**Anushree Bordia, Jeffrey Lierman, Samantha Misurda, Benjamin Wynia**

**Applied Data Science**

**Project Proposal**

## Project Need

The purpose of this project is develop an analytical model to predict the variation in price between the Zillow “Zestimate” for a home’s sale price, and the eventual sale price of the home. This model will help identify problems with the current Zestimate home price model, improving real estate price predictability for both buyers and sellers.

## The Dataset [1]

The dataset we will be exploring is part of a Kaggle competition sponsored by Zillow. We are provided with a record of home transactions taking place in the year 2016 in Los Angeles, Orange, and Ventura County. Our training set will consist of those sales occurring prior to October 15, and the test set of those thereafter. The files provided contain data fields with basic information, as well as the log error between the “Zestimate” and the actual sale price

## Hypothesis

We believe that home value can be modeled as a function of the following variables:

* Number of square feet
* Number of bedrooms
* Number of bathrooms
* Year constructed
* Home amenities
* Size of lot
* Land value
* Location and neighborhood quality
* Prevailing economic conditions, including median income, unemployment rate, expected population growth rate, crime rate, and mean time on the market

## Technical Approaches

Using the variables specified above, we will build the following models for predicting home value, and select the most accurate. Accurate in this case is quantified by the range between the actual sale price and the predicted sale price.

* Baseline simple linear regression model,
* Gradient Boosting,
* SVM,
* KNN,
* Radio Operating Characteristic (ROC) Analysis
* Entropy & Information Gain
* and Decision Trees/Forests.

## Project Execution Risks and Risk Mitigation Strategies

With the models specified above, the two most obvious risks are overfitting the model, and poor model selection. With building a regression model, there is the potential of training the model too specifically on the training data, and as a result, capturing a noisy or nonexistent relationship. In order to mitigate this risk, we may choose to build our training set using cross validation.

Secondly, when selecting features for our model, we must make sure that we are selecting relevant predictors. A strong relationship between two variables does not necessarily imply that they are functionally related, and thus useful predictors.

## Goals and Deliverables

Our goals will align closely with the preliminary round of the Kaggle challenge. Using R as a statistical platform, we plan to produce a model capable of predicting home sale prices on a test set of home sales in the final quarter of 2016. From the early work produced in the competition, Zillow has dropped their median margin of error from 14% to 5% in predicting the eventual sales price. Building a model that falls in that range would be an indicator of significant success.

Additionally, we will deliver a report (in HTML/CSS) that outlines the analysis performed and the visualization (via D3.js) of the findings therein.

If time allows, we may draw new features from additional external datasets, such as crime statistics.

## Updates and Amendments

*For Mini Week ending Sept. 12th, 2017*

* [Goals and Deliverables](#_mhoofms6r7qn) was extended to emphasize that to go above and beyond, this project will adopt D3 to provide superior data visualization.
* [Goals and Deliverables](#_mhoofms6r7qn) was extended to clarify that this project will use R as the dominant statistics language. This will be extended as appropriate by html, css, and JavaScript (e.g. D3.js) to improve the quality of the report.
* [Technical approaches](#_bdgz5gan6j36) were extended to include, but not remain limited to: Radio Operating Characteristic (ROC) Analysis, Entropy & Information Gain, and Decision Trees/Forests.

## References

[1] <https://www.kaggle.com/c/zillow-prize-1>