently participating	category
reportedcity	The city in which the insured property is located	string
smallbusinessindicatorbuilding	Indicator of whether the insured represents a small business	category
totalbuildinginsurancecoverage	Total insurance amount on the building	number
totalcontentsinsurancecoverage	Total insurance amount on the contents	number
totalinsurancepremiumofthepolicy	Total insurance premium of the policy	number

3. Data Wrangling

There are a few issues during the process of data wrangling:

- The “reportedcity” feature for Houston area has several different values in the original dataset, such as “HOUSTON”, “HOUSTON TX”, “HOUSTON HARRIS”, “S HOUSTON”, “SOUTH HOUSTON”, etc. When filtering the dataset for Houston area, I selected the records with the “reportedcity” string which contains “Houston” and the “propertystate” string which is “TX”.
- There are a lot of missing values for some features, such that the missing values for the features of “cancellationdateoffloodpolicy”, “obstructiontype”, and “basefloodelevation” are more than 70% of the total amount. Some other features have fewer missing values. Additional data exploration was conducted to determine the approach to handle this issue:
  - The information provided by some features can be estimated using other features, or the information provided by some features was determined to be not important, these features were removed. For example, “cancellationdateoffloodpolicy” contains the cancellation date of the policy, but similar information can be found in

    the feature of “policyterminationdate” which has zero missing values, so “cancellationdateoffloodpolicy” was removed from the dataset. Another example is that the information of zip code and census tracts is not important, so it was removed as well.
    - The information provided by some features, such as “basefloodelevation”, may be important even though there are a lot of missing values, and these features were further investigated during the exploratory data analysis (EDA) to determine the method to handle the missing values.
    - Some other features have only a small number of missing values (e.g., <5% of total amount), and those records were deleted from the dataset given that the dataset has more than 100,000 records.
- There are extreme values for some numeric features. For example, the insurance premium price feature has values less than 10 and higher than 50,000 dollars, which seem too extreme. Also, the feature of “elevationdifference” has a lot of values of 999, which seems incorrect. This issue was addressed during the EDA.
- Overall, the dataset has a lot of categorical but relatively fewer numeric features. Due to that this is a regression problem, the numeric features are better kept if possible.

4. Exploratory Data Analysis

A few issues about the dataset were found during the process of data wrangling, some of them were already addressed but others were further investigated during the EDA.

4.1 Improve Dataset Selection

The feature of “policycount” indicates the number of insured units, and it has values from 1 to 402. However, the number of value 1 accounts for more than 99% of total amount of records, so I only kept the records with “policycount” of 1.

The features of “policyeffectivedate” and “policyterminationdate” have the policy starting and cancellation dates. According to these two features, more than 95% of the total amount of records have the policy for one year period, while others have shorter or longer periods. To better predict annual insurance premium, I only kept the records with the policy period of one year long.

The two features, “policycost” and “totalinsurancepremiumofthepolicy”, are very related based on the scatter plot shown in Figure 1. “policycost” is calculated by adding together the insurance premium price, reserve fund assessment, federal policy fee, and HFIAA surcharge. Therefore, if the premium price is chosen as the target, “policycost” can be deleted, as well as reserve fund assessment, remove federal policy fee, and HFIAA surcharge from the dataset.

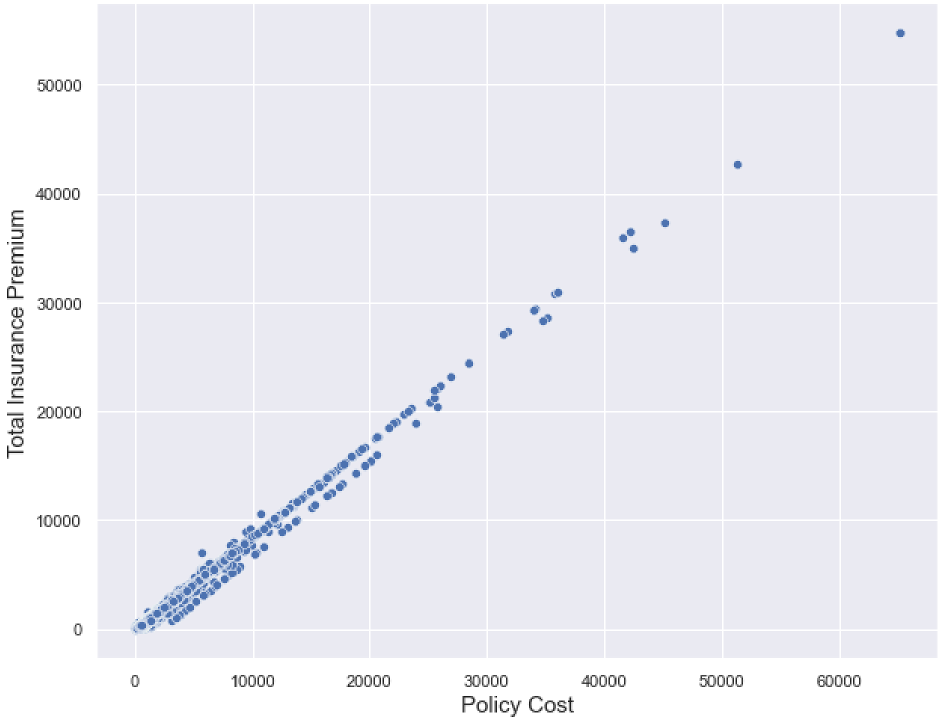


Figure 1. Relationship between the features of policy cost and total insurance premium

Additionally, the age of the building was calculated by subtraction the construction year of the building from year 2019. This is a new numeric feature, and it might be impacting the determination of insurance premium. Distribution of the building age values is shown in Figure 2.

