# Know Your Fair Housing Price

Class B Group I

## 1 Executive Summary

Buoyed by unprecedented advancements in big data and continue growth in US housing market, Property Tech (PropTech) expands its application from information gathering to analytics (Crew, 2021; Voices & Xu, 2020). Seizing the development opportunity, we design a data-driven product based on Ames housing dataset retrieved from Kaggle. With the accuracy-based team ranking at the 399th place in Kaggle competition (Appendix 1), we develop a house transaction price calculator for home sellers and seller agencies to mitigate the inaccurate estimation on sale price caused by information asymmetry existing in the housing market. For sellers, they are not confirmed on the expected sale price due to lack of professional knowledge and psychological factors. For seller agencies, their estimation for sale price is not as precise as expected. For accurate prediction, we take house features, market trading volume, macroeconomic and housing policy into consideration and utilize a stacking model. This model takes advantage of all outperformed machine learning (ML) models to maximize the prediction performance, which excels other ML methods in this study. It is found that the stacking model gives 47.84% more accuracy for sellers and 38.72% for seller agencies. In particular, extremely optimistic and pessimistic sellers are potentially gain higher benefits from our model. Besides, overall quality, living area and construction year of the house are all the important features of four base learners, which is consistent with previous study. These findings also offer practical implications to technology and property practitioners.

## 2 Business Scenarios

The focal business is the second-hand property market in the United States. To sell a house, the seller will first list the house at a property agency, who uses their social network to approach potential buyers, or the buyers' agents. In the end, the realtors can earn commissions from both the buyers and sellers. Nowadays, there is a growing trend for sellers to list for free with online platforms such as Zillow.

Home sales are going on at a record pace in 2021, with prices increasing by almost 20% compared to 2020 (CNN, 2021). Additionally, there are around 10 million second homes in the US, and with the ongoing pandemic, people will enjoy more flexibility to work from home (Crew, 2021). Hence, we believe that there will continue to be a growing trend in second-hand home sales in the coming few years.

The dataset consists of sales records for the home market from 2006 to 2010 in Ames, Iowa. Based on the dataset, we build a machine learning model that can potentially be applied to the following scenarios. The mechanism of modelling can be applied on the future data.

## 2.1 Government

Firstly, our model is helpful for governments to decide the tax base of property tax. In the US, a fixed allowance or a certain percentage of tax exemption is granted according to the valuation of the property, which is levied in excess of this part. When the property tax rate remains the same, the more tax must be paid for the house with higher valuation.

Besides, our model can motivate the governments to improve macroeconomic policies and policies about housing price regulation. Our model will predict the objective fair price of the house, and the government can use the fair price to compare with the market price of the house and therefore regulate the house price and prevent the real estate market from being too cold or overheated.

## 2.2 Bank

Firstly, our model can improve the work efficiency of property appraisers. It is time-consuming for the appraisers to evaluate the value of properties. Our model can give them a reference value or a benchmark value. Based on this value, the appraisers only need to adjust the price according to their own experience, economic environment and local laws and regulations to get the final price, which significantly improves the work efficiency of the appraisers and the capacity of the bank.

Besides, our model can be used to help banks to determine real estate loan amount. Kalfrin & Feeney (2021) noticed that housing appraisal will directly affect the loan amount of bank customers, and banks must deal with a large number of loan applications every day. For this reason, banks need an accurate and objective housing valuation model to help them decide whether the loan application is approved or not and loan amount.

## 2.3 Individuals

Our model can be applied to buyers or sellers for housing transactions to build a reasonable “bottom line”. Whether it is a buyer or seller, whether it is the primary market or the secondary market, for most individuals, it is difficult for them to have the ability to predict the fair price of the house due to information asymmetry. Hence, individuals usually lose their bargaining power when conducting house transactions or difficult to judge whether the price recommended by the property agents is a fair price. Our model can give them a reference price, to rationalize the psychological price of individual and improve their bargaining power in housing transactions.

