Springboard Capstone Milestone report

Dylan Distasio

May 23, 2017

Springboard Capstone Milestone report

Introduction

This capstone project is based on a Kaggle competition hosted by Sberbank. The full details from Kaggle are as follows:

Housing costs demand a significant investment from both consumers and developers. And when it comes to planning a budget-whether personal or corporate-the last thing anyone needs is uncertainty about one of their biggets expenses. Sberbank, Russia's oldest and largest bank, helps their customers by making predictions about realty prices so renters, developers, and lenders are more confident when they sign a lease or purchase a building.

Although the housing market is relatively stable in Russia, the country's volatile economy makes forecasting prices as a function of apartment characteristics a unique challenge. Complex interactions between housing features such as number of bedrooms and location are enough to make pricing predictions complicated. Adding an unstable economy to the mix means Sberbank and their customers need more than simple regression models in their arsenal.

In this competition, Sberbank is challenging Kagglers to develop algorithms which use a broad spectrum of features to predict realty prices. Competitors will rely on a rich dataset that includes housing data and macroeconomic patterns. An accurate forecasting model will allow Sberbank to provide more certainty to their customers in an uncertain economy.

The problem I will be looking to solve is building a model to attempt to predict realty price fluctuations in Russia based on the provided data.

Initial Exploratory Analysis

The training dataset contains 30471 observations of 292 potential features about each property (including the sale price), its surrounding neighborhood, and some features that are constant across each sub area known as a Raion.

The macroeconomic dataset contains 2484 observations of 100 potential features.

The two datasets were joined together based on the timestamp field common to both.

- Initial correlation analysis of the features selected regarding the property shows the strongest correlation between the square footage of the property, and the sales price.
- Initial correlation analysis of the features selected related to macroeconomic factors shows weak correlations with the sales price.
- There may be a time lag involved with the response of the sales price to macroeconomic factors which may be difficult to capture in many basic models.
- There are a number of potential features with a large % of missing values that may need to be massaged if used for building a model.

- I plan on beginning with a linear regression predictive model for home sales price based on a number of features selected related to the dwelling itself, compared to a second one that incorporates macroeconomic features.