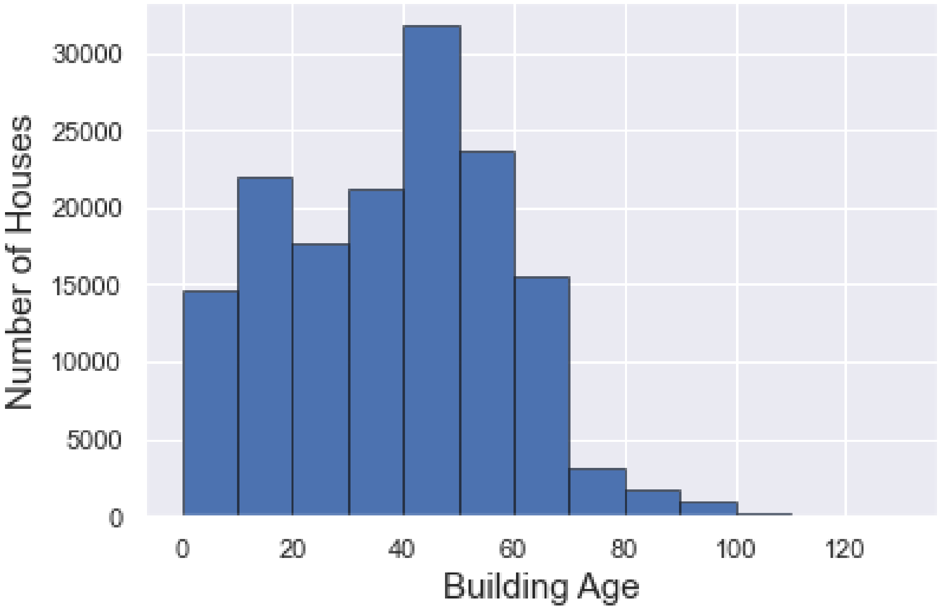


Figure 2. Distribution of building age

4.2 Handle Missing Values

For the numeric features with a lot of missing values, the correlations between the target (“totalinsurancepremiumofthepolicy”) and other numeric features were explored. The figure below shows the correlation heatmap. It shows that the correlation coefficient between the target and “basefloodelevation” is very small, as well as between the target and “lowestadjacentgrade”, and between the target and “lowestfloorelevation”. This suggests that these features may not provide much benefit in the prediction of the target. Moreover, these features have a lot of missing values (>70% of total amount), so they were removed from the dataset.

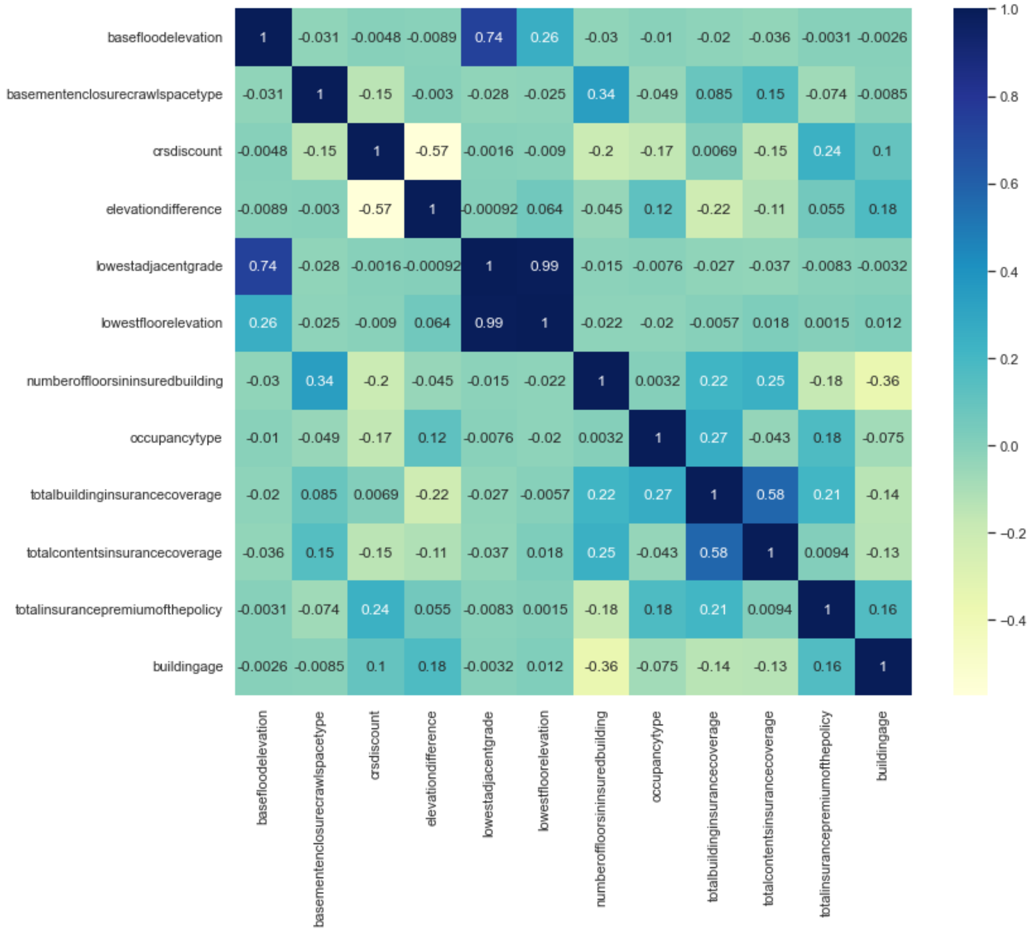
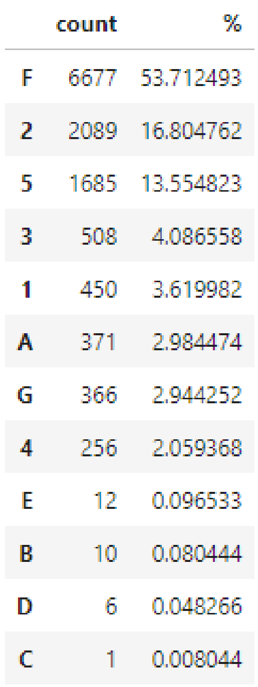


Figure 3. Correlation heatmap of numeric features

For categorical features, each feature was analyzed to determine how many unique values for each class in that feature. For the classes with less than 0.5% of total amount of records, they were removed from the dataset. For example, Figure 4 shows the count of each class in one categorical feature, and the classes of “E”, “B”, “D”, and “C” were removed from the dataset because they have negligible amounts of records. Removing them from the dataset might prevent any overfitting issue in the modeling.



	count	%
F	6677	53.712493
2	2089	16.804762
5	1685	13.554823
3	508	4.086558
1	450	3.619982
A	371	2.984474
G	366	2.944252
4	256	2.059368
E	12	0.096533
B	10	0.080444
D	6	0.048266
C	1	0.008044

Figure 4. Example of counts of different classes in a categorical feature

4.3 Handle Extreme Values

To handle the extreme values for some numeric features, such as the insurance premium, I used 95% confidence intervals to some of the features by keeping the records which are between 2.5 and 97.5 quantiles. This can remove some extreme outliers for those numeric features. For the numeric feature of “elevationdifference”, I used 99% confidence interval by keeping the records which is between 0.5 and 99.5 quantiles to keep a good number of records because this feature has some missing values.

4.4 Scale the Features with Several Orders of Magnitude

A new feature, called total coverage, was created by adding together the building insurance coverage and contents insurance coverage. This feature and the target (insurance premium) have a wide range value with more than several orders of magnitude, so logarithmic scale was conducted for these two features before the pre-processing and modeling.

