Predicting House Prices

THE BATTLE OF NEIGHBORHOODS

IBM CAPSTONE PROJECT

Table of content

Introduction

When attempting to explain house prices, one often considers square footage, number of bedrooms and bathrooms.

A dataset popularized by Kaggle is the Ames Housing data providing 79 features to predict the price of any given house sold in Ames, Iowa. This dataset contains vast information about the house itself, and also a few pieces of information about proximity to main road or railroads.

This study seeks to answer:

- What is the price tag on each relevant neighborhood feature.
- What kind of surrounding venues matter the most.

Data acquisition

Housing data:

- Data retrieved from Kaggle.
- The data set comprises of 1460 observations of houses sold in Ames.
- The data contains 79 features to explain the sale price of houses sold in Ames.
- The data was compiled by Dean De Cock for educational purposes.

Neighborhood data:

- Venue data in each neighborhood is retrieved using a Foursquare API.
- Latitudes and longitudes are gathered using an OpenCage Geocoder API.

Data cleaning

Missing values:

- For categorical features, e.g. Pool, Alley, Fence, Fireplace, Garage, Basement, are replaced with ‘None’ to obtain a new category of houses sold without the feature.
- For numerical features, e.g. LotFrontage, we assigned the mean value.

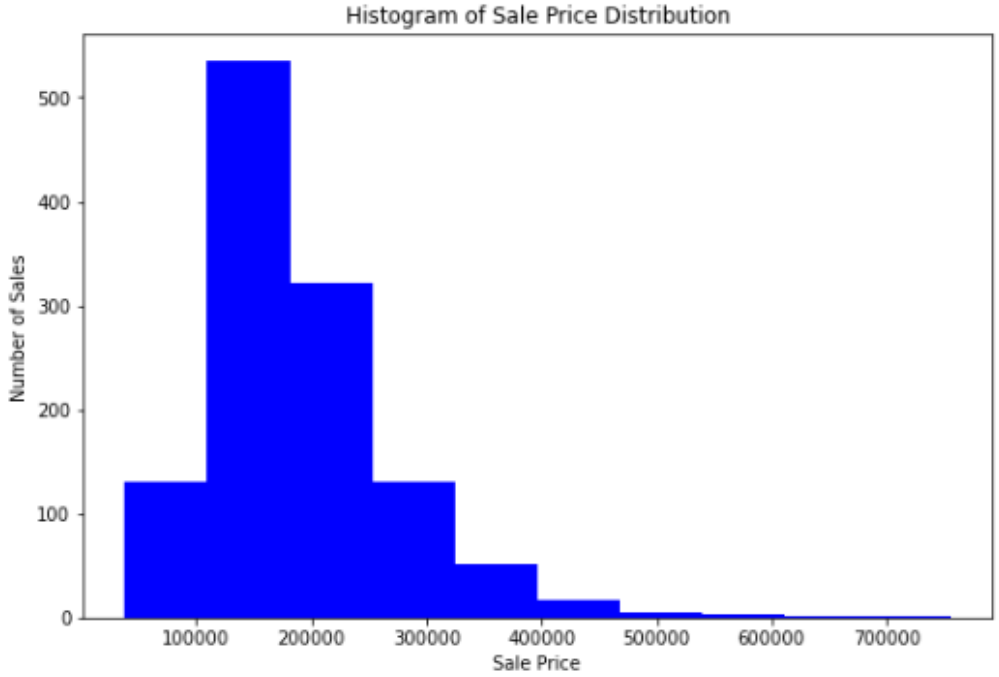
Redundant variables:

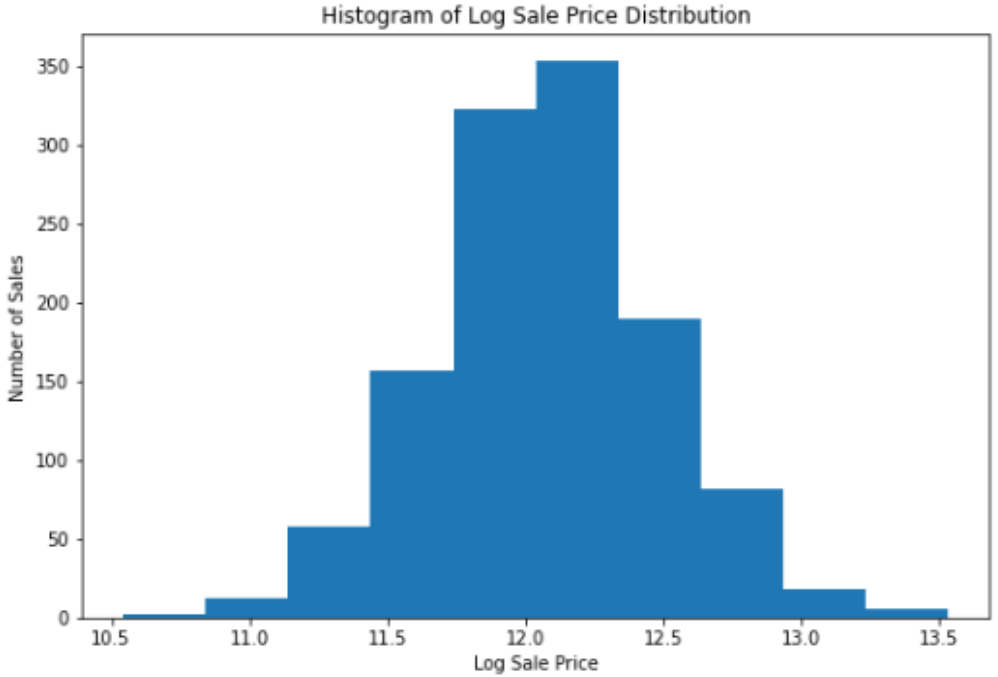
- We use the VIF factor to capture features that can be constructed through linear combinations of other features to alleviate strong multicollinearity.
  - Only GrLivArea, age and school have VIF factors above 10 in the final feature set.
- Exterior2nd and Condition2 are almost identical to Exterior1st and Condition1. Therefore, the second features were dropped.

Feature engineering

Log-transformation of target

- To better approach a gaussian distribution of the ‘SalePrice’ target feature, we take the natural log. The resulting distribution is approximately normal.





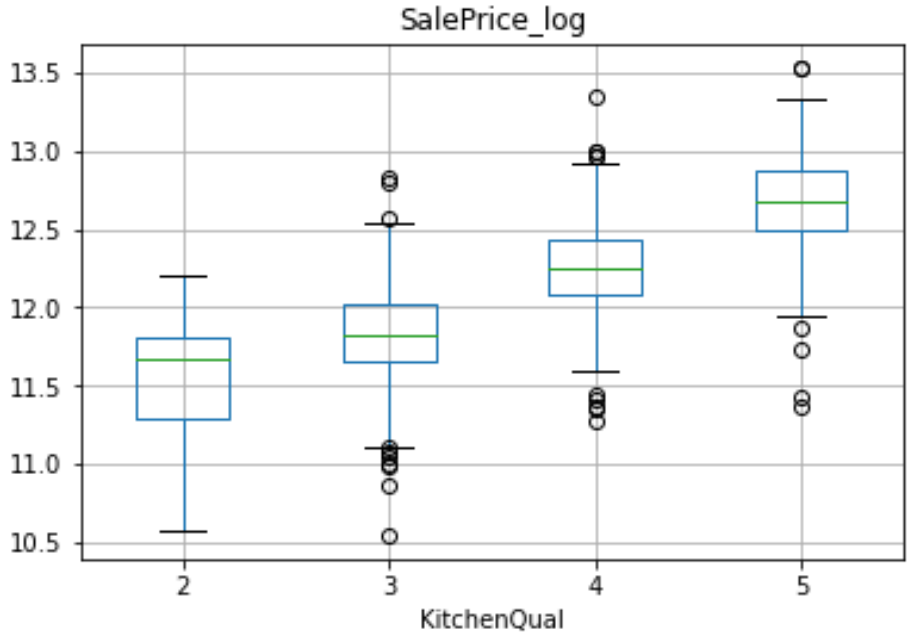
Feature engineering

Converting categorical variables

- All features with a ranking system were converted to numerical: {'Ex': 5, 'Gd': 4, 'TA': 3, 'Fa': 2, 'None': 1, 'Po':0} (e.g. see Kitchen Qual)

