Data Analytics 350 Airbnb Lab

Sally Hyde, Matthew Rinker

***Abstract*— In this report we explore a dataset taken from Airbnb and attempt to predict listing price from a variety of predictor variables. Our analysis is based upon linear regression and optimizing a model via principle component analysis.**

I. INTRODUCTION

In this lab, we expanded upon our dataset by conducting a sentiment analysis on the reviews for the Airbnb listings. We then used these values along with various measures about the listing to create a linear regression model that would predict the price. We did this first through a linear regression with all our response variables, and then we conducted principle component analysis (PCA) to reduce the number of response variables. We found that these variables actually had less impact on the price of a listing than we would expect. In fact, while we got a decent result from our first attempt with a generic linear model, we found that when testing under various cross validation methods, our model actually had a very poor fit. This was further reinforced after conducting our PCA.

II. DATA

The data used for this investigation was split over three data sets. The first called calendar was data taken from airbnb showing price data for specific listings and their availability. This dataset consisted of 1,308,890 rows and 4 columns. We did not incorporate the calendar dataset into our analysis although it was provided. The second dataset called listings included data about each specific listing as identified by their listing id. This dataset consisted of 3,585 rows and 95 columns. The final dataset was called reviews and was made up of data about each individual review about listings as identified by the listing id number. This dataset consisted of 68,275 rows and 6 columns.

To start, we did introductory analyses of the data. Looking at every numerical variable in our data, we identified variables which would be important to our methods later. To accomplish this we calculated some basic measures about each variable; we found the median value, mean value, minimum value, maximum value, standard deviation, and the variance of each variable. See Fig. 1 for this data.

	Median	Mean	Min	Max	Stdev	Var
accommodates	2.00	3.041283	1.00	16.00	1.778929e+00	3.164590e+00
availability_30	4.00	8.649930	0.00	30.00	1.043533e+01	1.088961e+02
availability_365	179.00	179.346444	0.00	365.00	1.421362e+02	2.020269e+04
availability_90	37.00	38.558159	0.00	90.00	3.315827e+01	1.099471e+03
bathrooms	1.00	1.221647	0.00	6.00	5.014871e-01	2.514893e-01
bedrooms	1.00	1.255944	0.00	5.00	7.530596e-01	5.670988e-01
beds	1.00	1.609060	0.00	16.00	1.011745e+00	1.023627e+00
cleaning_fee	50.00	68.380145	5.00	300.00	5.129783e+01	2.631468e+03
extra_people	0.00	10.886192	0.00	200.00	1.913777e+01	3.662543e+02
guests_included	1.00	1.429847	0.00	14.00	1.056787e+00	1.116799e+00
host_acceptance_rate	94.00	84.173089	0.00	100.00	2.177925e+01	4.743359e+02
host_listings_count	2.00	58.902371	0.00	749.00	1.711197e+02	2.928194e+04
host_response_rate	100.00	94.989082	0.00	100.00	1.251769e+01	1.566925e+02
host_total_listings_count	2.00	58.902371	0.00	749.00	1.711197e+02	2.928194e+04
maximum_nights	1125.00	28725.836820	1.00	99999999.00	1.670136e+06	2.789354e+12
minimum_nights	2.00	3.171269	1.00	300.00	8.874133e+00	7.875023e+01
monthly_price	2925.00	3692.097973	500.00	40000.00	2.899964e+03	8.409790e+06
number_of_reviews	5.00	19.044630	0.00	404.00	3.557166e+01	1.265343e+03
price	150.00	198.303004	10.00	7163.00	1.799912e+02	3.239681e+04
review_scores_accuracy	10.00	9.431571	2.00	10.00	9.318632e-01	8.683691e-01
review_scores_checkin	10.00	9.646293	2.00	10.00	7.627532e-01	5.817925e-01
review_scores_cleanliness	10.00	9.258041	2.00	10.00	1.168977e+00	1.366507e+00
review_scores_communication	10.00	9.646549	4.00	10.00	7.355070e-01	5.409705e-01
review_scores_rating	94.00	91.916667	20.00	100.00	9.531686e+00	9.085303e+01
review_scores_value	9.00	9.168234	2.00	10.00	1.011116e+00	1.022356e+00
reviews_per_month	1.17	1.970908	0.01	19.15	2.120561e+00	4.496781e+00
security_deposit	250.00	324.698212	95.00	4500.00	3.288731e+02	1.081575e+05
weekly_price	750.00	922.392377	80.00	5000.00	6.578218e+02	4.327295e+05

Fig. 1. Table containing data about each variable

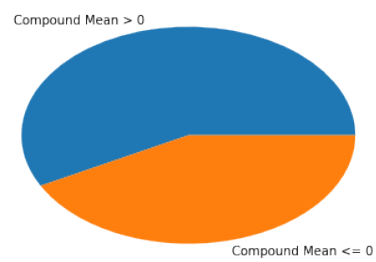

Fig. 2. Sentiment Analysis Results

III. RESULTS

To conduct our investigation into the Airbnb dataset we first calculated some new variables to aid our predictive model using sentiment analysis. We did this through two separate methods. First, we used the Python natural language tool kit (nltk) to provide a sentiment analysis score for each review. This was categorized into four separate scores: negativity, positivity, neutrality, compound. We also hand-coded a more simple version of a classifier. This used two corpuses provided with the dataset divided into positive words and negative words. We then calculated a simple positivity score and a simple negativity score. The formula for these scores were as follows: positivity_simple $= \frac{\text{\# positive words}}{len(\text{review})}$ and negativity_simple $= \frac{\text{\# negative words}}{len(\text{review})}$.