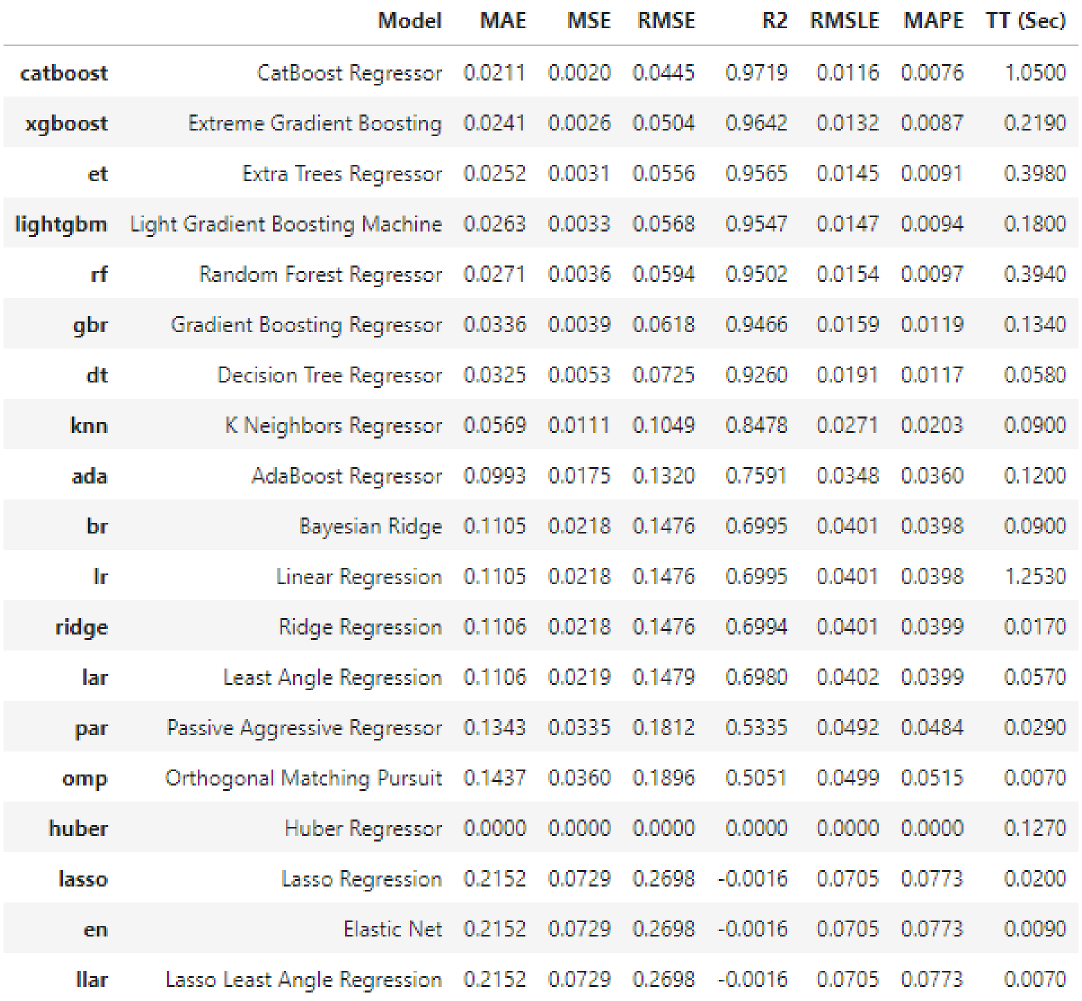
4.5 Encode Categorical Features

There are two types of categorical features: nominal and ordinal. For example, the number of floors is an ordinal categorical feature because its values are ordered, whereas flood zone is a nominal categorical feature without any intrinsic ordering. For nominal categorical features, dummy encoding was applied to encode them. For ordinal categorical features, they were specified different numeric values for different classes, and they were scaled before running the machine learning models.

5. Pre-processing

After data wrangling and exploratory data analysis, the dataset was finalized which has 12,332 records with 30 variables, including the target feature. Before any machine learning model was selected to evaluate the dataset, PyCaret was applied to determine the best models that give good performance for the dataset. PyCaret is an open-source and low-code machine learning library in Python that automates machine learning workflows.

Figure 5 shows the performance of multiple models to predict the flood insurance premium determined by PyCaret. Catboost is the best model given that it generates the best statistical values such as R2, MAE (mean absolute error), MSE (mean squared error), RMSE (root mean squared error), etc. Other good models include extreme gradient boosting, extra trees regression, light gradient boosting, random forest, decision tree regression, and K-nearest neighbors regression, which also provide very good statistical values.



	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
catboost	CatBoost Regressor	0.0211	0.0020	0.0445	0.9719	0.0116	0.0076	1.0500
xgboost	Extreme Gradient Boosting	0.0241	0.0026	0.0504	0.9642	0.0132	0.0087	0.2190
et	Extra Trees Regressor	0.0252	0.0031	0.0556	0.9565	0.0145	0.0091	0.3980
lightgbm	Light Gradient Boosting Machine	0.0263	0.0033	0.0568	0.9547	0.0147	0.0094	0.1800
rf	Random Forest Regressor	0.0271	0.0036	0.0594	0.9502	0.0154	0.0097	0.3940
gbr	Gradient Boosting Regressor	0.0336	0.0039	0.0618	0.9466	0.0159	0.0119	0.1340
dt	Decision Tree Regressor	0.0325	0.0053	0.0725	0.9260	0.0191	0.0117	0.0580
knn	K Neighbors Regressor	0.0569	0.0111	0.1049	0.8478	0.0271	0.0203	0.0900
ada	AdaBoost Regressor	0.0993	0.0175	0.1320	0.7591	0.0348	0.0360	0.1200
br	Bayesian Ridge	0.1105	0.0218	0.1476	0.6995	0.0401	0.0398	0.0900
lr	Linear Regression	0.1105	0.0218	0.1476	0.6995	0.0401	0.0398	1.2530
ridge	Ridge Regression	0.1106	0.0218	0.1476	0.6994	0.0401	0.0399	0.0170
lar	Least Angle Regression	0.1106	0.0219	0.1479	0.6980	0.0402	0.0399	0.0570
par	Passive Aggressive Regressor	0.1343	0.0335	0.1812	0.5335	0.0492	0.0484	0.0290
omp	Orthogonal Matching Pursuit	0.1437	0.0360	0.1896	0.5051	0.0499	0.0515	0.0070
huber	Huber Regressor	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.1270
lasso	Lasso Regression	0.2152	0.0729	0.2698	-0.0016	0.0705	0.0773	0.0200
en	Elastic Net	0.2152	0.0729	0.2698	-0.0016	0.0705	0.0773	0.0090
llar	Lasso Least Angle Regression	0.2152	0.0729	0.2698	-0.0016	0.0705	0.0773	0.0070

Figure 5. Best models to predict the insurance premium determined by PyCaret

6. Machine Learning Modeling

PyCaret provided several models which have good performance to predict the target feature using the dataset, and a few of them were selected in this section for further investigation.

6.1 Split and scale the dataset

Before any model evaluation was conducted, the dataset was split into two groups: train dataset and test dataset. The test dataset size was specified as 30% of total size.

All the features in the dataset have different units and data range, therefore standardization of the dataset was needed and beneficial for machine learning modeling. Except for the target feature and the dummy encoding categorical features, all other features were scaled using the standard scaler from scikit-learn library.

6.2 Modeling

Four different models, including K-nearest neighbors, random forest, extreme gradient boosting, and catboost, were selected to evaluate the dataset, and their performance to predict the insurance premium was analyzed and compared.

