1. Executive Summary

AirBnB is a vacation housing company that has a host driven platform that connects would-be vacationers with hosts who can provide the ability to rent rooms, homes, or even couches all over the world. This company has revolutionized the way that people think and shop for vacation or general housing rentals. In order to be successful as a platform, it can be argued that this company needs to ensure that all of it’s hosts are of genuine quality and provide a valuable experience to the guests who pay for the service and rentals. In this report I have created a model and procedure for AirBNB to use in order to predict which hosts will provide a valuable experience for their guests. The goal for this analysis is to hopefully give the executive team insight into which hosts need improvements prior to guests booking in order to increase the ratings the hosts will likely receive. This benefits AirBnb by increasing the overall satisfaction of their guests and could have positive externalities such as reducing refunds or complaints.

In this analysis I examined the characteristics of 29,980 AirBnB hosts. I chose average rating as a measure of a host’s value to their guests and used characteristics such as the number of rooms or beds in a listing, whether the host is verified, the number of bathrooms available, the hosts’ acceptance rate, etc. After removing several variables for various reasons such as missing values or apparent usefulness of the responses I cleaned the data in order to ensure the my chosen procedures would run. After cleaning the data I decided that I would utilize an ensemble procedure to make my predictions and evaluate my procedure by comparing the RNSE of my procedure compared to the baseline RMSE of the airBnB data.

The models and procedures I chose for my analysis included a backward elimination procedure that yielded an RMSE of 7.40, a forward elimination procedure which yielded a RMSE of 7.40, a ridge procedure which yielded a RMSE of 7.40, a lasso procedure which yielded a RMSE of 7.44, a bagged forest procedure which yielded a RMSE of 7.40, and a random forest procedure that yielded a RMSE of 7.31. Only Our random forest procedure was able to beat our baseline RMSE of 7.36, however by using an ensemble procedure we can possibly get a slightly better result. After running the ensemble procedure I found that this approach had a 0.0069 better RMSE than my best individual procedure.

Through the ensemble procedure process I was able to gain some insight into which host characteristics seemed to impact average rating the most. Characteristics such as classifying as a SuperHost, host response rate, being business travel ready, requiring a license, and price seem to have positive effects on average rating. This information can be useful because airBnB can identify when these characteristics are low or missing from a new host and suggest the host make the changes prior to new guests booking.

II. The Data

The data used in this analysis contains 29,980 airBnB hosts. The dataset contains the following variables.

|  |  |  |
| --- | --- | --- |
| feature name | Description | variable type |
| index | a reference column of indices for each listing (numbered 1 to 29985) | numerical |
| avg\_rating | the average rating of the Airbnb property (out of 100) | numerical |
| accommodates | how many guests can stay in the listing | numerical |
| amenities | list of amenties available in the listing | categorical list |
| availability\_30 | how many days out of the next 30 the listing is available for | numerical |
| availability\_365 | how many days out of the next 365 the listing is available for | numerical |
| availability\_60 | how many days out of the next 60 the listing is available for | numerical |
| availability\_90 | how many days out of the next 90 the listing is available for | numerical |
| bathrooms | number of bathrooms in the listing | numerical |
| bed\_type | description of the bed | categorical |
| bedrooms | number of bedrooms in the listing | numerical |
| beds | number of beds in the listing | numerical |
| cancellation\_policy | description of how strict the cancellation policy is | categorical |
| city\_name | the broader metro area the listing is in | categorical |
| cleaning\_fee | how much, if any, is the cleaning fee the host charges | numerical |
| country | country | categorical |
| experiences\_offered | whether there are "experiences" offered with the listing (t) or not (f) | categorical |
| extra\_people | additional charge for extra people in the rental | numerical |
| first\_review | when the first review for the listing was written | date |
| guests\_included | how many guests are included in the price of the rental | numerical |
| host\_acceptance\_rate | percent of stay requests the host accepts | numerical |
| host\_has\_profile\_pic | if the host has a visible profile picture | categorical |
| host\_identity\_verified | whether the host's identity has been verified using Airbnb's process | categorical |
| host\_is\_superhost | whether the host is a "superhost" | categorical |
| host\_listings\_count | how many total listings the host has | numerical |
| host\_response\_rate | percent of stay requests the host responds to | numerical |
| host\_response\_time | how long it takes the host to respond to requests | categorical |
| host\_since | date the host joined Airbnb | date |
| host\_total\_listings\_count | how many total listings the host has ever had | numerical |
| host\_verifications | ways the host has verified their identity | categorical list |
| house\_rules | free text field describing the rules in the residence | text |
| instant\_bookable | whether you can instantly book the airbnb (t) or not (f) | categorical |
| is\_business\_travel\_ready | whether the listing is available for business travel (t) or not (f) | categorical |
| is\_location\_exact | whether the listing reports the exact location (t) or not (f) (usually for privacy purposes) | categorical |
| latitude | the number of degrees west of the prime meridien | numerical |
| longitude | the number of degrees north of the equator | numerical |
| maximum\_nights | maximum nights you can book the listing for | numerical |
| minimum\_nights | minimum nights you can book the listing for | numerical |
| monthly\_price | price to rent the listing for a month | numerical |
| neighborhood\_overview | free text field written by the host describing their neighborhood | text |
| price | price to rent the listing for one night | numerical |
| property\_type | description of the type of dwelling the listing is in | categorical |
| require\_guest\_phone\_verification | whether the host requires a phone number to verify the guest's ID (t) or not (f) | categorical |
| require\_guest\_profile\_picture | whether the host requires the guest's profile picture (t) or not (f) | categorical |
| requires\_license | whether the listing is in a jurisdiction that requires the host to have a license (t) or not (f) | categorical |
| room\_type | description of the type of accomodation of the listing | categorical |
| security\_deposit | the amount of security deposit required to rent the listing | numerical |
| summary | free text field summarizing the description of the listing | text |
| transit | free text field describing nearby transit options for the listing | text |
| weekly\_price | price to rent the listing for a week | numerical |

