AMES, IOWA HOUSING MARKET

A DATA MUNGING CHALLENGE

EXECUTIVE SUMMARY

- The importance of proper data cleaning, munging, wrangling, manipulation cannot be overstated.
- The preponderance of messy data columns and features, the amount of errors and/or null values required careful selection and aggregation in order to arrive at accurate sale price predictions.

ADDRESSING NULL VALUES

Addressing columns with quality ratings.

Columns with nominal values from Poor - Excellent. Assigned ordinal values from 1 - 5. This process is important later as the columns are rated and grouped into 'good' and 'bad' features.

NaN in certain columns indicated the feature was not present so they were assigned with a zero value.

CORRELATION AND FEATURES

Check for correlation using a heat map.

Strong positive correlations in 'overall_qual', 'gr_living_area', 'garage_area', 'garage_cars', 'total_bsmt_sf', 'year_built', '1st_flr_sf', 'year_remod/add', 'full_bath', and 'garage_yr_built'.

There appear to be no strong negative correlations.

- Taken as individual features these correlations might not have a strong impact on our target (‘saleprice').

HOWEVER, WHEN GROUPED TOGETHER…

- Our engineered features have a strong correlations with our target.
- Strong positive correlations in 'overall_qual', 'good_features', 'gr_liv_area', 'exter_qual', 'garage_area', 'garage_cars', 'total_bsmt_sf', '1st_flr_sf', 'kitchen_qual', 'bsmt_qual', and ‘full_bath’.
- Strong negative correlations in 'age' and ‘bad_features'.

WHAT ARE 'GOOD FEATURES'? WHAT ARE 'BAD FEATURES'?

- The good feature group was comprised of the converted nominal features. Such as an 'excellent quality basement' or a 'built-in fireplace' were added to the 'good features' group based on wether they scored above a '4' or had a certain criteria, like 'built-in'.
- The bad feature group was also comprised of converted nominal features with the inverse criteria.
- Sometimes, the feature was either there or it wasn't. Like a garage or a pool.

MODELLING AND FITTING

A multiple linear regression was used to generate predicted home values on a blind set of data.

An R-squared score of 92% was achieved.

The relationship between all of our dependent variables (sale price) and our independent variables (good features, bad features, exterior condition, etc.) is a strong one.

CONCLUSIONS & FURTHER STUDY

Having a clear goal while feature engineering, data munging, and modeling are the key to accurate results with this dataset.

Knowing what features were important and how to leverage them to greater affect on the target helped immensely.

Grouping or clumping features together was fruitful and will be used again for other projects of this magnitude. Deliberately avoiding features that are inherently problematic, such as zip code or neighborhood, prevent our models from having a sociological bias.

Future projects with this dataset could include using other models besides linear regression such as Ridge/Lasso/Elastic Net or Random Forest.