Springboard Capstone Proposal

Dylan Distasio

May 22, 2017

Springboard Capstone Proposal

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.

Project Plan

1. 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.
2. Sberbank is the client and cares about the problem because accurately predicting realty prices in Russia would allow them to decrease uncertainty in the marketplace and make more accurate business decisions.
3. I will be using data provided by Sberbank to perform exploratory data analysis and build a predictive model using machine learning techniques. They have provided training and test datasets along with a dataset containing potentially useful macroeconomic information.
4. My approach to solving this problem will begin with exploratory data analysis and visualization to examine data quality, correlations between features, and other items of interest. Based on this analysis, I plan on comparing the performance of a number of different ML models based on selected features.
5. All code and visualizations, along with commentary will be provided in R markdown format including a deck generated from the raw output.