Fantastic Apartments and How to Find Them

PREDICTING POPULARITY FOR RENTAL LISTINGS

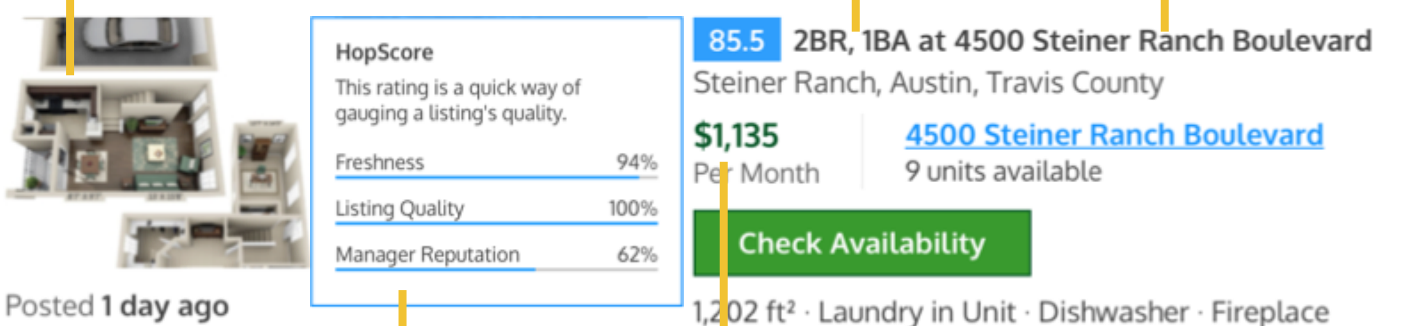
Tian Tan
Wei Ye
Jiaolong Yu
Alexander Groszewski
Tommy Huynh

Problem Description

Photos

Floorplan

Location



Freshness

Manager

Price

Amenities

Input Features & Output Target

5 numerical features

- Number of bedrooms
- Number of bathrooms
- Latitude
- Longitude
- Price

4 non-numerical features

- Building ID
- Manager ID
- Address
- Date created

3 array features

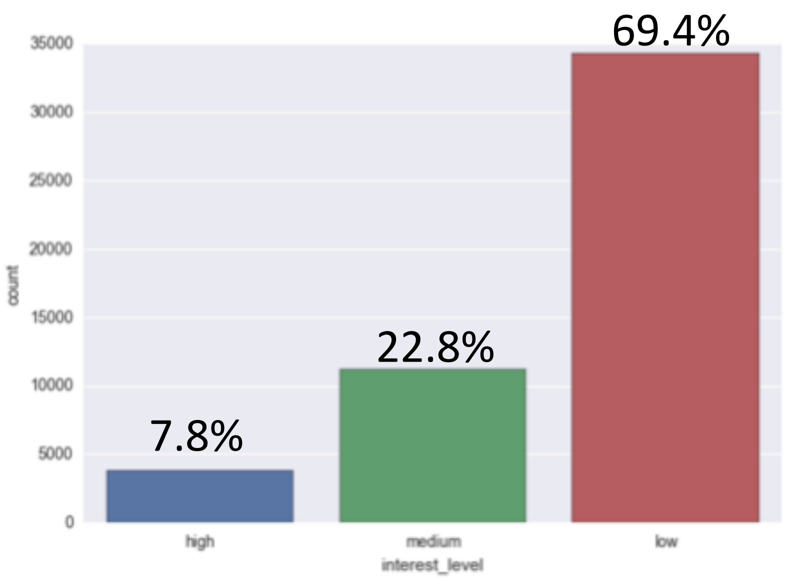
- Amenities
- Photos
- Textual description

3 target classes: “high”, “medium”, and “low”. (Ordinal Classification)

Key challenge: How to orchestrate different types of input features

More about Data Set

- About 50k listings (samples), 14 raw features (seemingly medium size)
  - #amenities/listing: 5.4 (avg), 39 (max)
  - #photos/listing: 5.6 (avg), 68 (max)
  - Description length: 90 words on average, up to 667 words
  - Actually very large
- Imbalanced class samples
  - But equal penalty if mispredicted
- Missing data
  - 7.3% have no photos
  - 16.8% have no building ID
  - Systematically missing
- Outliers
  - May be corrected