K-nearest Neighbors

GridSearchCV was implemented to fine tune the hyperparameters for K-nearest neighbors model, including number of neighbors, algorithm, weights, p value, and leaf size. Comparison of the predicted and actual insurance premium for K-nearest neighbors model is shown in Figure 6, which suggests that the model has a good performance with a R2 score of 0.9 and MAE value of 0.045. Note that MAE here represents the error after logarithmical scale, but even after converting it back to normal value, the actual MAE value is quite small: 10^0.045 = 1.1, which means that the estimation of insurance premium price is only ± 1 dollar compared to the actual premium price.

The learning curve with R2 score for the train and cross-validation datasets is shown in Figure 7, which suggests that R2 scores of both train and cross-validation datasets increase with the train set size and the variances for both datasets are very low. Overall, both datasets have low bias and low variance, indicating there is neither overfitting nor underfitting issue.

Random Forest

GridSearchCV was implemented to fine tune the hyperparameters for random forest regression model, including bootstrap indicator, number of estimators, max depth, and max features. Comparison of the predicted and actual insurance premium for random forest model is shown in Figure 8, which suggests that the model has a good performance with a R2 score of 0.95 and MAE value of 0.026. These statistical values are better than those for K-nearest neighbors model, indicating that random forest model performs slightly better than K-nearest neighbors model.