Several variables in this dataset contained many missing values. So much so, that the variables were not very useful. To decide which variables should be removed, we checked for any variables that were missing more than 20% of their values. The following variables were removed due to missing values or due to problems existing in regression or prediction procedures:

First\_Review

Index

Country

Host response time

Some variables made more sense to change rather than eliminate due to missing values. The following variables were altered if they had a missing value:

|  |  |
| --- | --- |
| Variable Name | Change made |
| Beds | If Missing Beds = 0 |
| Host acceptance rate | If missing Acceptance rate = 0 |
| Host Maximimum Nights | if missing, max nights = mean of max nights in data |
| Host minimum nights | if missing, minimum = 1 |
| price | if price is missing, price = median price of data |
| security deposit | if security deposit is missing, deposit =0 |
| monthly price | if monthly price is missing, monthly price = price\*30 |
| weekly price | if weekly price is missing, weekly price = price\*7 |
| is business travel ready | if business travel ready is missing, ==0 |
| host response rate | if host response rate is missing,== median response rate of data |

The remaining dataset contained 29,831 AirBnB hosts. The remaining dataset had no missing values and was partitioned into 70% training data and 30% testing data. Cross validation procedures were produced during the ridge and lasso regression procedures.

1. Data Analysis Procedures

The procedures performed in this report were run after the data cleaning and partitioning of the original dataset. These procedures were chosen in the hopes of not only obtaining a high level of accuracy but also retaining some inference power in order to gain insight into the characteristics that drive average rating.

The backward elimination procedure that was run with the remaining variables yielded a RMSE of 7.40. As can be referenced in Appendix 1, this procedure suggested that hosts who require license verifications and who are verified themselves tend to have higher average ratings. Interestingly hosts who have security deposits on their bookings as well as hosts who respond more tend to have higher average ratings as well. These seem to be characteristics that are easier to change as a host that do not require alteration to the physical rental space but could boost a hosts average rating and therefore increase guest satisfaction.

The forward elimination procedure agreed with out backward elimination procedure above. This procedure reported a RMSE of 7.40. As can be seen in Appendix2, the same variables were selected and seemed to suggest that hosts who are verified, require licenses, require security deposits and have a high response rate tend to have higher average ratings. Interestingly, it seems that hosts who are instant bookable and have multiple listings tend to have lower average ratings. It could be that hosts who have many listings and have set up instant booking have a more commercial approach to hosting and perhaps they lose the personal charm of community that airnB has in their mission. In a letter to shareholders airBnB stated that the key to their success is their community that creates a sense of belonging and positive community interaction. Commecialized airBnB hosts may lack the needed angagement that airBnB guests tend to enjoy.