Outline

- Problem Description
- Our Approach
  - Data Pre-processing
  - Model Selection
- Preliminary Results
- Lessons Learned

Location

- Extract zipcode from address
- Add external data based on zipcode
  - Population
  - Average income
  - Physical area
- Get adjusted price
  - Use KNN to find out average price for similar floorplans
  - Get the ratio of actual price to average price as adjusted price

Building & Manager Reputation Extraction

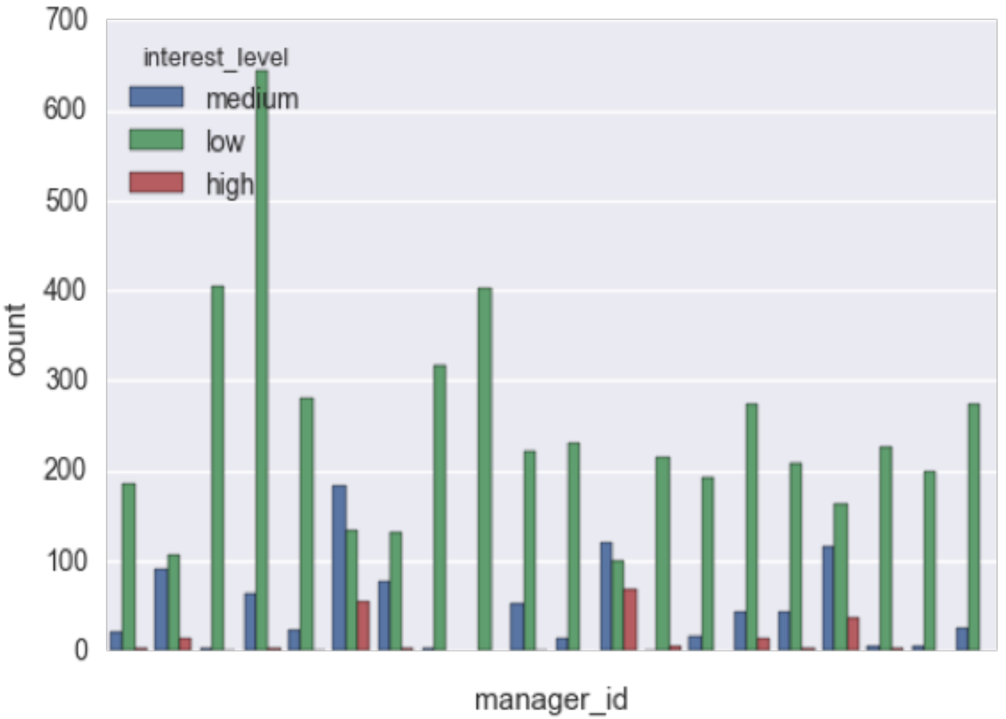
Group listings by manager/building ID

Think of distributions of classes as a "prior"

Example:

- Consider manager as "good" if

$$\frac{\#high_{manager}}{\#total_{manager}} > \frac{\#high_{dataset}}{\#total_{dataset}}$$



Amenities

- 2 approaches to process the list of features :

  1. TF-IDF score
  2. Extract common features

- Method 2 yields better results with GradientBoost based on experiments

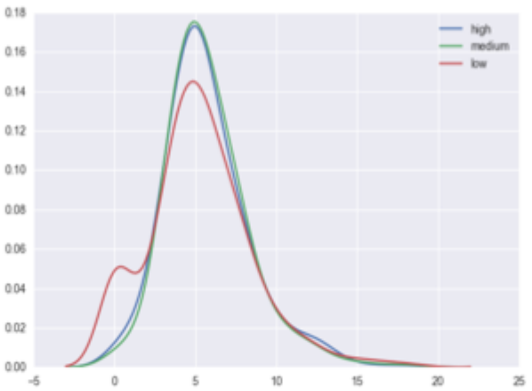
Photos

Approach 1: Count number of pictures

- Pros: simple
- Cons: missing information

Approach 2: Image classification using CNN

- Problem: 10 good photos + 1 bad photo= low interest
- Workaround: Manually select photos from each class
- Workaround: Use retraining to deal with small data set