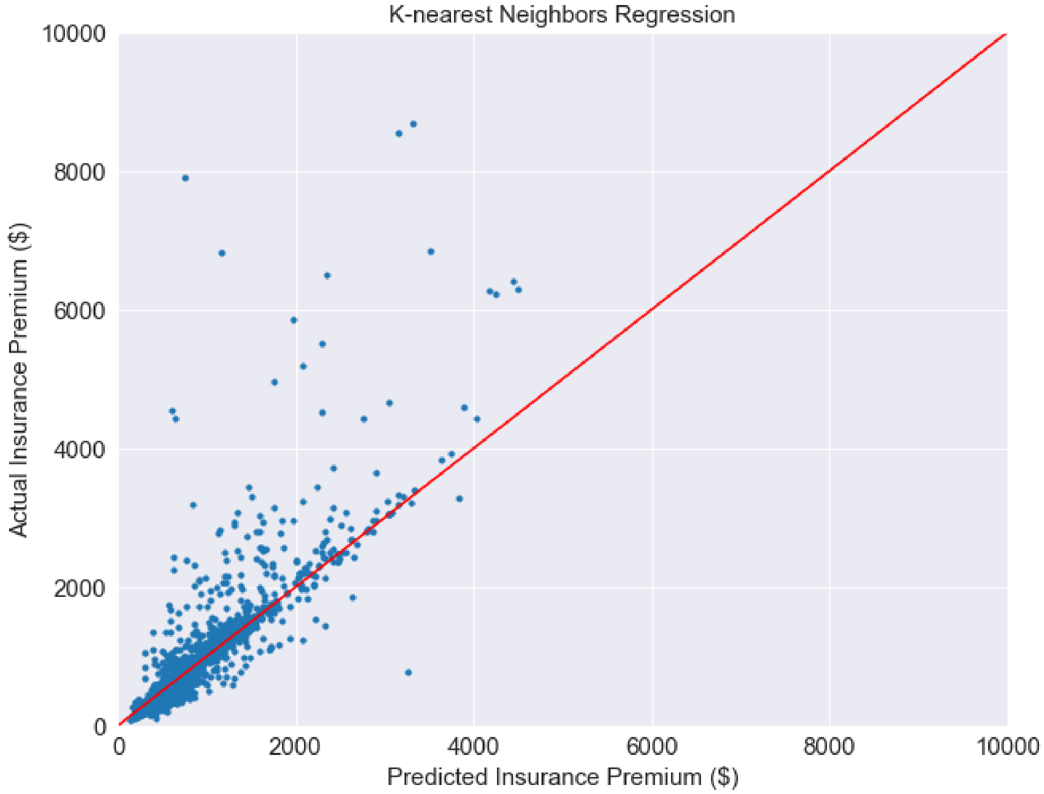


Figure 6. Predicted and actual insurance premium for K-nearest neighbors model

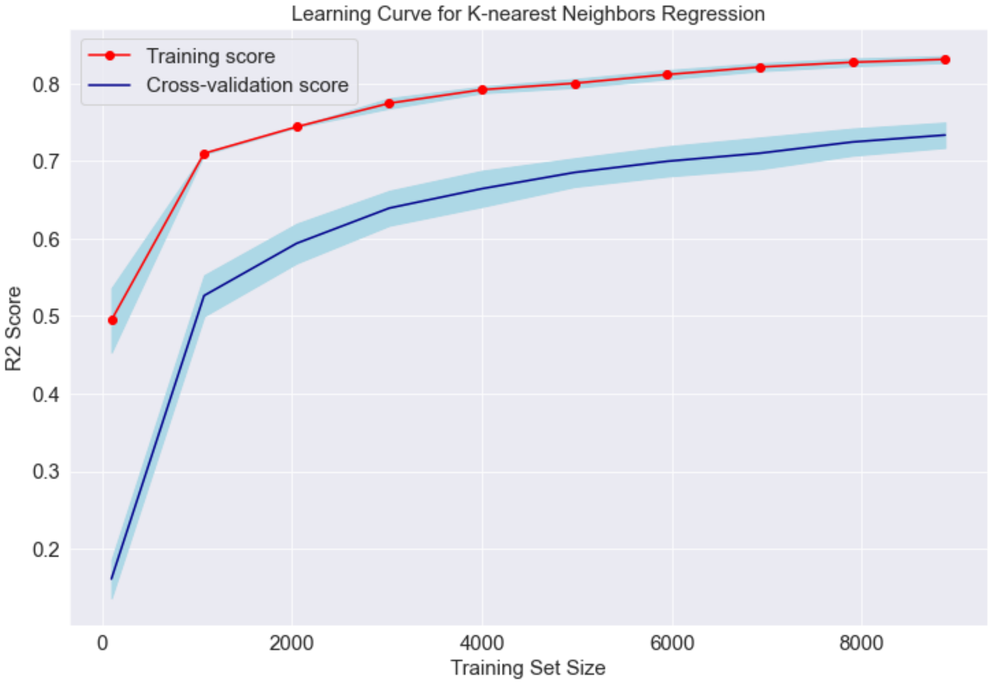


Figure 7. Learning curve with R2 score for K-nearest neighbors model

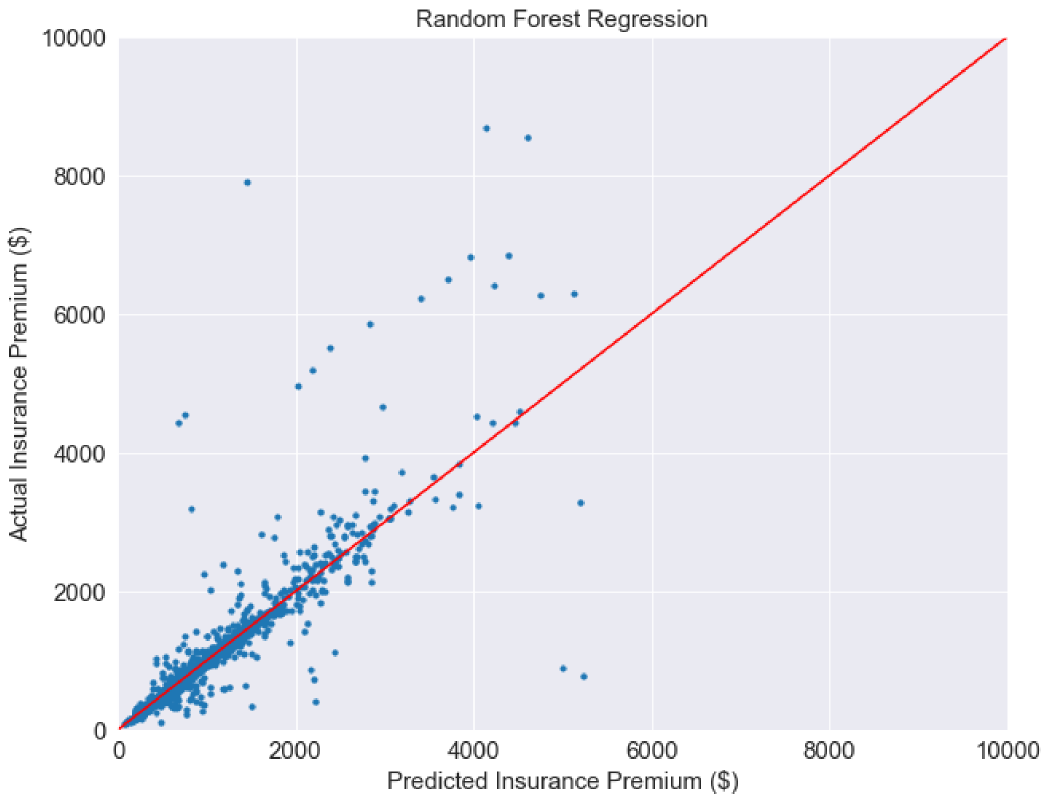


Figure 8. Predicted and actual insurance premium for random forest model

The learning curve with MSE score for the train and cross-validation datasets is shown in Figure 9, which suggests that mean squared error values decrease with the number of training set size for both training and cross-validation datasets, and they are converging to a very small value for both datasets. Overall, this very small mean squared error indicates that this model can make very good predictions for the target feature.

In addition, the feature importance function was applied to determine the top factors affecting the insurance premium price, which is shown in Figure 10. The top five factors are:

- Elevation difference: difference between the elevation of the lowest floor and the base flood elevation
- Insurance coverage: total insurance coverage for both building and contents
- Flood zone: NFIP specified flood zones used to rate the property
- Location of contents: the location where the contents are located within the structure
- CRS discount: the Community Rating System (CRS) flood insurance policy premium discount

Some features, such as occupancy type and rating method, seem not important for prediction of the insurance premium.