Moving on to the more predictive models, the ridge procedure suggested that some of the characteristics that were significant in our inferential procedures may not be super impactful, even if they have a significant relationship with average rating. As referenced in Appendix 3, our ridge procedure weighted most of out significant chosen variables from our backward and forward elimination procedure very low. This could suggest that even though there is a strong level of significance in relation to average rating, the overall impact on average rating could be very low.

The lasso procedure gives us a little more insight into the level of impact these characteristics have on average rating. This is because our lasso procedure will merely remove variables that likely will not have a significant level of importance in determining the average rating of a host. For example, in our ridge and elimination procedures, we determined that host verification had a significant but low impact on our average rating. This had suggested that verifying your identity as a host could have a positive impact on your average rating. We can see in Appendix 4, after running the lasso procedure, it seems that verification was not useful enough to the algorithm to justify keeping it. This now suggests that where there is probably a positive relationship its impact on average rating of the host is not significant. Even so, it seems many of our other characteristics remained in our model. Requiring license, host response rate, and being business travel ready remain in our model of characteristics that benefit average rating. Conversely, our lasso procedure determined that security deposits did not impact our average ratings and characteristics that suggest a negative effect on average rating still include host listing count and instant bookable. Still we could maintain this theory that the more commercialized hosts may tend to have lower average ratings. The overall performance of this procedure seemed to fall compared to our other procedures so far, with a RMSE of 7.44 we can assume that perhaps the variables excluded contained some information that could be somewhat useful to the predictive power of the model, but perhaps not enough to justify keeping them.

After running the ridge and lasso procedures I ran two forest procedures, of these two procedures the random forest procedure performed the best of all procedures in this report. The random forest procedure was implemented by running 100 randomly generated trees with 5 randomly selected variables each. The overall RMSE of this procedure was significantly better than our previous procedures with a reported RMSE of 7.31. The variable importance seemed to agree mostly with our other procedures which can be seen in Appendix 5A.

The final procedure I performed was an ensemble procedure. This is a very simple procedure that utilizes an average of all the predictions made by my individual procedures and becomes a more accurate prediction. The prediction accuracy is evaluated the same as before, with a RMSE calculation. The ensemble procedure can be represented by the following table.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Individual Procedures RMSE | | | | | |
| Backward | Forward | Ridge | Lasso | BF | RF |
| 7.4062 | 7.4062 | 7.4037 | 7.4444 | 7.406 | 7.3178 |
| Ensemble Procedure RMSE | | | | | |
| 7.3109 | | | | | |

1. Conclusion

AirBnB wrote a letter to shareholders shortly after the launch of their IPO. In this letter, the founders of the company told its shareholders that much of what holds the company afloat even during the 2020 pandemic, is their community. AirBnB provides value to its guests by not only giving them access to great hosts, but allows them to fight off a generational loneliness by bringing people together and really allowing its guests to experience new communities through sharing the homes with their hosts. In this report we found potential evidence through the variables related to average rating that commercialized hosts who have many listings and use instant booking may be missing out on opportunities to provide guests with the sense of community and personal experience of sharing a home. Through the process of running an ensemble procedure, I determined that hosts who have a more commercial approach to renting out their property could be receiving lower average ratings. I also found that a host could potentially obtain higher average rating by changing minor aspects of the hosting on airBnB’s platform. Hosts can increase their response rate, verify identity, turn off instant booking, Require guest have licenses, and overall try to engage more with guests prior to booking to increase their average rating. If these steps can be taken prior to a guest beginning the booking process the value of guest experience and the hosts average customer satisfaction could increase. This benefits AirBnB because it not only increases customer satisfaction for it’s guests but also could reduce the rate of refunds or complaints if the average guest is rating their hosts better.