Sample photos from low interest class

Manually selected photos from low interest class

Outliers

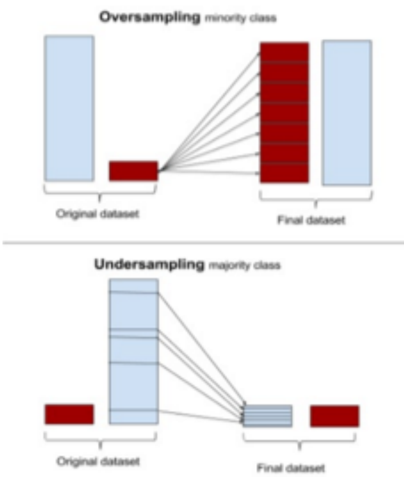
1807 data points have suspicious attributes:

- 0 longitude/latitude, or coordinates not in Manhattan
- Surprising low/high prices ($43 in Manhattan?)
- Strange floorplan(0 BR 10 BA?)

Fix them by looking at other attributes, and remove unfixable data points

Resampling

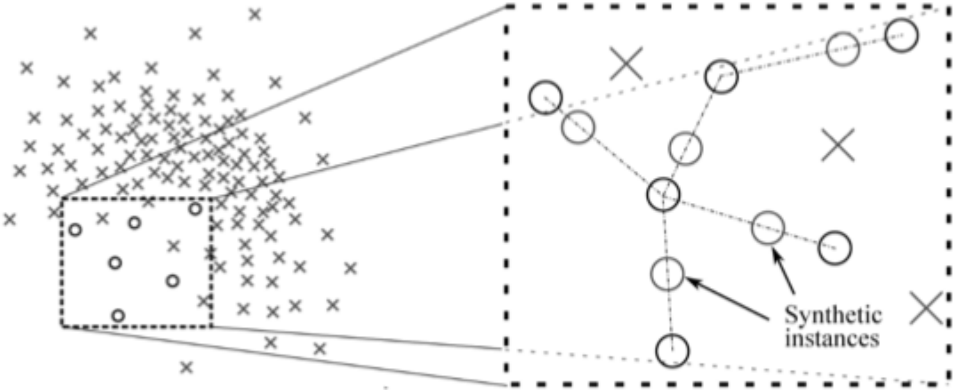
- Many classification algorithms will only perform optimally when number of samples in each class is roughly equal.
- Resampling can help offset this imbalance and arrive at a more robust and fair decision boundary.
- Resampling methods usually fall into one of three categories:
  - Under-sampling – removing instances of the majority class
  - Oversampling -increasing number of instances of minority class
  - Ensemble