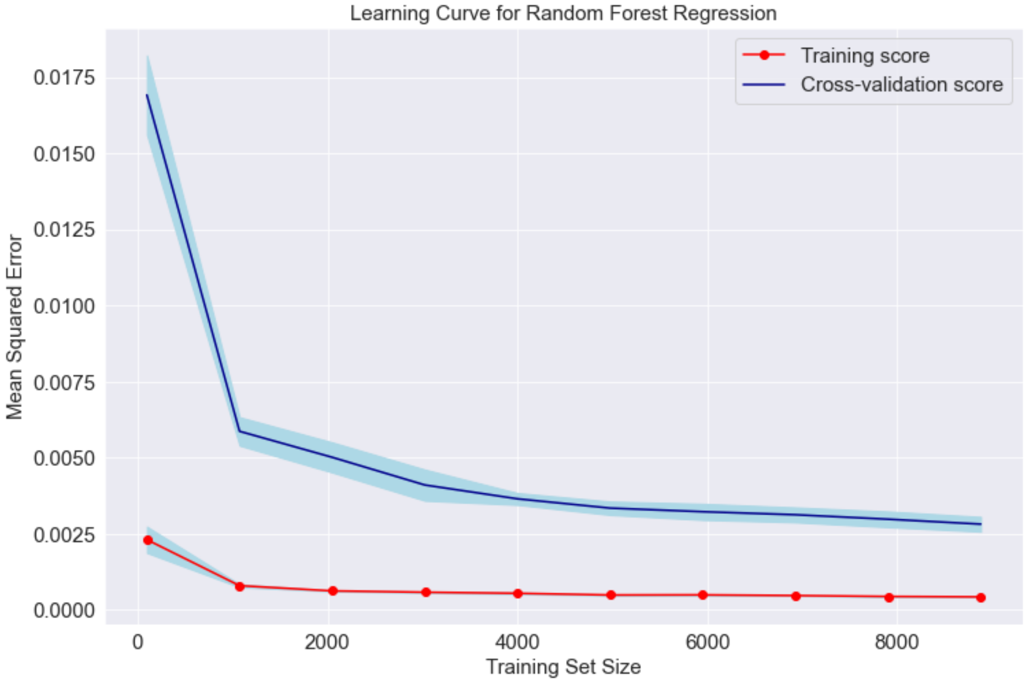


Figure 9. Learning curve with MSE values for random forest model

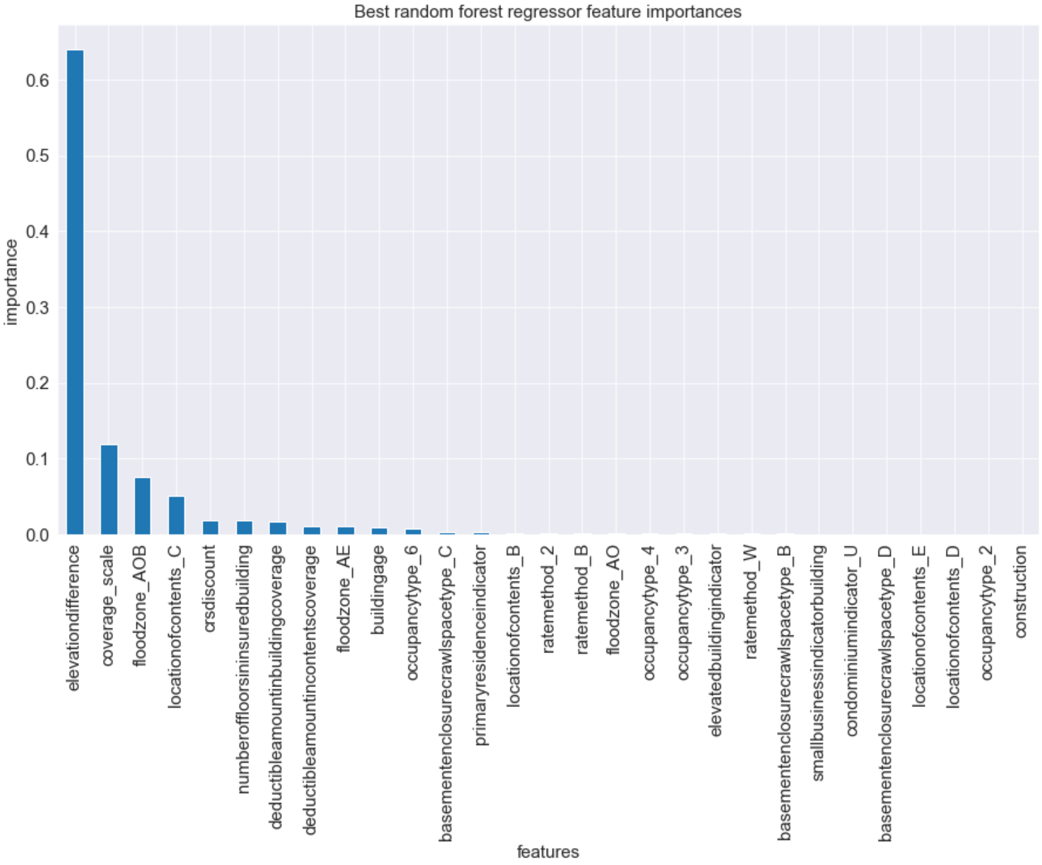


Figure 10. Feature importance for random forest model

Extreme Gradient Boosting

RandomizedSearchCV was implemented to fine tune the hyperparameters for extreme gradient boosting max depth, learning rate, subsample, colsample_bytree, colsample_bylevel, and number of estimators. Comparison of the predicted and actual insurance premium for extreme gradient boosting model is shown in Figure 11, which suggests that the model has a great performance with a R2 score of 0.97 and MAE value of 0.023. These statistical values are slightly better than those for random forest model, indicating that extreme gradient boosting model even performs slightly better.

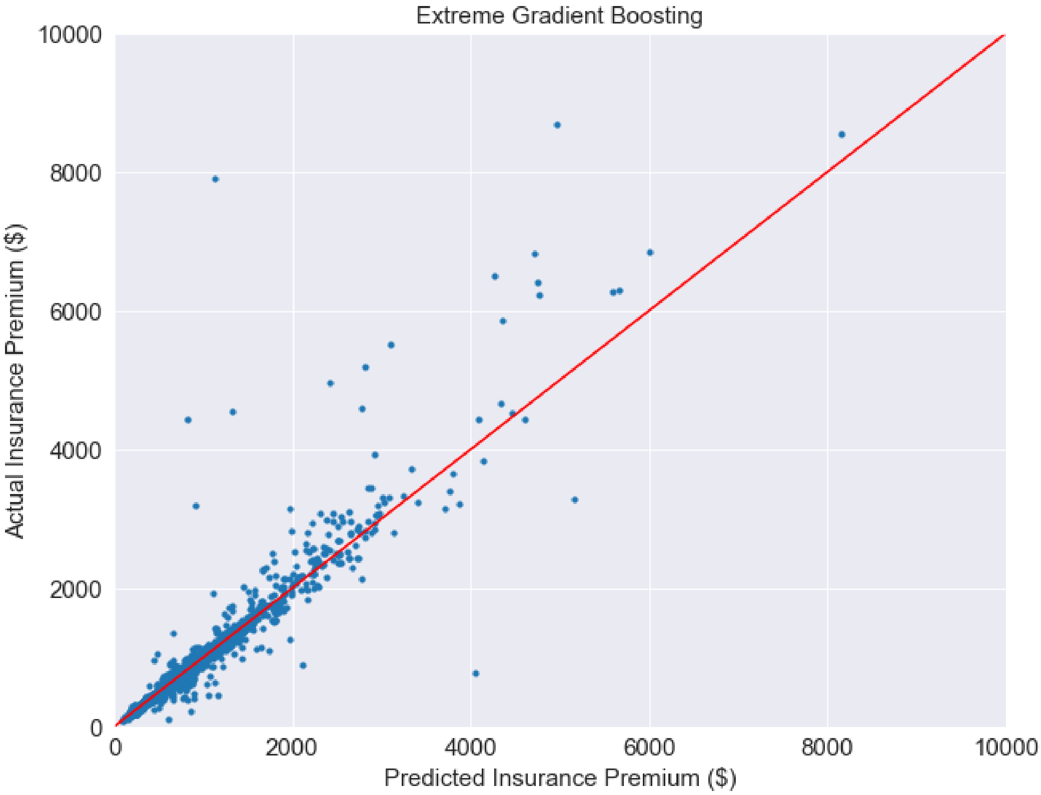


Figure 11. Predicted and actual insurance premium for extreme gradient boosting model

The learning curve with R2 score for the train and cross-validation datasets is shown in Figure 12, which suggests that R2 scores remain very high (close to 1) for the training dataset and increase with the train set size for the cross-validation dataset. Overall, R2 scores are very high for both datasets, indicating that this model can make very good predictions for the target feature.

In addition, the feature importance function was applied to determine the top factors affecting the insurance premium price, which is shown in Figure 13. The top five factors are:

- Elevation difference: difference between the elevation of the lowest floor and the base flood elevation

- Flood zone: NFIP specified flood zones used to rate the property
- Location of contents: the location where the contents are located within the structure
- Rate method: insurance policy rating method
- Elevated building indicator: indicator of whether a building meets the NFIP definition of an elevated building

The top five factors for the extreme gradient boosting model are quite different from those for the random forest model, except for the features of elevation difference and location of contents. In addition, more features have importance to some extent on prediction of the target feature in the extreme gradient boosting model.

