Proposal for STOR565 Project

Rental Listing Inquiries Prediction

1. **Group 7 members**

Hatip, Ahmet; Lan, Zhao; Qin, Di; Smith, Scott; Sun, Shuming.

1. **Project description**

The project will predict how popular an apartment rental listing is based on the listing content, such as price, number of bedrooms, text description, etc. The response variable, interest\_level, is defined by the number of inquiries that a listing received on the site.

1. **Data description**

The data set was posted on kaggle.com: https://www.kaggle.com/c/two-sigma-connect-rental-listing-inquiries. The original data comes from [renthop.com](https://www.renthop.com/), an apartment listing website. The train data set, which will be used for model building, is about 60 M, with 52730 observations and 15 variables. Besides, there is also a folder contains 78.5 GB image files.

Variable information:

* bathrooms: number of bathrooms
* bedrooms: number of bathrooms
* building\_id
* created date
* description
* display\_address
* features: a list of features about this apartment
* latitude
* listing\_id
* longitude
* manager\_id
* photos: a list of photo links
* price: in USD
* street\_address
* interest\_level: this is the target variable. It has 3 categories: 'high', 'medium', 'low'

1. **Type of machine learning involved and the techniques that will be used.**

The response variable is categorical, so classification will be used here and supervised learning will be involved in this project. Techniques will be used: variable selection, logistic regression, KNN, LDA, supporter vector machine.

1. **Challenges and brief plan**

Challenge 1 :How to extract useful information from text description and image files. For the former one, maybe we can use some text mining techniques, such as using Python to parsing text files and use some index factor to extract useful information.

Challenge 2: How to improve the prediction accuracy. We can try different methods and build different models to compare and choose one with highest accuracy. Since this is a complex issue, I guess the overall accuracy will not be high.