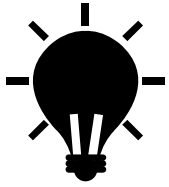
WASHINGTON DC AIRBNB HOSTING HELPER

By Rachael Friedman

TABLE OF CONTENTS

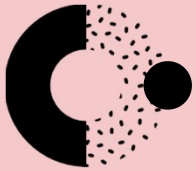







INTRO

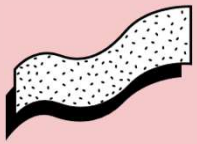
- Airbnb has been disrupting the hospitality industry since 2007.
- Hosts on Airbnb offer unique stays and local experiences for travelers that can't be replicated by a stay in a hotel.
- According to SmartAsset's 2020 study on the rental profits in 15 major US cities, the average expected annual profit of Airbnb hosts renting out a two-bedroom apartment after expenses is $20,619.
- For hosts renting out a one room in a two-bedroom home, they could expect to pay about 81% of their rent on average.
- Airbnb listing is available through Inside Airbnb and city venue information is available through the Foursquare API.



PROBLEM STATEMENT

The goal of this project is to help hosts understand what makes and Airbnb listing the most popular on the DC market and what to focus on to make their listing more competitive and increase their profits. There are a lot of apps out there to help hosts price their listing, but not a lot that look at what helps to make a listing popular in the first place.

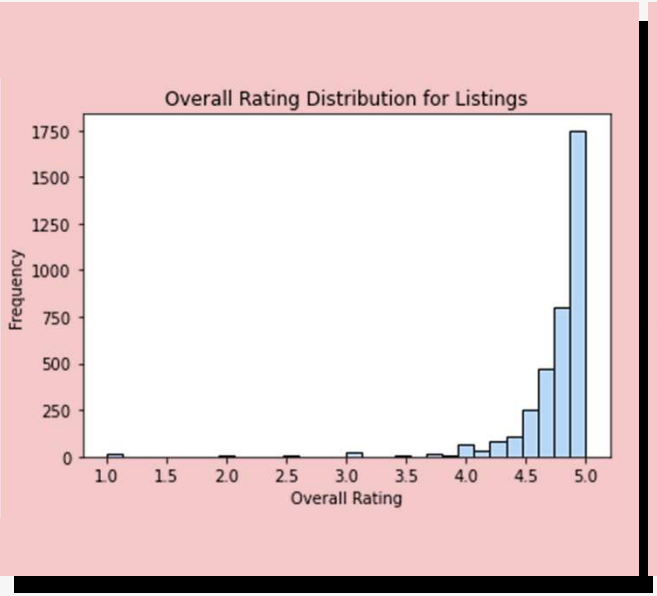
I will create a best predictive model to predict whether an Airbnb listing in DC will be considered popular or not compared to the current listing competition. I will also create a best interpretive model to help hosts understand what features they could improve on their listing to increase popularity. These models will be deployed together an app that hosts can use to make their listings as strong as possible.

