Ahmet Hatip

1. the list of members
   1. Hatip, Ahmet
   2. Lan, Zhao
   3. Qin, Di
   4. Smith, Scott
   5. Sun, Shuming
2. description of the problem of interest
   1. there are many apartment listings, and many of the apartments have different levels of interests. Our goal is to predict the different level
3. description of the dataset including data source, data size and variable information etc.
   1. data is from renthop.com and they give us information about apartment rents in New York
   2. data sources
      1. train.json- training set
      2. test.json- test set
      3. sample\_submission.csv- a sample submission file in preferred format
      4. images\_sample.zip- a sample of images of listings, size 100 listings
      5. Kaggle-renthop.7z- full list of images
   3. Variables
      1. Bathrooms: number of bathrooms
      2. Bedrooms: number of bedrooms
      3. Building\_id: id of the buildings
      4. Created: time the listing was created
      5. Description: description of listing
      6. Display\_address: address of listing that is displayed
      7. Features: a list of features for each listing
      8. Latitude
      9. Listing\_id: the id of the listing
      10. Longitude
      11. Manager\_id: id of manager
      12. Photos: a list of photo links
      13. Prices: price in USD
      14. Street\_address: real address
      15. Interest\_level: the response variable and the one we are trying to predict, categorical with high, medium, low
4. explanation on the type of machine learning involved and the techniques that will be used

We would use methods like

Logistic regression (& penalized logistic regression): this would be our most inflexible model

KNN classification: this would be our most flexible model

LDA & QDA: we can only use this if the variables are normally distributed but is may not be the case.

Support Vector Machines (SVM): the flexibility would be close to KNN but a bit more less

To reduce the number of variables we need, we may use techniques like

Lasso and ridge regression: to see which variables are appropriate and not and too make the model less flexible,

PLS and PCR: to shrink the dimension of the variables

If we use logistic regression we can use GAM to model certain areas differently and nonlinearly

In SVM we can change the kernel so that we can fit more nonlinear models

1. discussions on the potential challenges and a brief plan on how to handle the difficulties.

Some challenges that we may have is reducing the number of variables in the model and using variables that are relevant to the model. Also there are a lot of variables that mean the same idea like latitude and longitude, display\_address, street\_address, and building\_id.

Also some of the variables are just ids and some are just descriptions and lists of features. We may need to expand these variables and create new ones that extract the important information out of these variables

Another challenge would be trying to quantify or qualify the images that the listings has and put it in the model

Another challenge we would have is that we need to find an appropriate model for the data