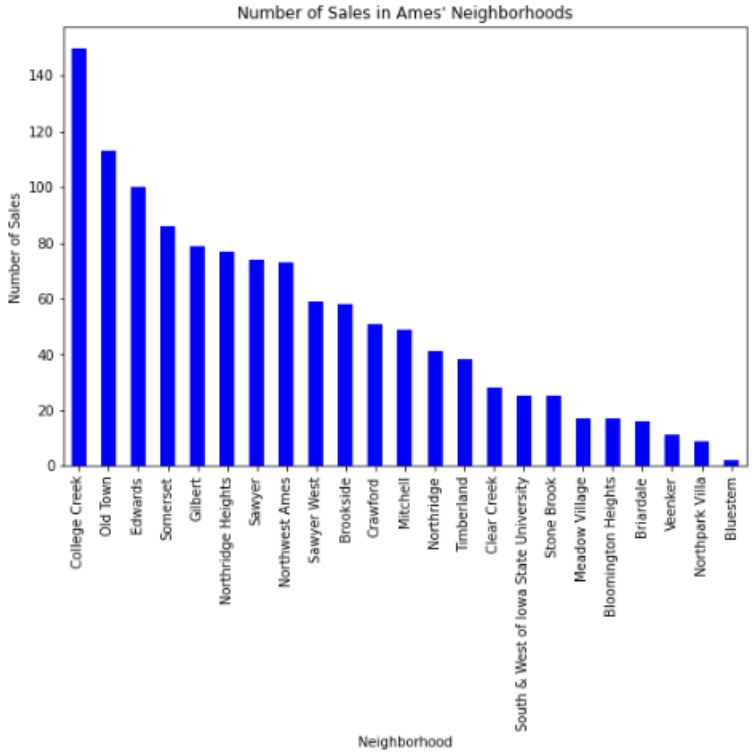
Age

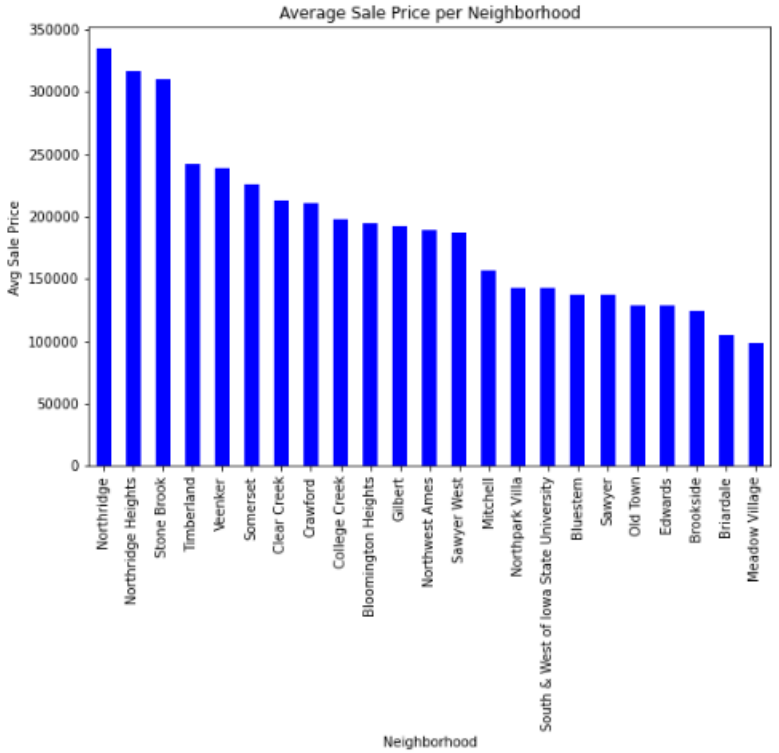
- A feature ‘Age’ was introduced by subtracting YearBuilt from YrSold. This variable will give an indication of whether the house is old or new.