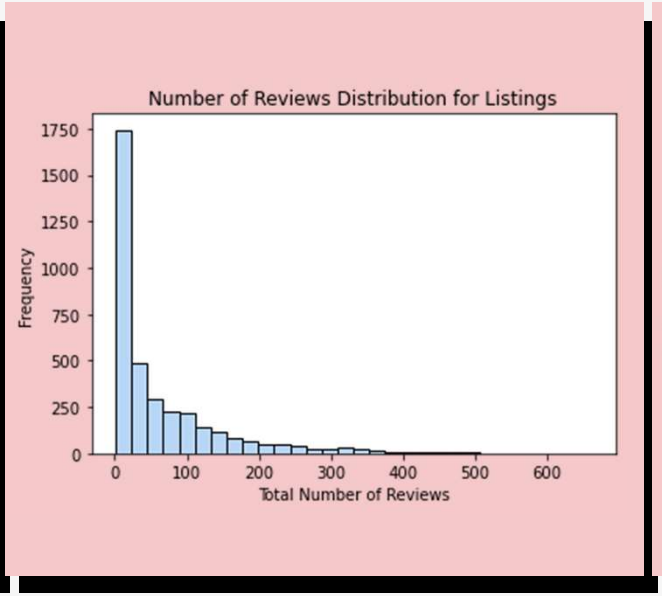


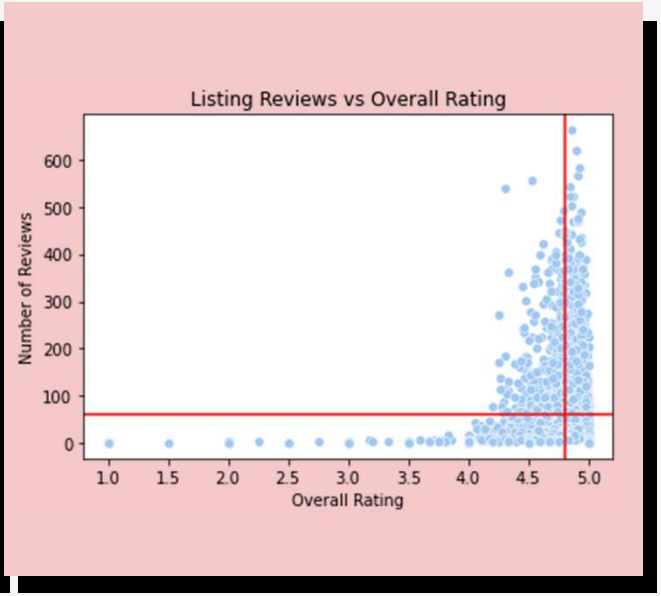
DATA PRE-PROCESSING

DATA COLLECTION	Pulled Washington DC Airbnb listing data from Inside Airbnb and collected DC venue data through a Foursquare API pull. Initial listing data had around 8,000 entries and 75 features. Venue data consisted of counts for 39 DC neighborhoods of about 10 common venues (ex: restaurants, clothing stores, nightlife spots).
FILTERING	Combined listing and venue information. Then filtered down data to contain only recent listings as defined by having a review in the past 12 months.
TARGET VARIABLE	Created the target variable of 'popularity' by looking at distributions of ratings and number of reviews. A popular listing in this dataset is defined as having a 4.8 overall rating or above and 60 or more reviews.
CLEANING	Rating and review features were removed from dataset. Performed other standard cleaning such as replacing nulls as necessary, removing outliers, and transforming data.
TEXT ANALYSIS	Word count and sentiment analysis were performed on listing name, description, neighborhood overview, and about the host and added as features to the dataset. Amenities were also broken out into a top 30 and added whether each listing had the amenity to the dataset as features.

