**Project Report**

**1. Introduction**

# Two Sigma Connect: Rental Listing Inquiries

What is the problem? Why is it important? What is your basic approach? A short discussion of how it fits into related work in the area is also desirable. Summarize the basic results and conclusions that you will present.   
  
**2. Problem Definition and Algorithm**   
  
2.1 Task Definition

Precisely define the problem you are addressing (i.e. formally specify the inputs and outputs). Elaborate on why this is an interesting and important problem.   
  
2.2 Algorithm Definition

### Preliminaries: decision tree learning

Decision trees are a popular method for various machine learning tasks. because it is invariant under scaling and various other transformations of feature values, is robust to inclusion of irrelevant features. However, they are seldom accurate. In particular, trees that are grown very deep tend to learn highly irregular patterns: they overfit their training sets, i.e. have low bias, but very high variance. Random forests are a way of averaging multiple deep decision trees, trained on different parts of the same training set, with the goal of reducing the variance. This comes at the expense of a small increase in the bias and some loss of interpretability, but generally greatly boosts the performance in the final model.

### Tree bagging

The training algorithm for random forests applies the general technique of [bootstrap aggregating](https://en.wikipedia.org/wiki/Bootstrap_aggregating), or bagging, to tree learners. Given a training set *X* = *x1*, ..., *xn* with responses *Y* = *y1*, ..., *yn*, bagging repeatedly (*B* times) selects a [random sample with replacement](https://en.wikipedia.org/wiki/Bootstrapping_(statistics)) of the training set and fits trees to these samples:

For *b* = 1, ..., *B*:

1. Sample, with replacement, *B* training examples from *X*, *Y*; call these *Xb*, *Yb*.
2. Train a decision or regression tree *fb* on *Xb*, *Yb*.

After training, predictions for unseen samples *x'* can be made by averaging the predictions from all the individual regression trees on *x'*: or by taking the majority vote in the case of decision trees.

This bootstrapping procedure leads to better model performance because it decreases the [variance](https://en.wikipedia.org/wiki/Bias%E2%80%93variance_dilemma) of the model, without increasing the bias. This means that while the predictions of a single tree are highly sensitive to noise in its training set, the average of many trees is not, as long as the trees are not correlated. Simply training many trees on a single training set would give strongly correlated trees (or even the same tree many times, if the training algorithm is deterministic); bootstrap sampling is a way of de-correlating the trees by showing them different training sets.

### From bagging to random forests

The above procedure describes the original bagging algorithm for trees. Random forests differ in only one way from this general scheme: they use a modified tree learning algorithm that selects, at each candidate split in the learning process, a [random subset of the features](https://en.wikipedia.org/wiki/Random_subspace_method). This process is sometimes called "feature bagging". The reason for doing this is the correlation of the trees in an ordinary bootstrap sample: if one or a few features are very strong predictors for the response variable (target output), these features will be selected in many of the *B* trees, causing them to become correlated. An analysis of how bagging and random subspace projection contribute to accuracy gains under different conditions.

Typically, for a classification problem with *p* features, √*p* (rounded down) features are used in each split.

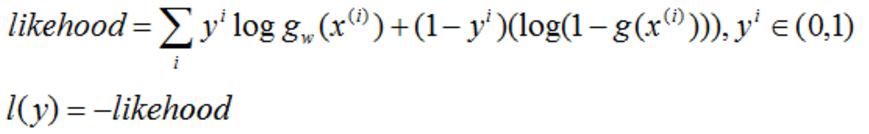
Data: bathrooms, bedrooms, latitude, longitude, price, num\_photos, num\_features, num\_description\_words, created\_year, created\_month, created\_day

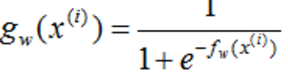
**3. Experimental Evaluation**   
  
3.1 Methodology

I use the logarithmic loss to evaluate my result. Log loss, aka logistic loss or cross-entropy loss.

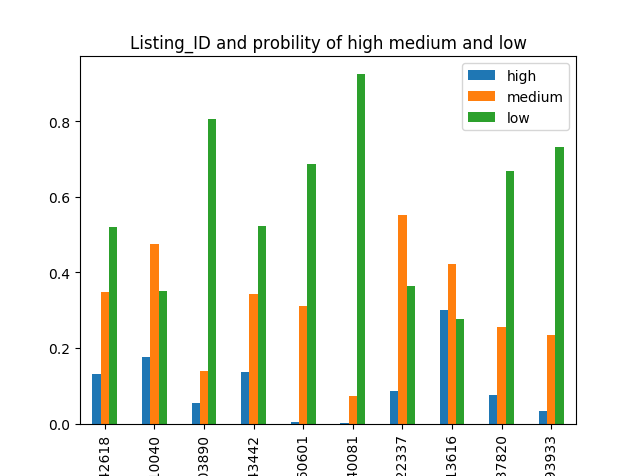
In order to calculate Log Loss the classifier must assign a probability to each class rather than simply yielding the most likely class. Mathematically Log Loss is defined as

It suits to multi-class tasks.





3.2 Results

Present the quantitative results of your experiments. Graphical data presentation such as graphs and histograms are frequently better than tables.   
  
3.3 Discussion

**4. Related Work**

Why is your problem and method better?

Use bagging, it decrease the overfitting and improve the performance of training  
  
**5. Future Work**

What are the major shortcomings of your current method? For each shortcoming, propose additions or enhancements that would help overcome it.   
  
**6. Conclusion**   
Briefly summarize the important results and conclusions presented in the paper. What are the most important points illustrated by your work? How will your results improve future research and applications in the area?   
  
**Bilbiography**   
Be sure to include a standard, well-formated, comprehensive bibliography with citations from the text referring to previously published papers in the scientific literature that you utilized or are related to your work.