## 2.4 Insurance Company

Our model can help insurance companies to objectively measure the amount of the subject matter insured in housing insurance and protect the interests of both insurance companies and customers. When insurance companies and customers conduct housing insurance transactions, it is easy for them to have contradictions on

the housing value (Roberte, 2021). A neutral and professional third party is needed to evaluate the objective fair value of the housing.

# 3 Conflicts

We mainly focus on the conflict where individual sellers and/or seller agencies cannot make a good estimate of the selling price for their houses.

For house sellers, their goal is to maximize the proceeds from selling the house under the market conditions within a target timeframe. Home sellers should figure out an appropriate listing price. If the seller is pessimistic about selling house, it is very likely that the seller may price it emotionally and significantly lower than the market price, thus losing a big chunk of potential profit. Another scenario is when the seller is optimistic about house sale and would like to set a higher price to maximize the profit, which is less attractive to potential buyers and the house remains unsold. To avoid over- or underpricing, a rule of thumb is to use the average listed price of all second-hand houses in its neighborhood as a reference. However, sellers may miss some unique features of their house that have potential to outperform their neighbors.

If the house sellers decide to hire a property agent, the house for sale will be exposed to more potential buyers, but it will incur some extra commissions and sellers are prone to information asymmetry between the sellers and property agents (Lopez, 2021). A traditional way for a property agent to price a house is to take an average of either those similar houses that are listed at the same time, or those that have been sold recently. These two methods are biased because they only take available reference houses into consideration. It is not an ideal method especially when the house has some special features that are rare on the market.

Our model considers different factors other than neighborhood. It returns a mean value for a house in the market, and sellers can set the final price tag based on whether it is urgent or not to make a deal.

# 4 Data Exploration and Preprocessing

## 4.1 Exploratory Data Analysis

Package *sweetviz* package is applied to do EDA to see the distribution, association and relationship between variables (Appendix 2). Some data quality issues like right skewed distribution problem and missing data problem are found and solved during data preprocessing.

## 4.2 Data Enrichment

Since external factors also affect the house pricing, three variables *Sentiment_t-1, Unemio, Mortgage* are enriched into the dataset (Table 1).

**Table 1** Data enrichment

| Enriched Variable | Definition |
|---|---|
| *Sentiment_t-1** | Proxy for house market trading volume |
| *Unemio* | Monthly total unemployment |
| *Mortgage* | Monthly national average contract mortgage rate |

*Sentiment_t-1 indicates the market sentiment in the last period.

## 4.3 Data Preprocessing

### 4.3.1 Dealing with missingness

1) Missingness at random

**Table 2** Imputation methods

| Methods | Variable Type | Examples |
|---|---|---|
| Mode | Categorical variable | *GarageCars, SaleType* |
| Mean | Numerical variable | *MasVnrArea, LotFrontage* |

2) Informative missingness

For variables like *ExterQual* and *GarageType* the missing value means there is no basement or garage in the specific house, thus we impute *NA* here with *None*. Thus, new variables *HasBsmt* and *HasGarage* are created to save this information rather than simply do label encoding.

### 4.3.2 Feature filtering

To increase the generalizability, interpretability and computation efficiency of the model, three kinds of variables are eliminated as follows.

**Table 3** Feature filtering

| Enriched Variable | Examples |
|---|---|
| Near-zero variance variables | *RoofMatl, Heating, Functional* |
| Variables have low correlation with saleprice | *YrSold, MoSold* |
| Multicollinear Variables | *GarageYrBlt* |

### 4.3.3 Dealing with too few observations

We perform lumping for some categorical variables, such as, *MSSubClass*, *LotShape* and *HouseStyle*.

### 4.3.4 Transform categorical variables into numeric forms

**Table 4** Encoding

| Enriched Variable | Examples |
|---|---|
| Label encoding | *ExterCond, BsmtQual, BsmtCond* |
| Dummy encoding | *LotShape, LotConfig, Neighborhood* |

#### 4.3.5 Dealing with skewness, outliers and wide range.

1) We take *log* on *SalePrice* to transform skewed *SalePrice* to approximately conform to normality.

2) We also use the *log(x+1)* transformation on the other numerical features like *LotFrontage*, *MasVnrArea* and *WoodDeckSF*.