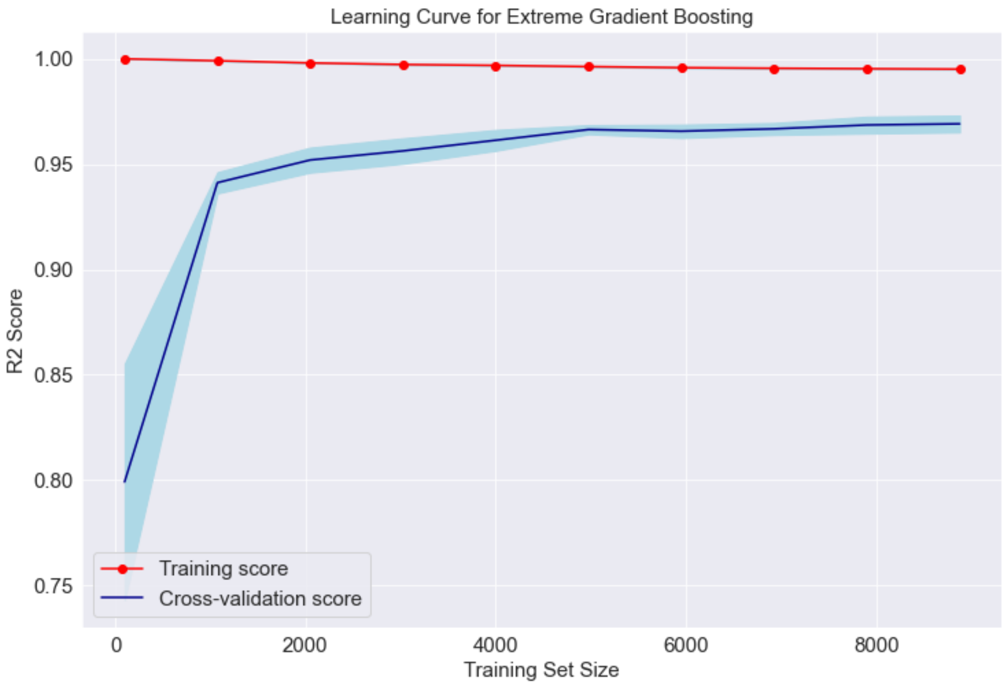


Figure 12. Learning curve with R2 scores for extreme gradient boosting model

Catboost

GridSearchCV was implemented to fine tune the hyperparameters for catboost model, including depth, learning rate, and iterations. Comparison of the predicted and actual insurance premium for extreme gradient boosting model is shown in Figure 14, which suggests that the model has a great performance with a R2 score of 0.97 and MAE value of 0.02. These statistical values suggest that the catboost model has a very high performance.

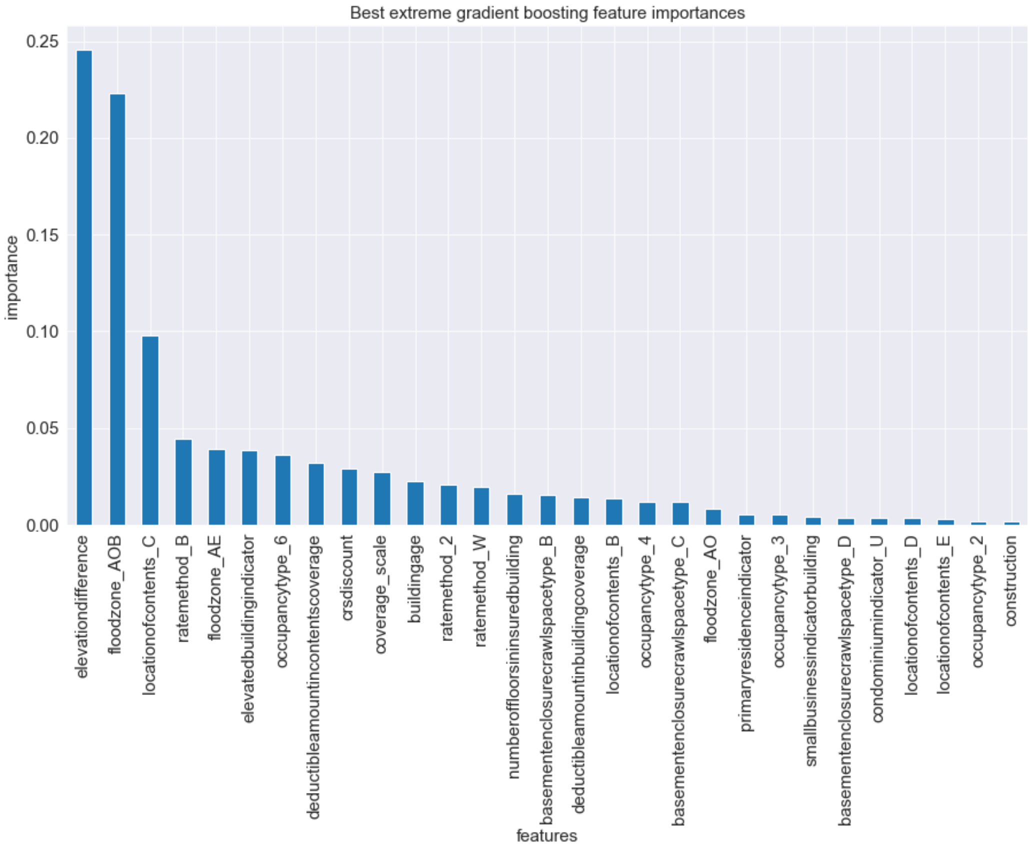


Figure 13. Feature importance for extreme gradient boosting model

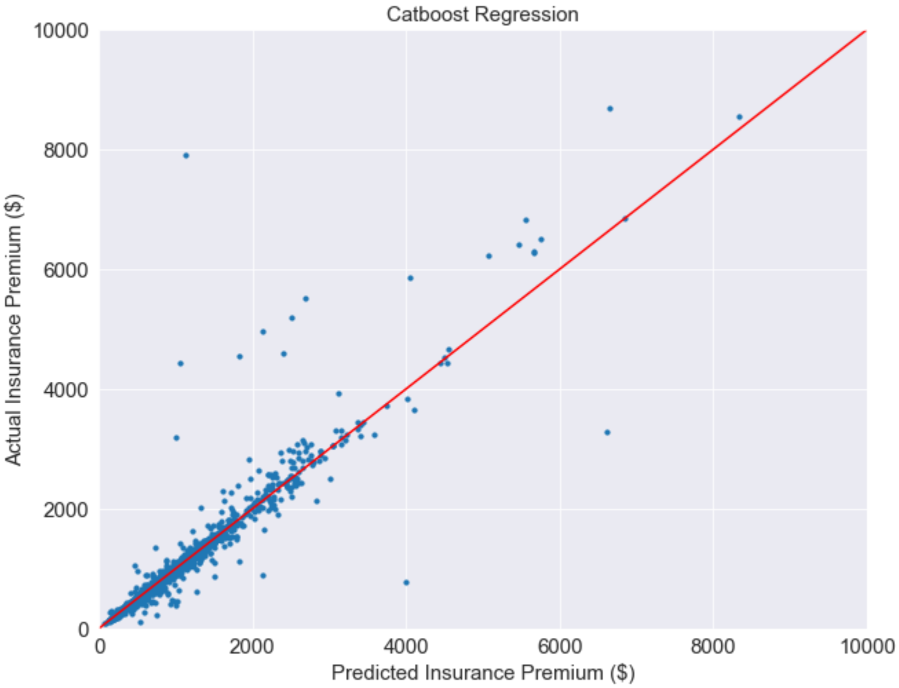


Figure 14. Predicted and actual insurance premium for catboost model

The learning curve with R2 score for the train and cross-validation datasets is shown in Figure 15, which suggests that R2 scores remain very high (close to 1) for the training dataset and increase with the train set size for the cross-validation dataset. Overall, R2 scores are very high for both datasets, indicating that this model can make very good predictions for the target feature.

