Data Mining: 36-462/36-662
Final Project

AirBnB Predictions

Deliverables and deadlines

Here are some key deliverables and deadlines upfront:

- Your predictions are due April 30th at 11:59pm, submitted on canvas.
- Your writeup is due May 3rd at 1pm, submitted on canvas.

Each team only has to submit one of everything. Hence you can read "you" throughout as "your team".

Disclaimer

This project is an open-ended work-in-progress. If you find any issues with the data or have difficulty interpreting something in the following description please post on Piazza and we will update this document as necessary.

Introduction

Your final project concerns various prediction tasks on AirBnB data from listings in Seattle. The data is from the following website:`http://insideairbnb.com/get-the-data.html`.

We are interested in understanding the dataset better through the lens of classification/regression/clustering algorithms:

1. We would like to predict the PRICE for listings. A natural use case for this regressor would be in helping people price new listings. This task uses the file `price.csv`.
2. Each AirBnB listing also has a REVIEW SCORE. I have discretized this variable into two classes $\{0, 1\}$ corresponding to {`undesirable, desirable`}. Our goal will be to design a classifier for this variable. Use the file `review.csv`.

Your project has a few parts:

1. Data exploration and pre-processing.
2. Building and validation of predictive algorithms.
3. Actual submission to a prediction contest.

4. Some follow-up analysis of your results.

5. Two (optional) extra-credit exercises.

More details

Download the files `price.csv` and `review.csv` from canvas and load them into your R session. This is the data that you will use for model building. Your prediction targets are the `price` variable, and the `review_scores_rating` variable.

Data pre-processing: You will need to do some pre-processing of your data. Particularly, consider appropriately encoding the categorical variables (and possibly drop/combine some of the categories). You will also need to turn the longitude-latitude features into a more useful form. One way to do this is by creating new features corresponding to the distance to popular landmarks (like pike place, the space needle, downtown and so on).

General advice for training and tuning your predictors: Split your data into a train and a validation set. Use the train set to fit and the validation set to evaluate and select a good model. Use small subsets of the data initially until you get a feeling for what works and what does not.

I have also provided you with a test set for each task. Using an external source to obtain the labels for the test set and using this to tune your model is considered cheating. Do not do this.

Making predictions

How can you make your predictions? You can use any of the techniques we have discussed in class. You can use any of the variables in the data set, and you can also consider constructing new variables by combining or transforming the variables that are present in the data set. You should not use external information sources for this project.

Submitting predictions

You will submit a single RData file with your predictions. This file should contain the following variables:

1. `price.guesses`: A single vector of price estimates. This vector should be as long as the number of test cases in `price_test.csv` (and in the same order).

2. `price.mse`: A single number indicating your best guess at the mean squared error for your predictions on the test set. This will give us an idea of how well your validation

has worked in setting your expectations. Good estimation of your performance is one goal of the project.

3. `review.guesses`: A single binary $\{0, 1\}$ vector of predictions of whether the test listings are undesirable or desirable. This vector should be as long as the number of test cases in `review_binary_test.csv` (and in the same order).

4. `review.acc`: A single number indicating your best guess at the 0/1 error for your predictions on the test set.

5. `team.name`: A string with your team's name. These will be revealed in class, so make it anonymous if you wish. Your report will link the team name to individual names for grading purposes.

To make this file, if you have the appropriate variables in your workspace, you can type

```
save(list=c("price.guesses","price.mse","review.guesses",
"review.acc", "team.name"),file="stat462final.RData")
```

This will create `stat462final.RData` file, which you can upload on canvas. (Please rename the file to include your team name before sending.)

Write-up

Along with your contest entries, you will submit a write-up of your work. This write-up should be a polished report, with figures and snippets of R code as you deem helpful. You don't need to submit your R code in its entirety. Your report should have the following sections (you can of course add subsections if you want), and should be no more than 8 pages.

Introduction: Describe your data set. What is the problem you are trying to solve? This can be quite brief.

Exploration: Exploration of your data. You don't need to do the typical "exploratory data analysis" that you might do in 36-401, but you should provide proper motivation for your work and explain any insights and exploration that led to your features and models. This can include unsupervised approaches that we learned in the last part of the term if you detect any interesting structure with them. (If you don't find anything interesting, then just describe what you tried. You don't need to artificially manipulate the data to find something that's not there.)

Supervised analysis: How did you make your predictions? Describe this process in detail. Again, you can use any of the classification/regression techniques that we learned in the first half of the course, or any other techniques as long as they are adequately described. What predictor variables did you include? How did you engineer features from the data? What technique did you use for prediction, and why did you choose it? If there were tuning parameters, how did you pick their values? Can you explain anything about the nature of the relationship between the predictors in your model and the predictions themselves?

Analysis of results: Once you have predictors you happy with, you should think about their performance a little more. At the least, you should try to address the following questions:

- What kinds of listings do you do well or poorly on?
- Suppose you had more time, what would your next steps be, i.e. what would you like to try next to improve your predictive performance?

Any other analysis of your results would be welcome here.

Evaluation

Your predictions will be evaluated against the true delays. You will be judged based on your MSE for the regression task and your mis-classfication error for the classification task.

The results from both contests will be revealed on the last day of class. The top 4 teams from each contest will receive extra credit. These teams will be asked to describe their prediction approaches and what worked. (Comments from everyone else are also welcome!)

Extra-Credit

You can choose to this (optional) exercise. The amount of extra-credit will be a function of the amount of effort you put in to this exercise but will be between 0 and 20% of the project grade.

1. Visit the website `http://insideairbnb.com/get-the-data.html`. Download the (detailed) `listings.csv` data for another city. This time you will need to do some pre-processing. In particular, remove any columns that do not exist in the datasets we gave you, if the cleaning fee is not present for a listing fill in 0, drop data points that are incomplete after this. You will need to re-think how you use the latitude-longitude features for this new city.

   Repeat your predictive analysis for this new city. To elaborate build and tune a classifier for `review_scores_rating` (to binarize it drop any listing with score > 90 and < 95 and label ≤ 90 as 0 and ≥ 95 as 1) and regressor for `price`).

   Comment specifically on the differences in the predictors for your chosen two cities (what features are differentially important and try to give reasonable explanations for this).

Cheating

Don't cheat. We know that there are ways to cheat on this final project. If we suspect you of cheating (e.g., if you have a remarkably low misclassification rate/low mean squared error, but your method is not really statistically motivated), then we reserve the right to give you a 0.