Appendix 1

Backwards Elimination Procedure

> summary(airbb.backward)

Call:

glm(formula = avg\_rating ~ availability\_365 + availability\_60 +

availability\_90 + bathrooms + bedrooms + beds + extra\_people +

host\_acceptance\_rate + host\_identity\_verified + host\_is\_superhost +

host\_listings\_count + host\_response\_rate + host\_since + instant\_bookable +

is\_business\_travel\_ready + latitude + longitude + maximum\_nights +

price + require\_guest\_phone\_verification + requires\_license +

security\_deposit, data = airbb\_rest)

Deviance Residuals:

Min 1Q Median 3Q Max

-77.661 -2.016 1.109 4.621 13.545

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.019e+02 3.643e+00 27.961 < 2e-16 \*\*\*

availability\_365 -3.469e-03 4.870e-04 -7.124 1.08e-12 \*\*\*

availability\_60 -4.697e-02 9.441e-03 -4.975 6.56e-07 \*\*\*

availability\_90 2.344e-02 6.463e-03 3.627 0.000287 \*\*\*

bathrooms 2.723e-01 1.026e-01 2.656 0.007923 \*\*

bedrooms 1.895e-01 8.767e-02 2.161 0.030690 \*

beds -3.593e-01 5.418e-02 -6.631 3.43e-11 \*\*\*

extra\_people -3.987e-03 2.316e-03 -1.721 0.085194 .

host\_acceptance\_rate 4.614e-01 2.175e-01 2.122 0.033887 \*

host\_identity\_verifiedTRUE 2.085e-01 1.166e-01 1.788 0.073752 .

host\_is\_superhostTRUE 3.400e+00 1.187e-01 28.655 < 2e-16 \*\*\*

host\_listings\_count -5.186e-03 8.249e-04 -6.287 3.31e-10 \*\*\*

host\_response\_rate 3.202e+00 4.062e-01 7.883 3.35e-15 \*\*\*

host\_since -2.237e-04 8.375e-05 -2.671 0.007575 \*\*

instant\_bookableTRUE -9.105e-01 1.116e-01 -8.157 3.62e-16 \*\*\*

is\_business\_travel\_ready 1.342e+00 2.003e-01 6.699 2.15e-11 \*\*\*

latitude -5.115e-02 1.324e-02 -3.862 0.000113 \*\*\*

longitude -5.426e-03 3.105e-03 -1.748 0.080550 .

maximum\_nights -5.171e-09 2.359e-09 -2.192 0.028394 \*

price 1.928e-03 3.255e-04 5.923 3.21e-09 \*\*\*

require\_guest\_phone\_verificationTRUE -6.182e-01 2.473e-01 -2.500 0.012424 \*

requires\_licenseTRUE 1.088e+00 1.437e-01 7.566 3.99e-14 \*\*\*

security\_deposit 2.163e-04 1.508e-04 1.435 0.151428

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 51.36465)

Null deviance: 1163780 on 20881 degrees of freedom

Residual deviance: 1071415 on 20859 degrees of freedom

AIC: 141539

Number of Fisher Scoring iterations: 2

Appendix 2

Forward Elimination Procedure

> summary(airbb.forward)

Call:

glm(formula = avg\_rating ~ host\_is\_superhost + availability\_60 +

requires\_license + instant\_bookable + host\_response\_rate +

price + availability\_365 + host\_listings\_count + is\_business\_travel\_ready +

longitude + beds + latitude + availability\_90 + bathrooms +

host\_since + require\_guest\_phone\_verification + bedrooms +

maximum\_nights + host\_acceptance\_rate + host\_identity\_verified +

extra\_people + security\_deposit, data = airbb\_rest)

Deviance Residuals:

Min 1Q Median 3Q Max

-77.661 -2.016 1.109 4.621 13.545

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.019e+02 3.643e+00 27.961 < 2e-16 \*\*\*

host\_is\_superhostTRUE 3.400e+00 1.187e-01 28.655 < 2e-16 \*\*\*

availability\_60 -4.697e-02 9.441e-03 -4.975 6.56e-07 \*\*\*

requires\_licenseTRUE 1.088e+00 1.437e-01 7.566 3.99e-14 \*\*\*

instant\_bookableTRUE -9.105e-01 1.116e-01 -8.157 3.62e-16 \*\*\*

host\_response\_rate 3.202e+00 4.062e-01 7.883 3.35e-15 \*\*\*

price 1.928e-03 3.255e-04 5.923 3.21e-09 \*\*\*

availability\_365 -3.469e-03 4.870e-04 -7.124 1.08e-12 \*\*\*

host\_listings\_count -5.186e-03 8.249e-04 -6.287 3.31e-10 \*\*\*

is\_business\_travel\_ready 1.342e+00 2.003e-01 6.699 2.15e-11 \*\*\*

longitude -5.426e-03 3.105e-03 -1.748 0.080550 .

