Dataset Quality

1. There was no sub-category such as tv, dvd, and etc. Due to limited time, did not have time to categorize them by product type, but this would be a good starting.
2. Definition not cleared for 1~2 fields, such as “Item Class”
3. Inconsistent “actual color”, aspect ratio and other fields

Data Processing

1. Html tags in product description that had to be removed
2. Removed all special characters
3. Considered only numeric and alphabetic characters
4. Used stemming
5. Used tfid to convert text field into matrix
6. Used the following columns for prediction: “Seller”, “Product Name”, “Item Class ID” and “Product Long Description”
7. Applied different weight factor for each column based on its importance to identify each product

Model

1. Considered “tag” as text field. Did not break them down into each tag level for simplicity
2. This is simple classification problem, so used linearSVC and randomforest to start with.
3. Multiple feedback loop was required for tuning parameters