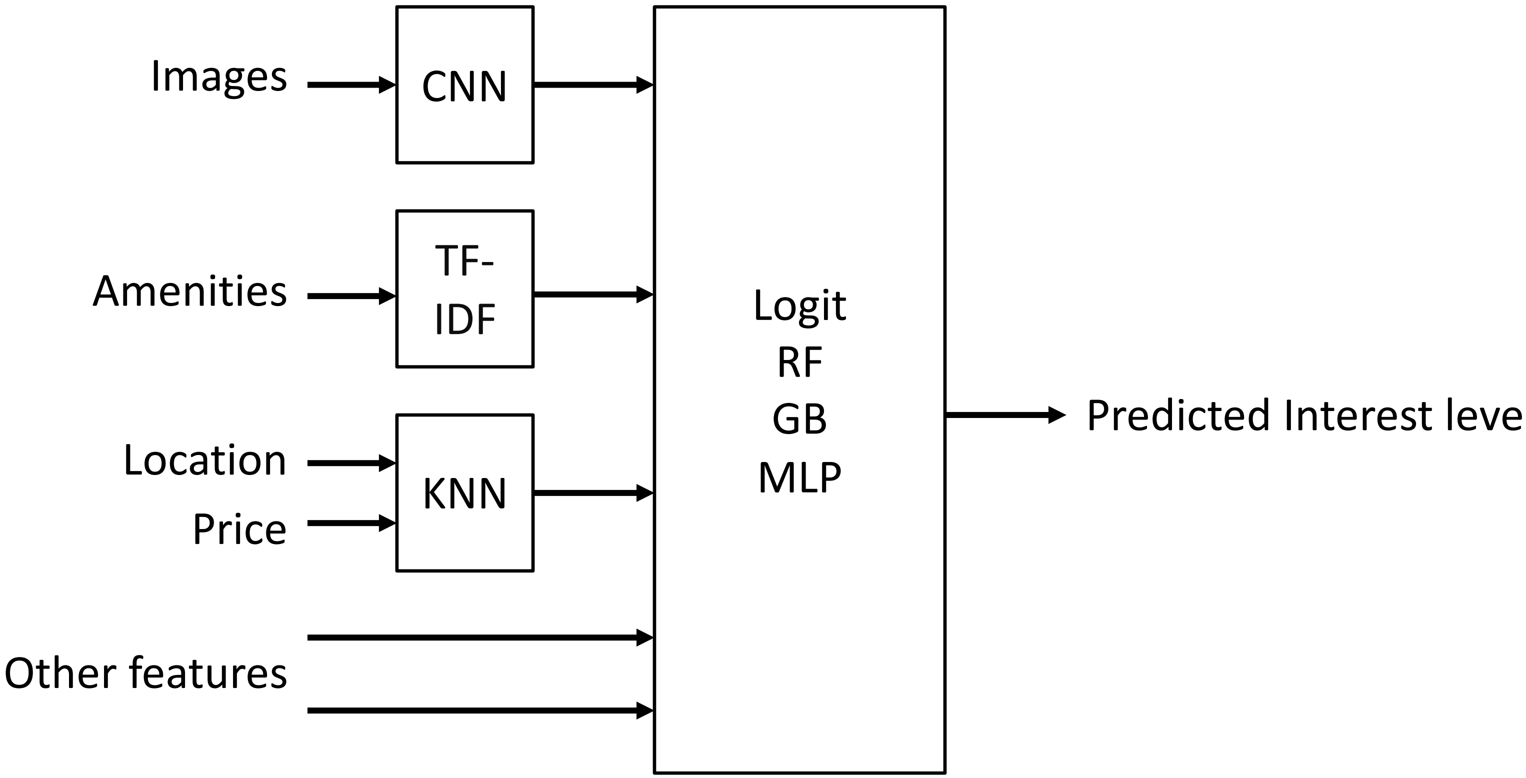
- Source: http://contrib.scikit-learn.org/imbalanced-learn/index.html

Synthetic Minority Over-Sampling Technique (SMOTE)

- Avoids creating multitude of redundant data.
- For each data point of minority class, find KNN and randomly create "synthetic" data point on vector between each neighbor.
- Can tune K and ratio of minority classes to majority class.



