Capstone Project Documentation

Members

- Melwin Jose – 200150505
- Rohit Nambisan -

1) Business Value of RentHop data set

RentHop makes apartment search smarter by using data to sort rental listings by quality. But while looking for the perfect apartment is difficult enough, structuring and making sense of all available real estate data programmatically is even harder. The project tries to predict the interest level of new listing receives based on the listing's features, building_id, manager_id, creation date and other input variables. Doing so will help RentHop identify potential listing quality issues, and allow owners and agents to better understand renters' needs and preferences.

Learning objectives:

- How Gradient Boosted Trees can be used for multi-class classification.
- Detect the most significant features that helps in the classification task.
- Use text processing and other methods to reduce the dimensionality/cardinality.
- Learn how to use Model-Stacking to improve accuracy.

3) Papers/Tutorials

- XGBoost: A Scalable Tree Boosting System
- Gradient Boosted feature selection
- Model Stacking: http://blog.kaggle.com/2016/12/27/a-kagglers-guide-to-model-stacking-in-practice/

6 ???

8) Feature Selection

Direct Features

- bathrooms
- bedrooms
- building_id
- latitude
- longitude
- manager_id
- log(price)

Engineered Features

- price_diff_bedrooms: price – mean price of apartments with same number of bedrooms
- total_rooms: bathrooms+bedrooms
- price_diff_rooms: price – mean price of apartments with same number of total_rooms

- bathbed: bathrooms/bedrooms
- bed_price: price/bedrooms
- room_price: price/total_rooms
- distances to the following places:
  - Parks: Central Park, Prospect, Pelham, Van Cortlandt, Flushing
  - Universities: New York University, Borough of Manhattan Community College, Columbia Unviersity, Hunter College, Kingsborough Community College, Bernard M Baruch College, Brooklyn College, New York, New York City College of Tech, City College, Touro College, John Jay College of Criminal Justice, Pace university, The New School, Fashion Institute of Technology
  - Subway Stations: Times Square, Grand Central, Herald Square, Union Square, Penn
- nphotos: number of photos
- freq_features: one hot encoding of hardwood floors, laundry, common laundry, no fee, pre-war
- display_addr: similar to display_address but with reduced cardinality using text processing
- price_diff_addr: price – mean price of apartments in the neighborhood