We then took these scores for each review and aggregated them for each listing. This gave us a mean score in each of the following categories: negativity, neutrality, positivity, compound, simple positivity, and simple negativity. In Fig. 2 we have created a pie chart to show the proportion of listings which had compound means above and below zero.

We then took these new variables and added them to our listings dataset. These variables along with more review data would become our explanatory variables for our linear regression. We first ran a regression using all of our explanatory variables. The statistics of the resulting model can be

R^2	Adjusted R^2	# Observations	DF
0.704	0.703	3585	10

Fig. 3. Data for the Linear Regression Model

Variable	Coeff	Std Err	P
host_response_rate	0.2150	0.147	0.144
review_scores_rating	1.3622	0.421	0.001
review_scores_accuracy	-7.7411	3.246	0.017
review_scores_cleanliness	17.2991	2.830	0.000
review_scores_checkin	-6.4321	3.783	0.089
review_scores_comm	0.8751	3.963	0.825
positivity_mean	215.8845	78.441	0.006
negativity_mean	-103.9290	31.241	0.001
positivity_simple_mean	190.3089	96.403	0.048
negativity_simple_mean	-9.6821	8.440	0.251

Fig. 4. Coefficiant Data for the Linear Regression Model

seen in Fig. 3 and the statistics of each response variable can be seen in Fig. 4.

Next, we tested the fit of our model using cross validation with various cost functions. The results of this can be seen in Fig. 5. As you can see, the cost function results are not particularly promising, which is interesting considering a generalized linear model of the variables produced fairly good results.

Additionally, the correlation between each of our response variables can be found in Fig. 6.

Next, since there were three main groups of explanatory variables that determined price, we used PCA to reduce the number of explanatory variables in our data set and ran another linear regression. We split our dataset in a 70:30 split for our training and test sets for this evaluation. The resulting model can be found in Fig. 7. The coefficient data for the PCA variables can be found in Fig. 8.

We also created another correlation plot to show the correlation of each PCA variable with each other. This plot can be see in Fig. 9.

Finally, we ran our PCA reduced set through the cross validation methods we looked at before. The results of this can be found in Fig. 10. Interestingly, this method performed far worse than a standard linear regression. This is visible in the R^2 values of both models. The standard linear regression with all response variables performed far better with an R^2 value of 0.75 versus the R^2 of 0.002 from our PCA model. However, this does not necessarily mean that the generic linear model is better. Outside factors could have played a role as well as overfitting from the training set. Ultimately, this is our most interesting result as it shows that as of now

	mape	mse	r_squared	rmse
Test/Train	71.389242	12062.598845	0.029287	109.829863
KFold	71.589634	12058.478554	0.029618	109.811104
LOOCV	71.861671	12044.880427	0.030713	109.749171

Fig. 5. Cost Functions vs. Cross Validation Method

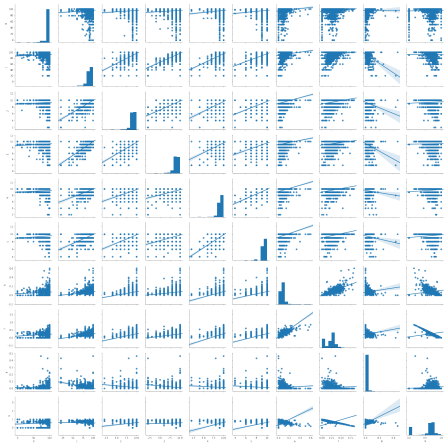

Fig. 6. Correlation Between Response Variables

R^2	Adjusted R^2	# Observations	DF
0.003	-0.000	1076	3

Fig. 7. Data for the Linear Regression Model after PCA

there is no good model to predict the price of a listing. This coincides with our preconceived ideas surrounding the dataset. However, when running our own cost functions against the PCA dataset we got the opposite of the values we expected and of the values we had gotten when running the linear regression.

IV. CONCLUSION

From our exploration of the data from Airbnb we found that our preconceived notions were likely correct. We think that the variables chosen likely have little impact on the pricing of a listing. Instead we think that the main variables

Variable	Coeff	Std Err	P
x1	0.9231	3.024	0.760
x2	-6.2992	3.999	0.115
x3	3.2779	6.157	0.595

Fig. 8. Coefficiant Data for the Linear Regression Model after PCA

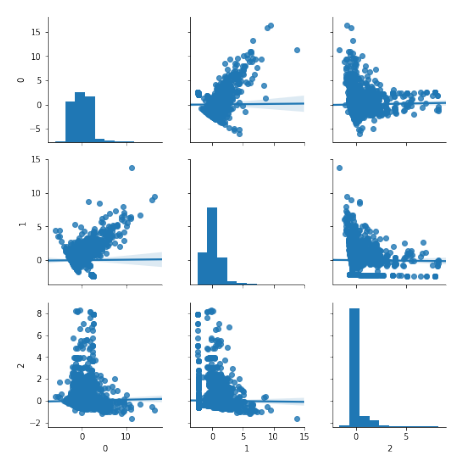

Fig. 9. Correlation Between Response Variables After PCA

Cost Function	Score
R^2	0.9997
MSE	3.4713
RMSE	0.0311
MAPE	100.0177

Fig. 10. PCA vs Cost Functions

that would affect pricing would be less quantifiable such as amenities, location, and time of year. There is very likely some omitted variable bias since our model lacks these significant factors. We believe that the omission of these variables made our model less effective. That being said, we acknowledge that the review score does have a small impact on the pricing, as guests would likely be more willing to pay a higher price for a listing with good reviews.