Model Overview



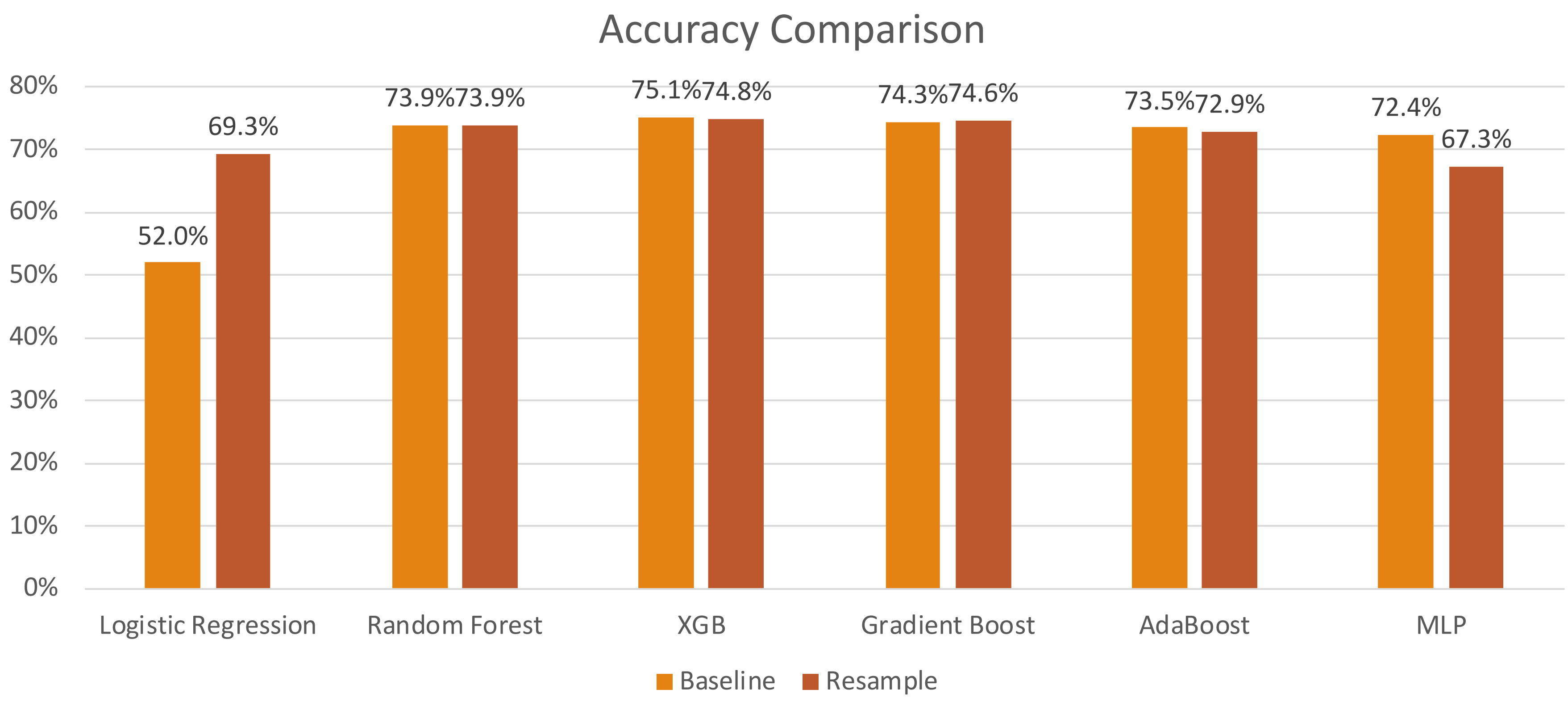
Model Selection

Model	Pros	Cons
Logistic Regression	Regression nature	Relies on monotonicity
Ensemble	Known for good accuracy	Many hyper parameters
Neural Network	Known for good accuracy	Poor interpretation