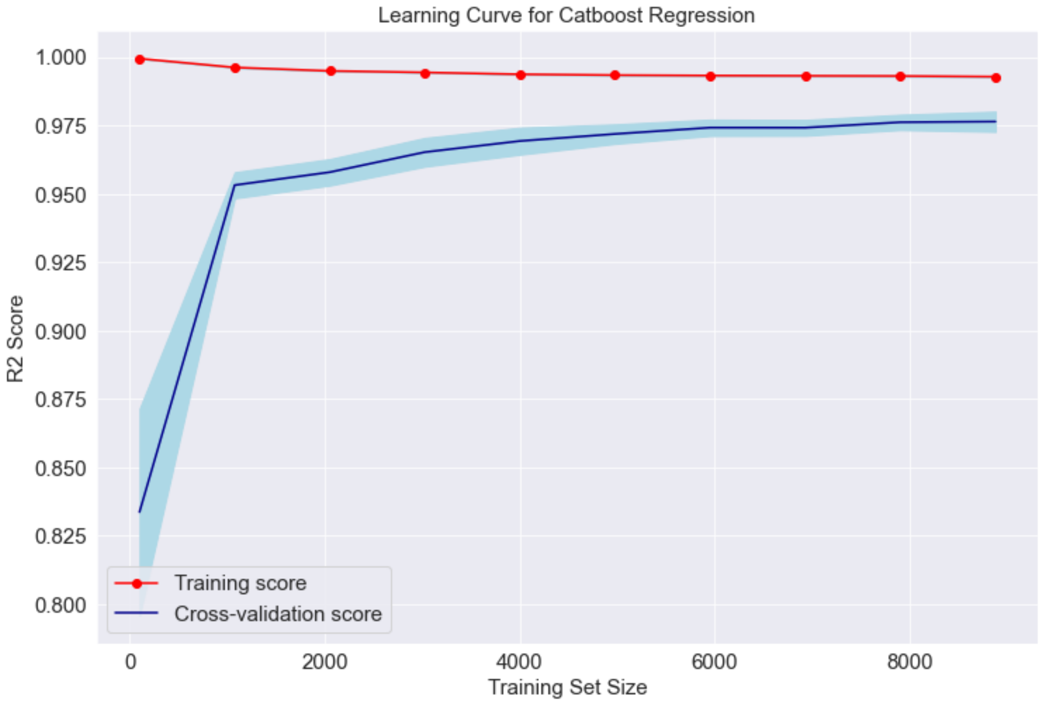


Figure 15. Learning curve with R2 scores for catboost model

In addition, the feature importance function was applied to determine the top factors affecting the insurance premium price, which is shown in Figure 16. The top five factors are:

- Elevation difference: difference between the elevation of the lowest floor and the base flood elevation
- Insurance coverage: total insurance coverage for both building and contents
- Flood zone: NFIP specified flood zones used to rate the property
- Location of contents: the location where the contents are located within the structure
- Number of floors: the number of floors for the insured structure

The top four factors are the same as those estimated for the random forest model, except for the 5th factor. It is number of floors for the catboost model whereas it is the CRS discount. Overall, it provides similarly important features as the random forest model, and the elevation difference is the most important factor affecting the insurance premium, which is determined by the random forest, extreme gradient boosting, and catboost models.

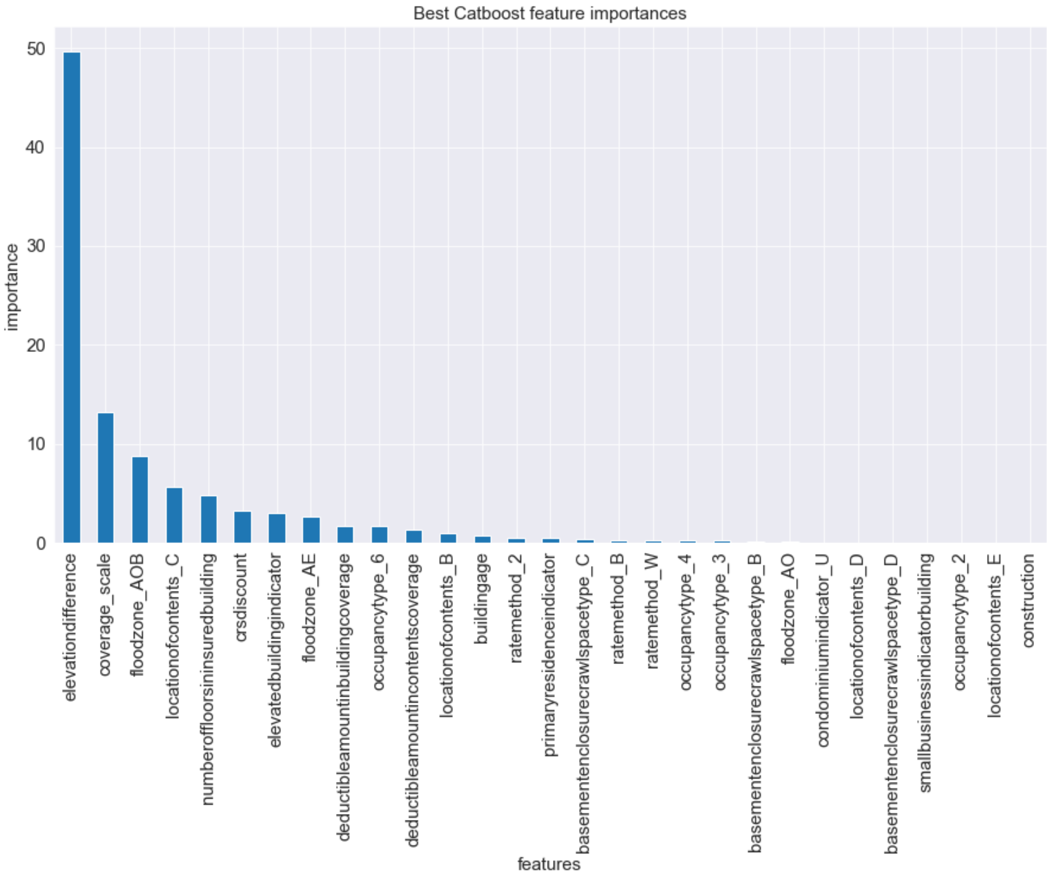


Figure 16. Feature importance for catboost model

7. Summary and Future Work

The 2019 data obtained from NFIP open dataset, managed by FEMA, was used to predict the flood insurance premium price for Houston area. Data wrangling and exploratory data analysis were conducted to clean, analyze, transform, and pre-process the dataset and to generate the final dataset for the modeling. Four machine learning models were selected to evaluate the dataset. All the models perform well. Among them, catboost regression, extreme gradient boosting, and random forest regression perform better than K-nearest neighbors regression given that those models generate better mean absolute error and R2 values. Among all features, elevation difference is determined to be the most important feature to predict the insurance premium. Other features, such as total coverage, certain flood zones, number of building floors, and location of contents, also play an important role in predicting the insurance premium.

The dataset has some limitations which may affect the modeling performance. There are missing, incorrect, and anomalous values in the dataset. Additionally, flood insurance premium is also impacted by other factors, such that whether the building has any submitted claims in the previous years; however, those factors are not in the dataset.

For the future work, ensemble modeling is preferred to combine multiple machine learning models in the prediction process, which offers a solution to overcome the technical challenges and limitations from a single estimator.

Data source: https://www.fema.gov/openfema-data-page/fima-nfip-redacted-policies-v1