3) We do standardization.

4) We recheck the association between variables.

# 5 Machine Learning Questions & Solution

## 5.1 Machine Learning question formulation

We built several machine learning models for regression to predict the sales price of residential homes in Ames, Iowa, US. The data set consists of around 1500 records for training and another 1500 records for testing, describing multiple aspects including house features, market trading volume, macroeconomic and policy factors in the given period. The log transformation of *SalePrice* is served as the response variable in our experiments, and Kaggle's score uses Root Mean Squared Error (RMSE) to evaluate the model performance. The formulation is described as follow.

$$\min(RMSE),\ where$$
$$RMSE = \sqrt{(\log(predicted\ SalePrice) - \log(true\ SalePrice))^2}.$$

## 5.2 Machine Learning methods

Multiple machine learning regression models are built, compared, and finally selected for stacking. In particular, we apply 5-fold cross-validation to tune the hyperparameters.

### 5.2.1 Multiple Linear Regression

Multiple linear regression, including lasso, ridge, and polynomial regression with elastic net are employed in our experiments. EDA result shows that the feature "*sentiment_t-1*" has a polynomial effect (Appendix 3) on "*SalePrice*". Hence, we incorporate feature "*sentiment_t-1_sqrt*" and "*sentiment_t-1_cube*" into our models. These new features are calculated as follows.

$$sentiment\_t - 1\_sqrt = sentiment\_t - 1^2,$$

$$sentiment\_t - 1\_cube = sentiment\_t - 1^3.$$

#### 5.2.2 Principal Component Regression (PCR)

We implemented principal component regression by combining principal component analysis (PCA) and linear regression due to the limitation of the Apache Spark package, and around the first 100 principal components are selected after PCA for linear regression.

#### 5.2.3 K-nearest Neighbors Regression (KNN)

KNN regression model estimates the association between different houses and predicts "*SalePrice*" by averaging those observed values within their neighborhood.

#### 5.2.4 Support Vector Machine Regression (SVM)

Three SVMs with linear, polynomial, and radial basis function (RBF) kernel are employed and compared in our experiments to capture non-linear relationship between house features and sales price.

#### 5.2.5 Tree-Based Methods

The tree-based methods including decision tree, random forest, Gradient Boosting Machine (GBM), and XGBoost are used for advanced sales price estimation. To improve the interpretability, feature importance is extracted from these tree-based methods for further explanation.

#### 5.2.6 Multilayer Perceptron (MLP)

Given the relatively small and simple data set, we build a feedforward multilayer perceptron with two

hidden layers to avoid overfitting, and hyperbolic tangent activation function and adaptive learning rate are employed to achieve better performance.

#### 5.2.7 Stacking

The stacking model is an ensemble algorithm of some strong models in our prior experiments. Given, PCR, random forest, GBM, and XGBoost are served as base-models considering consistent input and high performance, we choose linear regression as meta-model, as complex models may lead to overfitting.

### 5.3 Benchmark Models Description and Overall Performance

#### 5.3.1 Benchmark Models for Seller

When looking for an opportunity to sell houses, house owners assess the value of their property through the comparative market analysis (CMA) (Folger, 2022). We assume that the most important factors for them are physical locations (*Neighborhood*) and building types (*BldgType*). According to these two features, the houses in the city can be divided into different groups. Then the average unit house price for each group can be calculated. The total value of the property is equal to average unit house price times the floor space. Because

of personality differences between different sellers, an optimist and a pessimist may make different estimates of the value of the same property. Therefore, a coefficient is added to illustrate this variable (pessimistic: 0.8; slightly pessimistic: 0.9; common: 1.0; slightly optimistic: 1.1; optimistic: 1.2). Besides, an assumption is made that pessimistic, slightly pessimistic, common, slightly optimistic and optimistic home sellers' proportions are 0.05, 0.15, 0.6, 0.15, 0.05 for calculating models' improvements compared to home sellers' benchmark, as well as overall home seller benchmark model performance (Appendix 4).