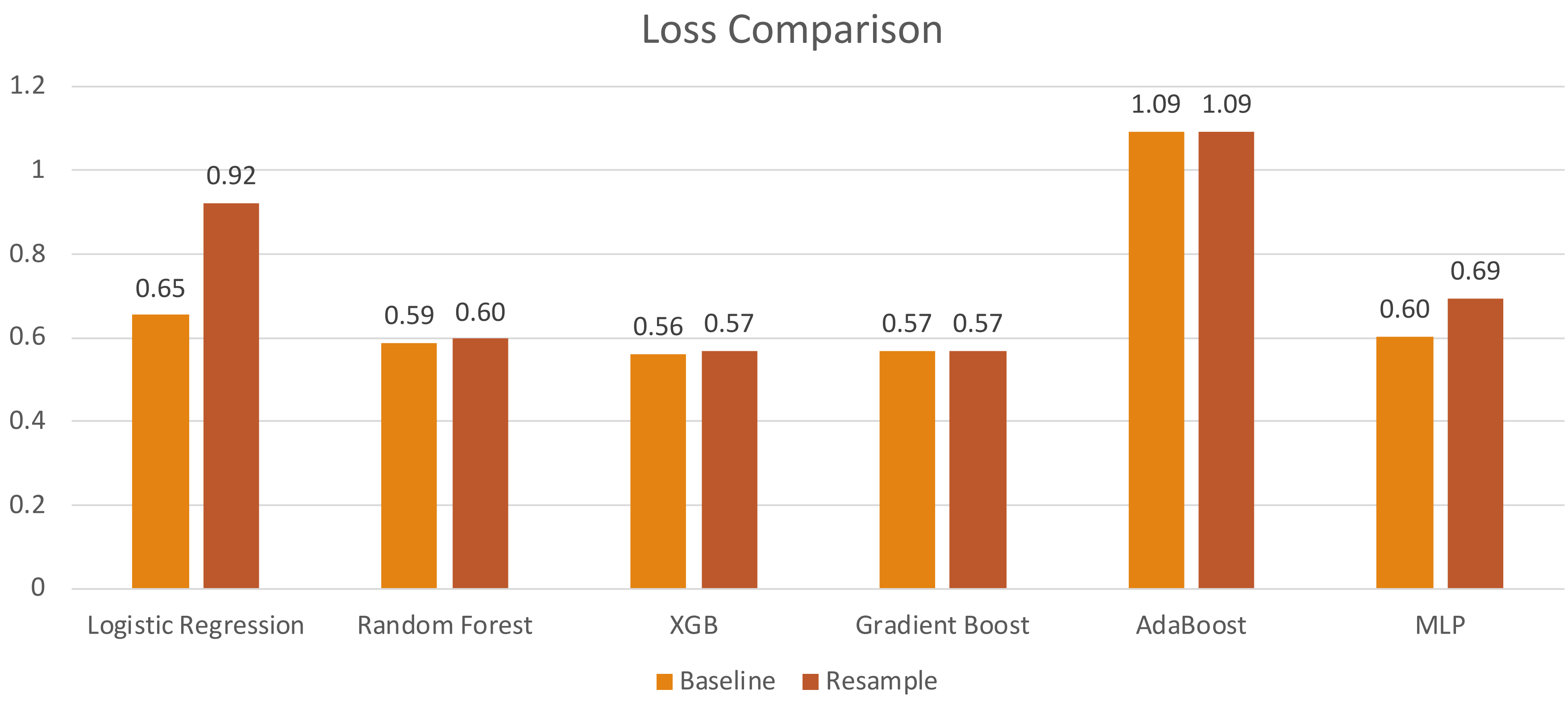
Outline

- Problem Description
- Our Approach
  - Data Pre-processing
  - Model Selection
- Preliminary Results
- Lessons Learned

