***Agenda – ML Project Meeting with Thomas L.***

1. General Project Targets
   * Have initial EDA/feature selection completed tomorrow (2/20) and consolidate findings for to start initial model generation on (2/21)
   * Have initial model set end of Monday (2/24) and then consolidate on Tuesday (2/25) and move forward on presentation and/or refinements of earlier portions
   * Thought is to have analysis and presentation in a solid state by Friday and that Friday/over the weekend most of what we are doing is “spell check”/minor cleanup of presentation and code.
2. Initial EDA / Modeling
   * General Observations
     1. House prices are right skewed, a log transformation appears to make it mostly normal
     2. Some of the numeric columns that have high pairwise correlation are sqft and number of rooms, so probably makes sense to pick one or the other or some transformation (e.g., sqft per room)
     3. Some categorical columns appear to only have meaningful differences for a subset of the categories
        1. For example kitchens that are Excellent are a lot higher versus other values for kitchen quality (which are somewhat close)
        2. Another example is looking at incorporate neighborhoods -> some are noticeably higher but many are tightly bunched together -> so instead of dummifying the entire spectrum – thought may be to just dummify the top 3-5 that seem to be a “notch above” others (i.e., is this house in one of the top 3-5 area? – 1 or 0 within a single dummy variable)
   * EDA
     1. Some features that appear significant:
        1. Sqft per room
        2. Number of baths
        3. Number of bedrooms
        4. Whether an item is in a “top neighborhood” based upon median price?
        5. Whether a home is newer/more recently re-modeled
           1. Newer construction (e.g., past couple of decades) seem to show some separation from homes built prior to 1980
        6. ***Placeholder for others to discuss***
3. ***Questions***
   * Recommendations for handling outliers other than log transformation + looking to see if those outliers have some distinguishing feature that hasn’t been incorporated?
   * Feedback on “selective” dummification (e.g., only have a 0/1 for a subset of the categories values that appear to show meaningful separation from the rest of the pack)?
   * Question about incorporating models that we haven’t covered in class (e.g., Random Forest, Decision Tree)?

***Comments***