### 5.3.2 Benchmark Models for Seller Agency

Real estate agents apply CMA to predict sale price of house. Compared with home sellers, they have more access to market information and better grasp of various related factors, therefore make assessment more precisely. To better simulate the prediction process, KNN is used in this paper to represent the prediction made by agents. KNN can predict the category of house based on other houses with multiple similar house features, which is more accurate than simple method used by home sellers. Considering seller agents can capture the market trend in some degree, market sentiment (*Sentiment_t-1*) is incorporated in the prediction. The assumption is that *1+0.1* (standardized market sentiment)* is the coefficient of the influence factor.

### 5.3.3 Overall Performance Evaluation

The following table reports the overall performance of our machine learning models with data enrichment, and all are compared with benchmark models illustrated above. Stacking achieves 47.84% and 38.72% improvements beyond benchmark home seller and agency, respectively, which demonstrates the effectiveness of our model.

**Table 5** Overall performance evaluated by RMSE

| Model | RMSE | Improvement (Home Seller) | Improvement (Agency) |
|---|---|---|---|
| **Seller Benchmark** | **0.23498** | - | - |
| **Seller Agency Benchmark** | **0.20002** | 14.88% | - |
| **Stacking** | **0.12257** | **47.84%** | **38.72%** |
| Multiple Linear Regression | 0.13232 | 43.69% | 33.85% |
| PCR | 0.13226 | 43.71% | 33.88% |
| KNN | 0.17275 | 26.48% | 13.63% |
| SVM-RBF | 0.12821 | 45.44% | 35.90% |
| Decision Tree | 0.19527 | 16.90% | 2.37% |
| Random Forest | 0.14207 | 39.54% | 28.97% |
| GBM | 0.13086 | 44.31% | 34.58% |
| XGBoost | 0.12759 | 45.70% | 36.21% |

| | | | |
|---|---|---|---|
| MLP | 0.14479 | 38.38% | 27.61% |

# 6 Findings and Discussion

## 6.1 Findings for Features and Application

### 6.1.1 Feature Importance

Since the stacking model performs best, we use it as our final model, and the analysis of feature importance is also based on the base learners of stacking including GBM, PCR, Random Forest and XGBoost. We focus on the Top 10 important features of each model then also consider the ranking of the importance of features in each model, we set the weight of the feature that is most importance in each model to 10, the weight of the feature with the second important to 9, and so on.

**Table 5** Feature importance

| | Weight in models | | | | |
|---|---|---|---|---|---|
| | GBM | PCR | RF | XGBoost | Total |
| **OverallQual** | 10 | 10 | 10 | 8 | **38** |
| **GrLivArea** | 9 | 2 | 9 | 9 | **29** |
| **YearBuilt** | 5 | 4 | 6 | 7 | **22** |
| **LotArea** | 4 | - | 3 | 10 | **17** |
| **ExterQual** | - | 8 | 8 | - | **16** |
| **GarageCars** | - | 9 | 5 | - | **14** |
| **KitchenQual** | - | 6 | 7 | - | **13** |
| **TotalBsmtSF** | 8 | 3 | - | - | **11** |
| **1stFlrSF_log** | - | - | 4 | 6 | **10** |
| GarageFinish | - | 7 | - | - | 7 |
| BsmtFinSF1 | 7 | - | - | - | 7 |
| OverallCond | 6 | - | - | - | 6 |
| BsmtUnfSF_log | - | - | - | 5 | 5 |
| BsmtFinSF1_log | - | - | 2 | 3 | 5 |
| BsmtQual | - | 5 | - | - | 5 |
| LotFrontage_log | - | - | - | 4 | 4 |
| GarageCond | 3 | - | - | - | 3 |
| Unemio | - | - | - | 2 | 2 |
| CentralAir_N | 2 | - | - | - | 2 |
| FullBath | - | 1 | - | - | 1 |
| FireplaceQu | - | - | 1 | - | 1 |

| | | | | | |
|---|---|---|---|---|---|
| Sentiment_t-1 | - | - | - | 1 | 1 |
| 1stFlrSF | 1 | - | - | - | 1 |

As we can see from the above table, the three most important features are also *OverallQual*, *GrLivArea* and *YearBuilt*. Besides, *LotArea*, *ExterQual*, *GarageCars KitchenQual*, *TotalBsmtSF*, and *1stFlrSF_log* are also relatively importance.