beds -3.593e-01 5.418e-02 -6.631 3.43e-11 \*\*\*

latitude -5.115e-02 1.324e-02 -3.862 0.000113 \*\*\*

availability\_90 2.344e-02 6.463e-03 3.627 0.000287 \*\*\*

bathrooms 2.723e-01 1.026e-01 2.656 0.007923 \*\*

host\_since -2.237e-04 8.375e-05 -2.671 0.007575 \*\*

require\_guest\_phone\_verificationTRUE -6.182e-01 2.473e-01 -2.500 0.012424 \*

bedrooms 1.895e-01 8.767e-02 2.161 0.030690 \*

maximum\_nights -5.171e-09 2.359e-09 -2.192 0.028394 \*

host\_acceptance\_rate 4.614e-01 2.175e-01 2.122 0.033887 \*

host\_identity\_verifiedTRUE 2.085e-01 1.166e-01 1.788 0.073752 .

extra\_people -3.987e-03 2.316e-03 -1.721 0.085194 .

security\_deposit 2.163e-04 1.508e-04 1.435 0.151428

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 51.36465)

Null deviance: 1163780 on 20881 degrees of freedom

Residual deviance: 1071415 on 20859 degrees of freedom

AIC: 141539

Scoring iterations: 2

Appendix 3

Ridge Regression Procedure

> summary(airbb.ridge)

Length Class Mode

a0 100 -none- numeric

beta 3300 dgCMatrix S4

df 100 -none- numeric

dim 2 -none- numeric

lambda 100 -none- numeric

dev.ratio 100 -none- numeric

nulldev 1 -none- numeric

npasses 1 -none- numeric

jerr 1 -none- numeric

offset 1 -none- logical

call 4 -none- call

nobs 1 -none- numeric

>

>

Appendix 4

Lasso Regression Procedure

> summary(airbb.lasso)

Length Class Mode

a0 88 -none- numeric

beta 2904 dgCMatrix S4

df 88 -none- numeric

dim 2 -none- numeric

lambda 88 -none- numeric

dev.ratio 88 -none- numeric

nulldev 1 -none- numeric

npasses 1 -none- numeric

jerr 1 -none- numeric

offset 1 -none- logical

call 4 -none- call

nobs 1 -none- numeric

>

>

Appendix 5

Random Forest Procedure

> importance(random.forest)

%IncMSE IncNodePurity

accommodates 9.2063921 28060.335

availability\_30 5.6829870 26269.157

availability\_365 6.8742118 45538.324

availability\_60 7.7280368 33347.023

availability\_90 7.5930243 35894.079

bathrooms 5.2517470 16305.306

bedrooms 5.7771325 15948.311

beds 9.6330606 17859.469

cleaning\_fee 8.1826603 44828.294

extra\_people 3.0683547 29846.832

guests\_included 4.6668107 16650.621

host\_acceptance\_rate 2.2453347 4526.583

host\_has\_profile\_pic 2.8190858 382.585

host\_identity\_verified 0.1723703 10684.508

host\_is\_superhost 14.2791384 43958.806

host\_listings\_count 9.1407920 34197.023

host\_response\_rate 3.5992098 29455.018

host\_since 3.7270491 76297.132

host\_total\_listings\_count 9.9739573 31189.849

instant\_bookable 3.1366024 11489.544

is\_business\_travel\_ready 3.8532441 2986.688

is\_location\_exact 0.5094126 9825.872

latitude 7.7768225 78182.241

longitude 6.1964542 84994.550

maximum\_nights 3.1324144 35563.750

minimum\_nights 4.0650240 33337.268

monthly\_price 13.2522202 60914.196

price 12.8889266 54927.826

require\_guest\_phone\_verification 2.9442161 1955.712

require\_guest\_profile\_picture 3.1515035 1768.604

requires\_license 2.8111581 5408.569

security\_deposit 3.9796301 28527.359

weekly\_price 11.6365142 59817.712

>