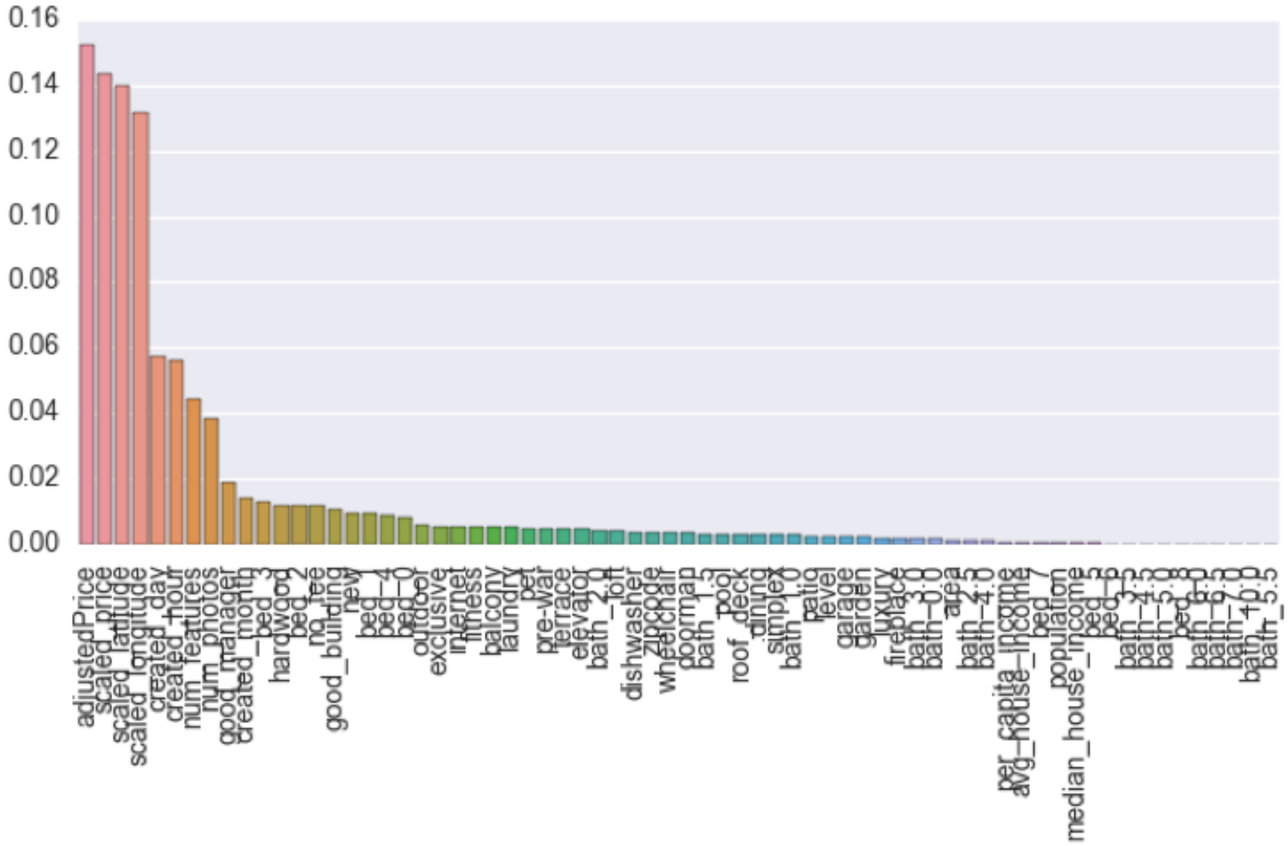
Results--Accuracy



Results--Loss

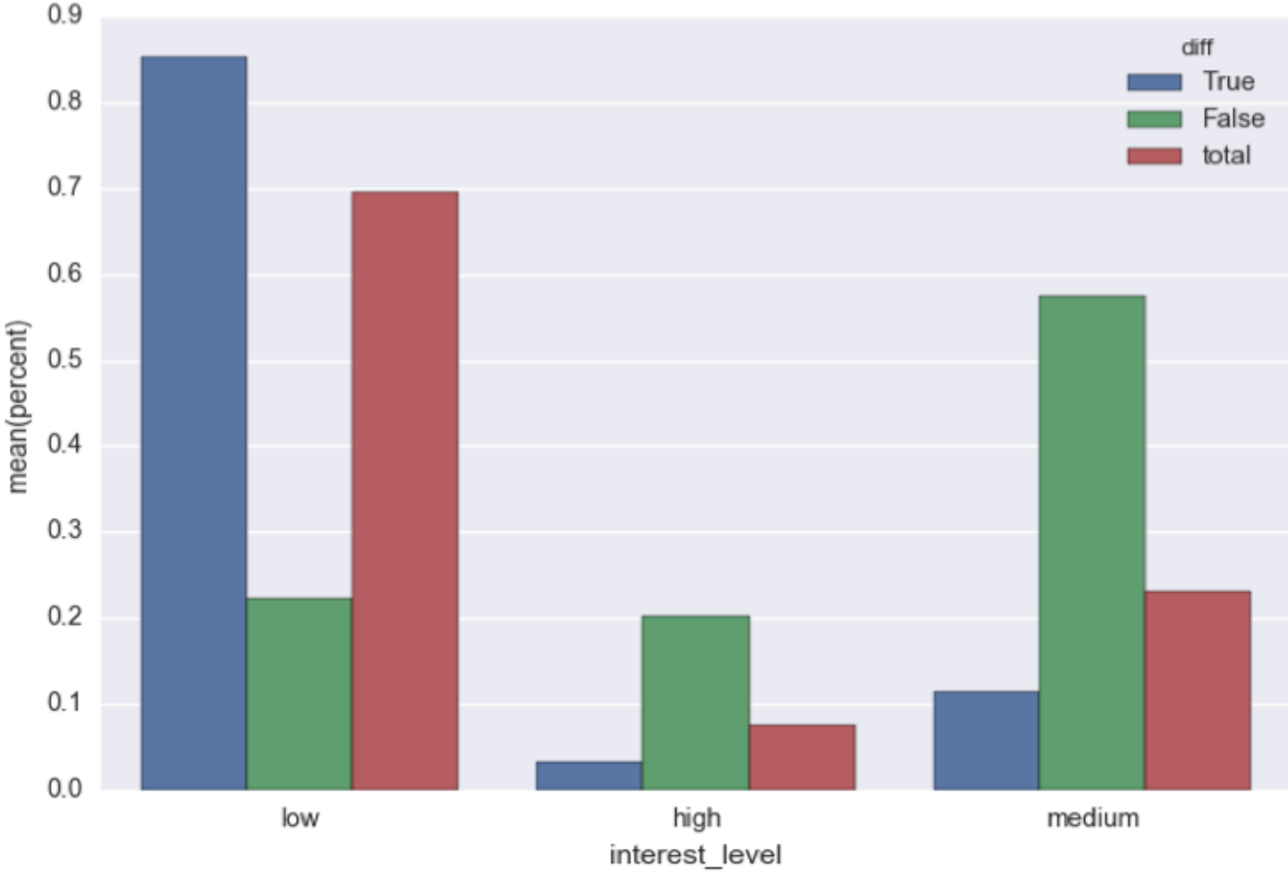


Feature Importance



Takeaways: Money Matters Most; Timing is also important

Sensitivity and Specificity



Our current model is too pessimistic

Outline

- Problem Description
- Our Approach
  - Data Pre-processing
  - Model Selection
- Preliminary Results
- Lessons Learned

Lessons Learned

- Data mining is a cooperation between human and machine
  - Manually select typical photos to train
  - Manually extract synonymous amenities
- For image classification, retraining should be a preferred path
  - Smaller data set, faster training, decent accuracy
- Tensorflow is hard to use
  - Neural network, as a tool, is still in its infancy stages
- Tips for fast sublease
  - Low Price
  - Good location
  - Right timing