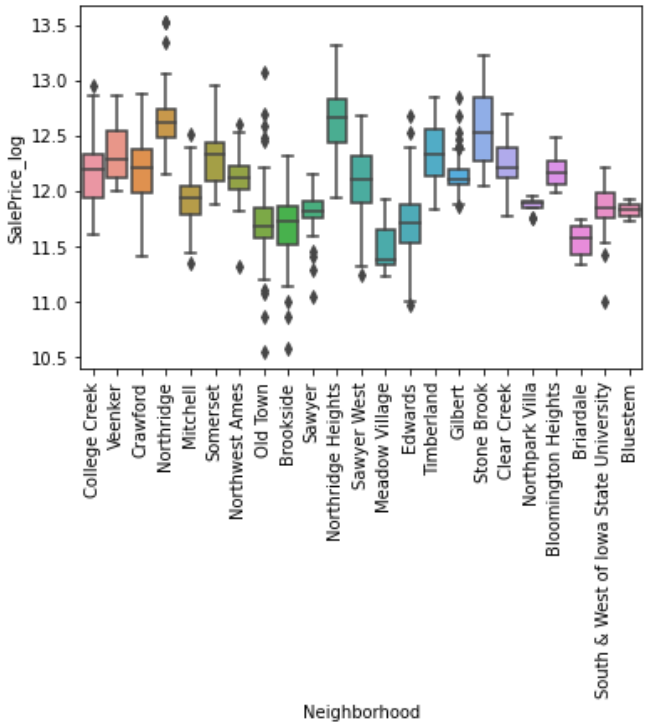


EDA – Neighborhood prices

- Most sales in College Creek and highest average sale price in Northridge.
- Almost no sales in Bluestem and cheapest neighborhood on average is Meadow Vlg.
- Northridge areas and Stone Brock also spike out in the boxplot as being more expensive.