In conclusion, the factors that have the greatest influence on the price of the house are the overall quality of the house, living area and age of the building. In addition, the overall external quality of the house, garage, kitchen and basement have a great impact on housing price.

### 6.1.2 Application Performance and Usage

From previous analysis, it is found that our model achieves 47.84% and 38.72% improvements beyond benchmark home seller and agency, respectively. Furthermore, it is found that the model is more beneficial to home sellers with more extreme personality (Table 7).

**Table 7** Model Performance for Different Scenario

| Stakeholders | Home Seller with Different Personality | | | | | Seller Agency |
|---|---|---|---|---|---|---|
| | Pessimistic | Slight Pessimistic | Common | Slight Optimistic | Optimistic | |
| Kaggle RMSE | 0.30612 | 0.23985 | 0.22062 | 0.24452 | 0.29292 | 0.20002 |
| Improvements | 59.96% | 48.90% | 44.44% | 49.87% | 58.16% | 38.72% |

Thanks to the superiority of our product, home sellers know the fair price of a house in a more accurate way, therefore avoid potential loss caused by information asymmetry and make a transaction in a short period of time. In the long run, our model is helpful for the government to maintain the order of the real estate brokerage market and reduce the occurrence of intermediaries using information asymmetry to mislead buyers and sellers. We expect that our product can gain the investment from venture capital because of our outstanding product performance and future potential growth.

## 6.2 Limitation

When building the model, due to lack of data with at most five years, time series factors cannot be considered. Besides, in data preprocessing, we delete some near-variance variables, which may be potentially important, causing the missing information and accuracy to our model.

Besides, our model assumes that the house features dominate the price, other than the features we enrich, since the original dataset focuses more on the house features, which causes that the marginal effect of enriched feature is minor.

### 6.3 Further Improvement

The limitations on data enrichment should be addressed by collecting more data in terms of instance and attributes. For more instance, since we only have a dataset with nearly 1,500 houses, to improve the accuracy and generalizability of our model, more houses should be included, for example, some houses in other cities in Iowa. For more attributes, more information of the house is needed, such as the listing date and the decoration cost. In addition, the accurate location of the house is also necessary, since buyers focus more on the environment around their houses, including but not limited to the infrastructure, education and noise.

The algorithm-based adjustment will be performed to mitigate the minor marginal effect. First, we will introduce more features related to market sentiment, policy and macroeconomics. In addition, we will adjust the weight of features by direct cross-validation or ensemble models to strike balance between the effect of house features and external factors.

## 7 Reference


Bahney, A. (2021, December 28). *The housing market was on a wild ride this year. here's what to expect in 2022*. CNN. Retrieved February 13, 2022, from https://edition.cnn.com/2021/12/27/homes/us-real-estate-market-2021/index.html

Crew, P. (2021, October 6). *PACASO Second Home Market Report - summer 2021*. Pacaso. Retrieved February 13, 2022, from https://www.pacaso.com/blog/second-home-market-report

Folger, J. (2022, February 8). Comparative market analysis. Investopedia. Retrieved February 15, 2022, from https://www.investopedia.com/terms/c/comparative-market-analysis.asp

Kalfrin, V., & Feeney, C. (2021, December 16). *From the outside in: What do home appraisers look for in a house?* HomeLight Blog. Retrieved February 15, 2022, from https://www.homelight.com/blog/what-appraisers-look-for-in-a-house/

Lopez, L. A. (2021). Asymmetric information and personal affiliations in brokered housing transactions. *Real Estate Economics, 49*(2), 459-492.

