House prices: Advanced Regression Technique

Team:

Beibhinn Gallagher

Anusha Singamaneni

Abstract:

In this project, we will be conducting an exhaustive search for the best predictive model for Housing data. This project will be using the Ames Housing Dataset and will be preprocessing, fitting, and testing various models to determine which machine learning regression model best predicts house prices. The data was collected by the Ames City Assessor's Office of assessments and the testing of the Root Mean Squared Logarithmic Error (RMSLE) was organized through Kaggle.

Introduction:

Purchasing a home remains one of the biggest purchasing decisions that individuals make in their lifetime. Our project aims to predict sale prices of houses in Ames, Iowa by using various aspects of residential homes. The insights gathered could be used by individuals to complement their decision-making process when purchasing a house. This helps to maximize the value users can gain while keeping to a budget. Findings could also be used by a property agent to improve the probability of a sale through proper marketing of key variables. The dataset that we are working on consists of mainly categorical variables stored as integers and factors. We would be focusing on descriptive and predictive analytics, with suggestions on how additional data could be obtained to conduct more experiments to optimize the purchase. The dataset used is obtained from Kaggle. Both team members have worked with the Boston Housing Dataset in previous classes and this dataset is intended to be a more updated and more challenging application of regression models.

The objective of this project was to predict the final sales price for each house using advanced regression techniques. Submissions to this Kaggle competition are evaluated based on the Root Mean Squared Logarithmic Error (RMSLE) for the predicted value and the observed sales price of each house. It was our goal to create a comprehensive project that includes an exhaustive search of models that would best fit this data and increase our status on the leaderboard in the process. In our report we have recorded our methods of searching for models as well as our corresponding RMSLE scores obtained.

Data and Related Work:

This competition and data was organized by Dr. Dean DeCock at Truman State University. Dr. DeCock had worked with the Boston Housing Dataset which is 506 instances and 14 variables for many years before becoming dissatisfied with the outdated data from the 70's (original dataset). Dr. DeCock wanted to organize a modern version of the Boston Housing Dataset that had a reasonably large number of variables and observations so that students would have to go beyond simple algorithms to construct a final model. He obtained data from the Ames City Assessor's Office of assessments made between 2006 and 2010. The training dataset has a total of 1460 observations with 80 variables while the test dataset has one less variable since we have excluded the predicted variable, SalePrice. Both datasets are combined together to ensure consistency in pre-processing. The 20 continuous variables relate mostly to various area dimensions for each observation. The 14 discrete variables typically quantify the number of items occurring within the house such as bedrooms and bathrooms. There are a large number of categorical variables (23 nominal, 23 ordinal) associated with this data set. They range from 2 to 28 classes with the smallest being *STREET* (gravel or paved) and the largest being *NEIGHBORHOOD* (areas within the Ames city limits). In order to work with this dataset we will need to preprocess the data. There are missing data and categorical variables that will need to be converted to dummy variables for some models. Previous work with this data set consisted of the processes of using a simplistic model, a more complex cross validated model, and mainly focused on the statistical aspects of regression modeling and not on machine learning methods. In our project we have started in a similar fashion applying simple learning models and graduating to more complex and higher performing models.

The below files are made available for the analysis:

i. train.csv – this is the entire training set, consisting of 1460 rows and 80 columns
ii. test.csv – this is the test set, with 1459 rows and 79 columns
iii. data_description.txt – which contains the explanation for each column in the data set.

In a related paper, “A Net Over Your Head: A Neural Network Approach to Home Price Predictions” by Hongtao Sun and Ji Hoon "Andy" Kim they outlined a similar process to our own on a Sberbank Russian Housing Market dataset also hosted through a Kaggle competition.

Their process was testing the following models: Principal Component Regression, Lasso and Ridge Regression, Regression Forests, XGBoost Gradient Boosting, and lastly, Neural Networks. Similar to what we will find in our own work below, their XGBoost Gradient Boosting performed best compared to previous models. However, this team constructed a Neural Networks model that outperformed in RMSLE compared to the other models. After reading this research paper it is clear that next steps for this project, if it were to continue, would be constructing Neural Networks. It is important to note that our project's RMSLE at about 0.13711 to 0.23564 are significantly lower than this study's RMSLE at about 0.4 when comparing the same models. However, their model does significantly improve RMSLE to about 0.25 when Neural Networks are used.