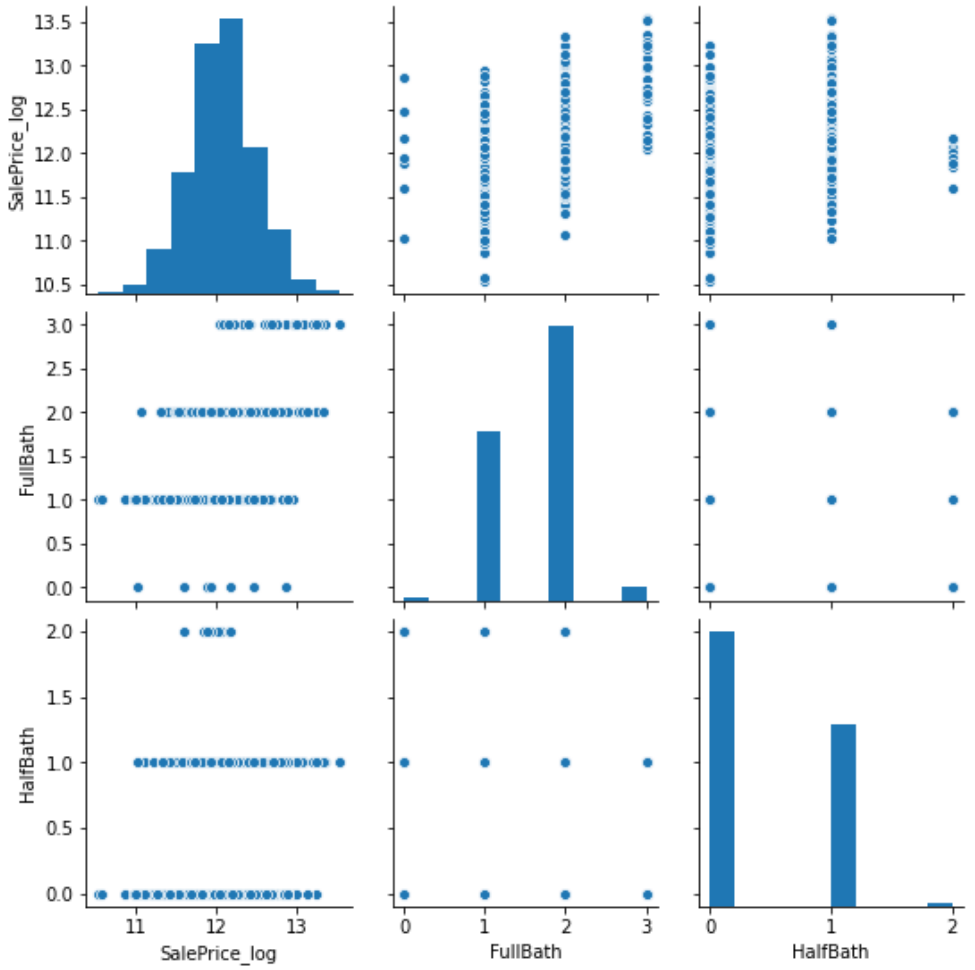






EDA – Bath variables

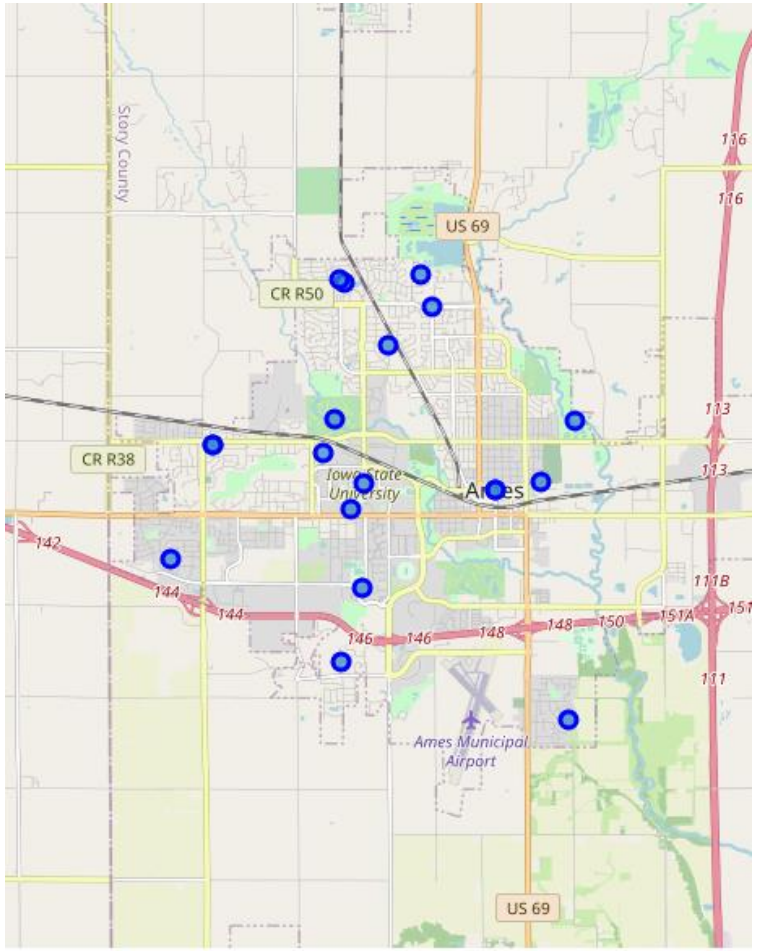
- We refer to the Feature Selection and EDA notebook for reference to the full exploration of categorical and numerical features relation to the sales price.
- An example is the Bath variables
- FullBath display a positive relation to SalePrice_log
- HalfBath shows no clear pattern
- HalfBath was therefore dropped from the analysis



EDA - Folium map of neighborhoods

Notable from map:

- Northridge and Northridge Highs are located next to the CR R50 label. Stonebrook is the marker just below US 69.
- Meadow Village in the middle of Iowa State University with College Creek just below.
- Michell in the bottom right, next to the airport.
- Timberland on the left side of the airport.
- Edwards furthest to the left.
- The expensive areas are therefore located in the northern part of Ames.



Linear regression - Price tags

- Ames house buyers are therefore willing to pay extra to live near universities (This is most likely an effect of College Creek being the 7th most expensive neighborhood and the location of Iowa State University).
- Recreational venues are concentrated in cheap city center neighborhoods (Meadow Vlg and Edwards), which explain the negative sign.
- Transportation options are also greater in the cheaper city center neighborhoods.
- In summary, it would seem to indicate that Ames housing is priced with a higher price tag on houses in the outskirts of the city center, and buyers put a larger price on houses in the vicinity of more schools, parks and leisure venues.

	coef	std err	t	P>|t|	[0.025	0.975]
restaurants	-0.0125	0.011	-1.094	0.274	-0.035	0.010
transport	-0.0462	0.011	-4.022	0.000	-0.069	-0.024
school	0.0567	0.014	4.104	0.000	0.030	0.084
golf	-0.0002	0.006	-0.027	0.979	-0.012	0.012
park	0.0444	0.010	4.491	0.000	0.025	0.064
recreational	-0.0289	0.008	-3.828	0.000	-0.044	-0.014
cultural	-0.0137	0.010	-1.408	0.159	-0.033	0.005
leisure	0.0288	0.008	3.702	0.000	0.014	0.044

Model building – Selection

- As we are working with a continuous target variable, regression models are appropriate.
- Linear regression is naturally included. But as we have a large amount of potentially still redundant features, Lasso and Elastic net regressions are also introduced.
- We also include Random Forest regression. Given the many binary features in the data set, a tree-based model makes a lot of sense. As it uses a random sample of the data, we are less likely to overfit on the training data.
- Moreover, a Support Vector Regression (SVR) is introduced. The benefits of this model is that SVR acknowledges the presence of non-linearity in the data set. This model could therefore provide a more proficient prediction model.
- Finally, a Gradient Boosting model is selected for testing. As it build trees one at a time and correct errors made on previously trained trees, this model could potentially provide a similar and potentially better model than the random forest.