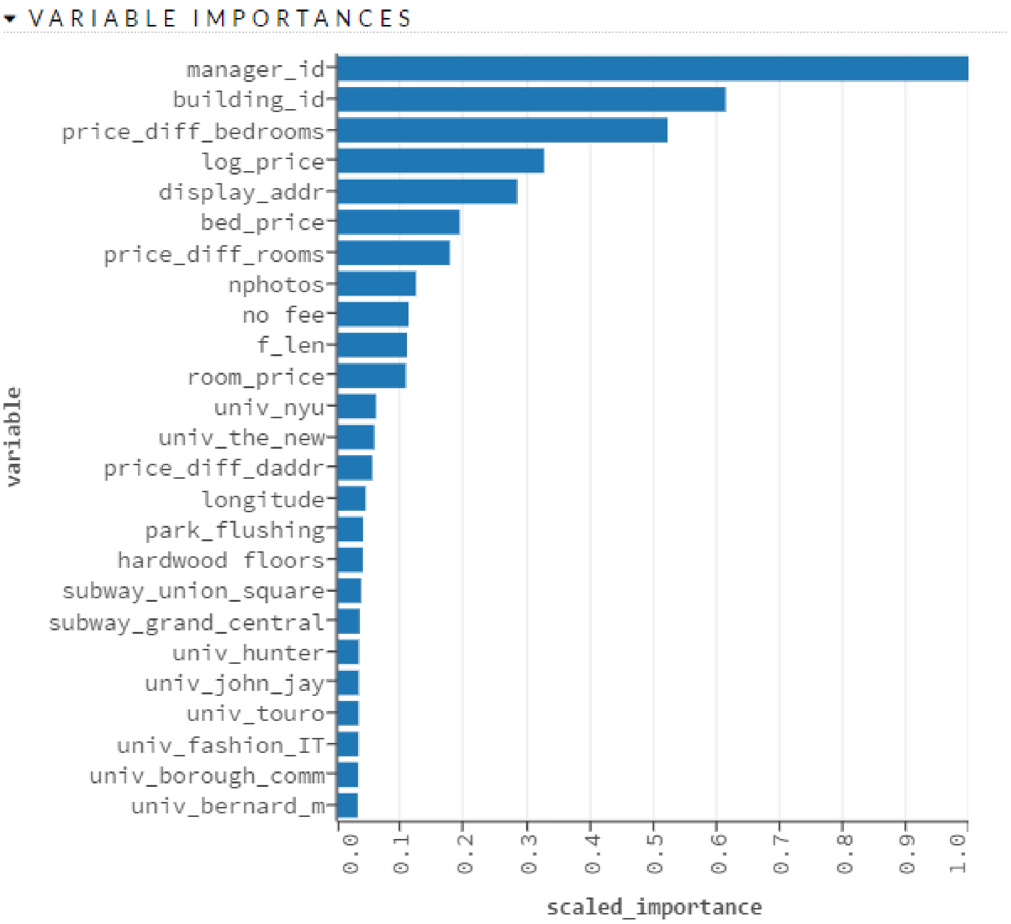
9) Solution

Text Processing for cardinality reduction

- The ‘features’ columns contains 1556 unique features, and a lot of them are duplicates. For example, "hardwood", "hardwood floor", "hardwood flooring" are duplicates of "hardwood floors". Similarly, other duplicate features can also be mapped to reduce the unique types of features. The list `freq_features_map` contains the mapping which is used to reduce them to 778 unique features.
  - Code: final_submission.R starting @line:388
- Similarly, the `display_address` has 8826 categories. As this column contains information other than the address and it can be processed to reduce the cardinality and to detect if the string contains non-address details. If it does, it is marked as “SKIP”. A list of non-address words is stored in `skip_words` and these are remove along with non-alphabetic characters/words to reduce the cardinality to 1916.
  - Code: final_submission.R starting @line:733

Gradient Boosted Machine for feature Selection

- To select the features that were played the most significant role in predicting the results, we have used the “variable importance” plot from h2o’s GBM. Variable importance is determined by calculating the relative influence of each variable: whether that variable was selected during splitting in the tree building process and how much the squared error (over all trees) improved as a result. More info: section “*8.1 Relative input of input variables*” in the paper “*Greedy Function Approximation: A Gradient Boosted Machine*”



- The above strategy was used to select the 45 most significant features by building a model on 100+ features, both direct and engineered ones, and dropping all those whose importance level was below threshold.
- Code: final_solution.R @line:890

Hyper-Parameter Tuning

- To find the best hyper-parameters for the XGBoost we train and tested the model on the following:

parameter	values	comments
nrounds	200, 400, 600, **800**, 100	number of rounds
subsample	0.2, 0.4, 0.6, **0.8**	subsample ratio of the training instance
colsample_bytree	0.3, 0.4, **0.5**, 0.6, 0.7	subsample ratio of columns when constructing each tree
min_child_wieght	50, **100**, 200	minimum sum of instance weight (hessian) needed in a child. If the tree partition step results in a leaf node with the sum of instance weight less than min_child_weight, then the building process will give up further partitioning.
eta	0.0125, **0.025**, 0.05,	learning rate

The ones in bold gave the best accuracy with 10-fold cross validation and were used to train the level-2 XGBoost

Stacking

- In order into improve the accuracy of the classification and to reduce the variance, Model Stacking have been used. Our stack consists of the following two levels:
    - Level-1: It builds the following models on the train and make prediction both train and test.
        - Generalized Linear Model
        - Neural Network – 2 hidden layers
        - Random Forest
        - H2O's Gradient Boosted Machine
    - Level-2: uses the results from level-1 along with the train data to make predictions on the test data
        - Trained 40 XGBoost's with different seeds and depths (10-fold cross-validation). And the results were averaged.
- Code: final_solution.R @line:948

10) Project Materials

- Github link
    - Folder Structure:
        - papers/ : research papers
        - final_submission.R : our solution to the problem
        - lvl-1_<model>_[train|test].csv : predictions from models from level-1
        - ../data/*.json : train and test data
- Member Contribution
    - Code:
        - Melwin:
            - Text Processing for cardinality reduction of display_addr and features
            - Feature Selection
            - Model Stacking
        - Rohit:
            - Feature engineering of places
            - Hyper-Parameter Tuning