Capstone Project Proposal

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 biggest expenses. Therefore, many companies such as Zillow, a real-estate Marketplace, or, 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. Recently, Sberbank has challenged Kagglers to develop algorithms which use a broad spectrum of features to predict realty prices. They provided a rich dataset that includes housing data and macroeconomic patterns.
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. An accurate forecasting model will allow Sberbank to provide more certainty to their customers in an uncertain economy.
The dataset has been released in the Kaggle (https://www.kaggle.com/c/sberbank-russian-housing-market)

Acquiring Data:

The Sberbank Russian Housing Market dataset has recently been released by Sber Bank, the oldest and the largest Russia's bank (https://www.kaggle.com/c/sberbank-russian-housing-market/data). The datasets was released for kagglers to propose some models to forecast the house prices. The train dataset contains more than 30000 transactions from August 2011 to June 2015 in the Russian House Market for some years, and the macro dataset provides the daily economic parameters in Russia. A test data also published to evaluate the kagglers' models. As the real prices in the test file is hidden by competition holders, therefor for the purpose of this capstone project, only the train and macro data is used.

The train data contains 30471 rows with 391 columns for each data point, which has been collected from August 2011 to June 2015. The dataset has been riched with different types of features such as house characteristics, demographics, cultural information, country economical information, etc. The macro data includes 100 economical parameters in Russia for 2484 successive days.

Approache:

My approach to solve this problem includes 4 major steps that are mentioned below in detailed, based on their priorities.

Exploratory Data Analysis:

EDA consists of an exploration to better see the correlations and facts that are hidden in data. Many stories could be told showing different types of visualizations. I would seek for the most important features and their effects in this problem, House Price Prediction.

Also from the initial EDA also we could find some things called bad data, outliers and hidden missing values, that must be handled in next stage.

Data Wrangling:

First of all, a deep data wrangling must be done to prune the dataset from missing values, bad data and outliers. For handling the missing values, we have different ways:

1- drop-out missing values: When we have a very few outliers or missing values it could be acceptable, but in most of the cases that's a very bad idea.

2- Independent Imputation of columns: We can fill missing values in each column with a function (example; mean, median) of that column independently. It does very fast but not consider the correlations with other variables.

3- Regression Imputation: We can apply some Regression models such as Linear or knn Regression to estimate the missing values with respect to correlated features. It's more precise but very slow method.

4- Decision Tree based Algorithm: Decision tree has the potential to handle with missing values. The best Machine Learning algorithm to do this is the Extreme Gradient boosting which has been recently published. xgboost is a package of libraries that implemented this algorithm.

Feature Engineering:

A very effective way to boost up the model performance is to use the art of Feature Engineering. That's to some extent heuristic, but it worth considering. Here I extended the timestamp column into different variables, such as, year, month, quarter, etc. This is because the house price has a periodic characteristic with respect to month, quarter, year, etc. Therefore, these features could guide the model to better learned.

Machine Learning:

The price prediction is a Regression problem. There are a lot of Algorithms to handle this problem, such as Linear Regression, the simplest one, Decision Tree, Random Forest Gradient boosting, Xgboost, etc. Fortunately, all of these useful algorithms are implemented in sci-kit learn library

package. Therefore, I will start tuning each model hyper parameters using grid search cross validation. At the end a comparison between different models could be done on their Regression power and Forecasting power.