Roberte, L. (2021, December 7). *What is an appraisal for homeowners insurance?* The Balance. Retrieved February 15, 2022, from https://www.thebalance.com/insurance-appraisals-5104944

Voices, V., & Xu, L. (2020, July 21). *Modernizing real estate: The Property Tech Opportunity*. Forbes. Retrieved February 13, 2022, from https://www.forbes.com/sites/valleyvoices/2019/02/22/the-proptech-opportunity/?sh=5b7e18055826

*Get started here*: Data Sets | Federal Housing Finance Agency. (n.d.). Retrieved February 15, 2022, from https://www.fhfa.gov/DataTools/Downloads

Iowa Workforce Development, L. M. I. D. (2022, January 26). Iowa unemployment insurance recipiency rate (quarterly). Retrieved February 15, 2022, from https://data.iowa.gov/Workforce/Iowa-Unemployment-Insurance-Recipiency-Rate-Quarte/r6dy-q2sw

Google. (n.d.). Google trends. Retrieved February 15, 2022, from https://trends.google.com/trends/?geo=US

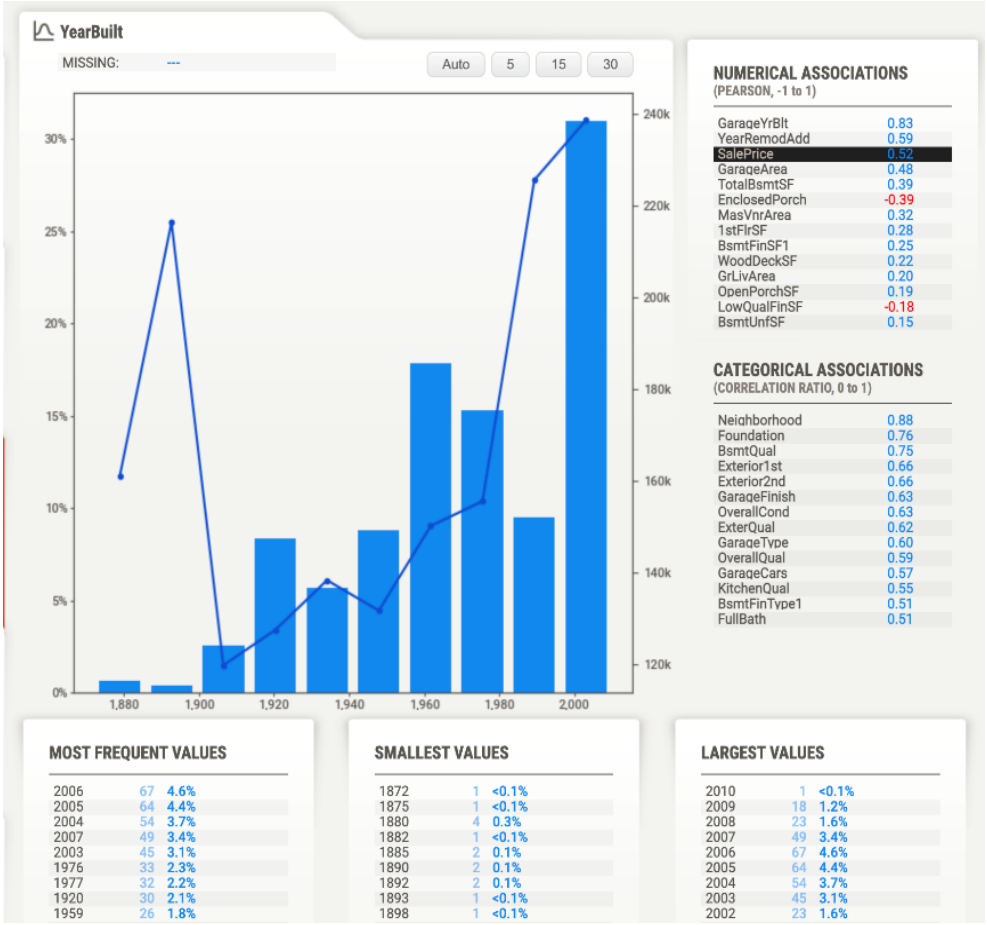
# Appendix

1. Rank on Kaggle

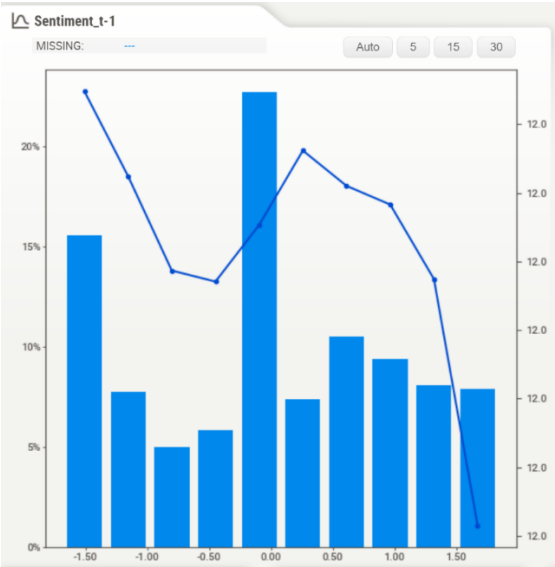
| | | | | | |
|---|---|---|---|---|---|
| 399 | **Nirvana1010** | | 0.12257 | 9 | 3d |

Your Best Entry!
Your most recent submission scored 0.12257, which is the same as your previous score. Keep trying!

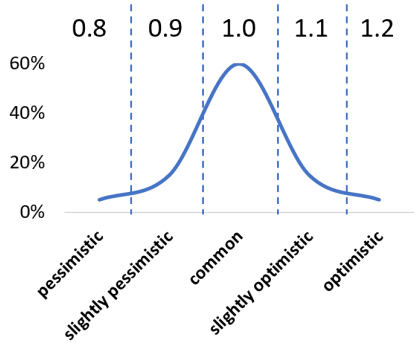
2. EDA with package *sweetviz*



3. Distribution of *Sentiment_t-1* and relationship with *SalePrice*




4. Personality Coefficients




5. Model performances and comparison with and without data enrichment