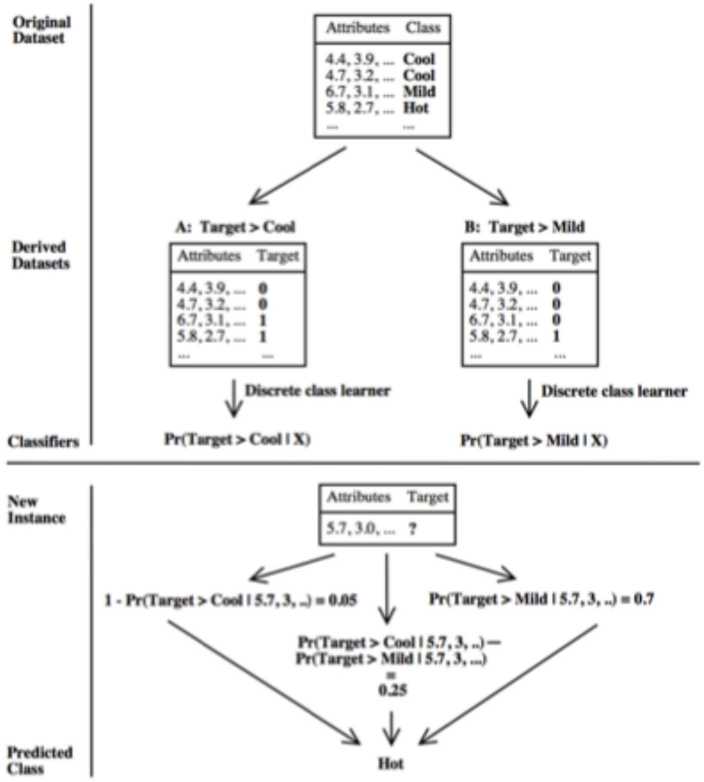
Ongoing/Future Work

- Further tuning of resampling parameters/devising an ensemble approach
  - Investigate why resampling does not help
- Ordinal classification
  - Target classes in our problem is ordered
  - “high” > “medium” > “low”
  - Ordinal classification can leverage ordered target
- Devising an ensemble approach

Q & A

Ordinal Classification

- Target classes are ordered
  - “high” > “medium” > “low”
- Misprediction can be quantified in terms of class distance
  - Mispredicting a low-interest sample to be high is worse than mispredicting it to be medium
- Approach 1: Associate cost function with class distance
- Approach 2: Use multiple 2-class classifier*



*Frank, Eibe, and Mark Hall. "A simple approach to ordinal classification." *European Conference on Machine Learning*. Springer Berlin Heidelberg, 2001.