Model building – Spot check CV

Model	Negative mean absolute error	Stdev
Linear Regression	-3.626.332.352	10.878.997.056
Lasso Regression	-0.318790	0.0176
Elastic Net	-0.318790	0.0176
Random Forest	-0.105817	0.0064
Support Vector	-0.127016	0.0148
Gradient Boosting	-0.096784	0.0053

- Spot checking algorithms with cross validation on the training set results in the following mean absolute errors.
- The linear regression errors are most likely a result of a large amount of features compared to limited data, or simply breaches of GM-assumptions (e.g. correlated features).
- Random Forest, Support Vectors and Gradient Boosting performs better, and model tuning will be performed on those.

Model building – Tuning

- Random Forest
  - Number of estimators: tuned to 280.
  - Max depths: 25.
- Support Vector
  - The regularization parameter, C, is tuned to 0.5.
  - The kernel to ‘linear’.
- Gradient Boosting
  - Learning rate tuned to 0.1.
  - Number of estimators: 180.
  - Loss function: ‘huber’.

Model evaluation

- The tuned models perform slightly better than the un-tuned models.
- The gradient boosting model performs better than random forest and SVR.
- An average of all three models combined outperforms the individual prediction model.
- We observe some degree of overfitting for Random Forest and Gradient Boosting.
- Taking the average results in an average error of $14.335 on house prices in Ames
- Surprisingly, for SVR, we see the train error > generalization error. This could indicate that it is an unknown fit, or that the test set contained easy cases to price, and the train set contained harder cases.

	Train Neg MAE	Test Neg MAE
Random Forest	-7581.4	-17262.531536
Support Vector	-17709.8	-15930.752618
Gradient Boosting	-9391.56	-15504.138400
Combined		-14335.539862

Feature importance

- From a random forest feature importance output, we observe a clear indication that overall quality and above ground living area were key in explaining house prices.
- The first Foursquare feature was restaurants with a VI of 0.00689 in the 25th rank.
- From this it is evident that the neighborhood venues features does not add much in explaining the house prices. They only serve as indicators of a (small) positive or negative impact by proximity to the specific venues.

	Feature	Variable Importance
2	OverallQual	0.330655
12	GrLivArea	0.126630
9	TotalBsmtSF	0.088130
20	GarageArea	0.045721
19	GarageCars	0.040806
38	age	0.032982
1	LotArea	0.030514
4	YearRemodAdd	0.020636
0	LotFrontage	0.020093
13	FullBath	0.015834
3	OverallCond	0.013621
6	BsmtQual	0.012727
29	MoSold	0.012191
22	WoodDeckSF	0.011665
5	MasVnrArea	0.011639
23	OpenPorchSF	0.011603
18	FireplaceQu	0.009329
42	MSZoning_RL	0.007667
16	TotRmsAbvGrd	0.007517
14	BedroomAbvGr	0.007293
117	MasVnrType_BrkCmn	0.007273
8	BsmtExposure	0.007212
15	KitchenQual	0.006893
30	restaurants	0.006734

Conclusion

Questions: What is the price tag on each relevant neighborhood feature and what kind of surrounding venues matter the most?

- Based on a linear regression we found that house buyers are paying a statistically significant premium on proximity to parks, leisure and school venues, while significantly discounting proximity to transportation and recreational venues. The cheapest neighborhood on average, Meadow Village, is also a neighborhood in central Ames, meaning lots of transportation options and recreational venues – as well as restaurants, which also appear with a negative (insignificant) coefficient.
- The model does not improve the fit from a model based on a simple neighborhood category. This could mainly be due to the degrees of freedom being different. However, the model does tell a story about how certain venues make the neighborhoods differ in price. The results are therefore useful to e.g. potential house buyers in Ames, as the model can give a fair estimate of the price of any house given a list of the feature inputs with an average error of 14 thousand dollars.