Several experiments are conducted to evaluate our models' performance (Table 8). House properties with and without macroeconomic factors (data enrichment) are fitted in our models to demonstrate the effectiveness of enrichment data. Multiple linear regression with enrichment data achieves 1.4% improvement when comparing the same model without data enrichment, highlighting the significance of data enrichment.

**Table 8** Overall performance evaluated by RMSE between different data set

| Model | RMSE | Improvement (Home Seller) | Improvement (Agency) |
|---|---|---|---|

| | | | | |
|---|---|---|---|---|
| Benchmark | **Home Seller** | **0.23498** | - | - |
| | **Agency** | **0.20002** | **14.88%** | - |
| With Data Enrichment | Multiple Linear Regression | 0.13232 | 43.69% | 33.85% |
| | PCR | 0.13226 | 43.71% | 33.88% |
| | KNN | 0.17275 | 26.48% | 13.63% |
| | SVM-RBF | 0.12821 | 45.44% | 35.90% |
| | Decision Tree | 0.19527 | 16.90% | 2.37% |
| | Random Forest | 0.14207 | 39.54% | 28.97% |
| | GBM | 0.13086 | 44.31% | 34.58% |
| | XGBoost | 0.12759 | 45.70% | 36.21% |
| | MLP | 0.14479 | 38.38% | 27.61% |
| | **Stacking** | **0.12257** | **47.84%** | **38.72%** |
| Without Data Enrichment | Multiple Linear Regression | 0.13555 | 42.31% | 32.23% |
| | PCR | 0.13232 | 43.69% | 33.85% |
| | KNN | 0.17098 | 27.24% | 14.52% |
| | SVM-RBF | 0.13197 | 43.84% | 34.02% |
| | Decision Tree | 0.2128 | 9.44% | -6.39% |
| | Random Forest | 0.14168 | 39.71% | 29.17% |
| | GBM | 0.13042 | 44.50% | 34.80% |
| | XGBoost | 0.12727 | 45.84% | 36.37% |

| | | | | |
|---|---|---|---|---|
| | MLP | 0.13564 | 42.28% | 32.19% |
| | **Stacking** | **0.12471** | **46.93%** | **37.65%** |