Methods:

In order to complete our project, it was necessary to preprocess the data before working with models. We completed exploratory visualization and analysis before completing machine learning applications. In order to ensure that we have found a well fitting model, we performed an exhaustive search. This has included using many methods and models.

In order to preprocess the data, we conducted preliminary analysis to determine how many null values were in our dataset. Columns that had a significant amount of null data were dropped. The columns dropped had greater than 250 missing values out of 1460 rows. The remaining null values were processed by filling the na values with the mean or mode of the column values depending on if the data was categorical or numerical.

The data consisted of both numeric and categorical data and in order to efficiently use all features of the data, the categorical data was converted into numerical values. This process was completed by using the LabelEncoder() to transform the categorical data. After the data is represented by all numerical values, we were able to analyze the correlations between variables and SalePrice. It was found that the following had the highest positive correlation to SalePrice: OverallQual, GrLivArea, GarageCars, GarageArea, TotalBsmentSF, 1stFlrSF, FullBath, TotRmsAbvGrd, and YearBuilt with correlations ranging from .79 to .52. The following had the highest negative correlation to SalePrice: ExterQual, BsmtQual, KitchenQual, and GarageFinish with correlation ranging from -0.64 to -0.54. This negative correlation does not necessarily mean what the first impression would imply because these features have been transformed into numeric values. In these cases there is a negative correlation because the value of 1 is rated as excellent while a higher number such as 6 is rated as poor, thus causing the negative correlation. To model the data, as related work for this dataset, we followed the method of starting with simple models and progressing to more complex models.

Experiments:

To start modeling the data, we applied simple models. We started by applying basic LinearRegression() and achieved an accuracy of 84.85% on the training data . This accuracy is decent so we then try Lasso Linear Regression and Ridge Regression. Lasso produces an accuracy of 84.85% on the training data, and Ridge results in a slightly lower accuracy of 84.81% on the training data . The next steps taken were to apply more complex models. Using Cross-Validated Grid Search for KNeighborsRegressor, an accuracy of 100% was achieved on the training data using the best parameters: KNeighborsRegressor(n_neighbors=8, weights='distance'). As suspected, this model overfits the data and the Root Mean Squared Logarithmic Error (RMSLE) is 0.23564 on the validation set. We then tried the same process of using cross-validated grid search to find the best parameters for RandomForestRegressor which achieved 98.05% accuracy on the training data using the parameters: RandomForestRegressor(max_depth=59, max_features='sqrt'). This model resulted in an improved RMSLE score of 0.15010 on the validation set. Random Forest Regressor is clearly the best model at this point in the project. In order to ensure the best possible model selection, we then used RandomizedSearchCV to determine the best regression model which was the following:

```
xgboost.XGBRegressor(base_score=0.25, booster='gbtree', colsample_bylevel=1,
             colsample_bynode=1, colsample_bytree=1, gamma=0, gpu_id=-1,
             importance_type='gain', interaction_constraints='',
             learning_rate=0.1, max_delta_step=0, max_depth=2,
             min_child_weight=1, missing=nan, monotone_constraints='()',
             n_estimators=900, n_jobs=0, num_parallel_tree=1, random_state=0,
             reg_alpha=0, reg_lambda=1, scale_pos_weight=1, subsample=1,
             tree_method='exact', validate_parameters=1, verbosity=None).
```

This model resulted in 98.28% accuracy on the training data and the best RMSLE score of 0.13711 on the validation set.

Conclusion:

In this project it was found that xgboost.XGBRegressor was the best model for predicting SalePrice of homes. After having success with Random Forest it was a logical next step to use xgboost.XGBRegressor. Random Forest is a bagging based algorithm where only a subset of features are selected in order to build a forest of decision trees, whereas XGBoost utilizes an optimized gradient boosting algorithm through parallel processing, pruning, and regularization to avoid overfitting. The best model resulted in 98.28% accuracy on the training data and the best RMSLE score of 0.13711 on the validation set. Next steps to further expand this research would include constructing Neural Networks as seen in the related work.

Resources:

SUN, HONGTAO, and JI HOON "ANDY" KIM. “A Net Over Your Head: A Neural Network Approach to Home Price Predictions.” *Stanford.edu*, 2017, cs229.stanford.edu/proj2017/final-reports/5191322.pdf.