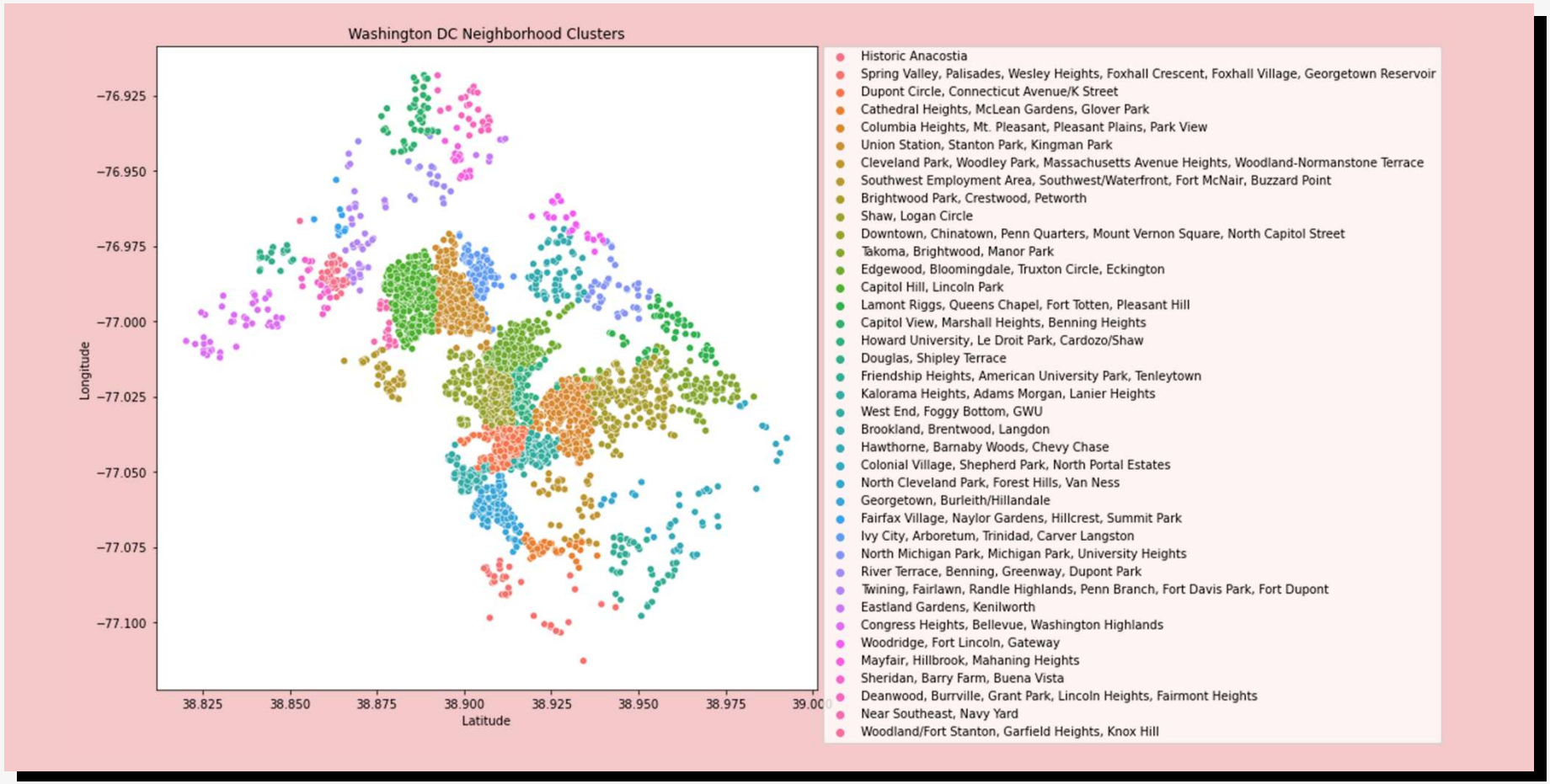
EDA – RATINGS & REVIEWS



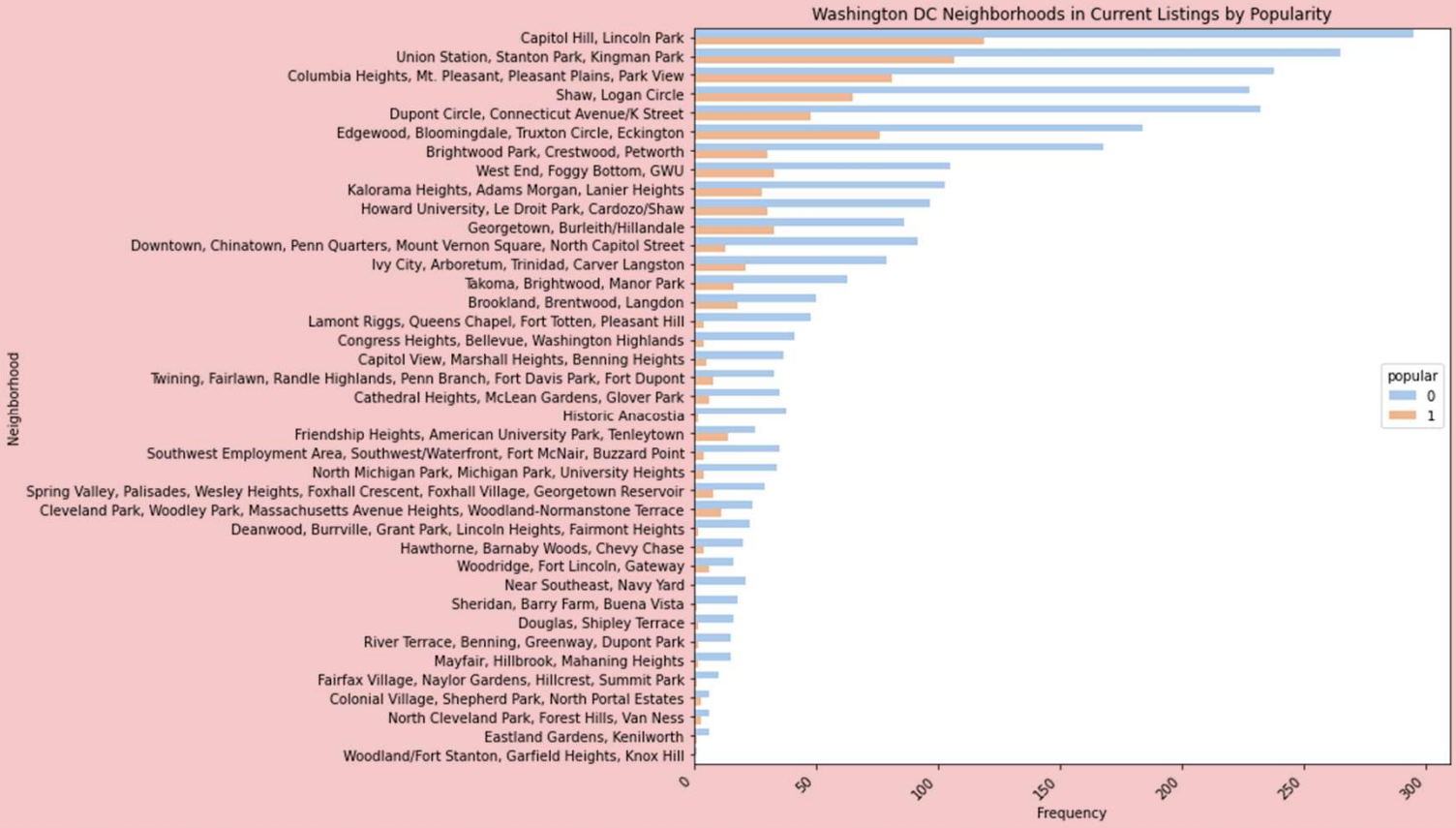




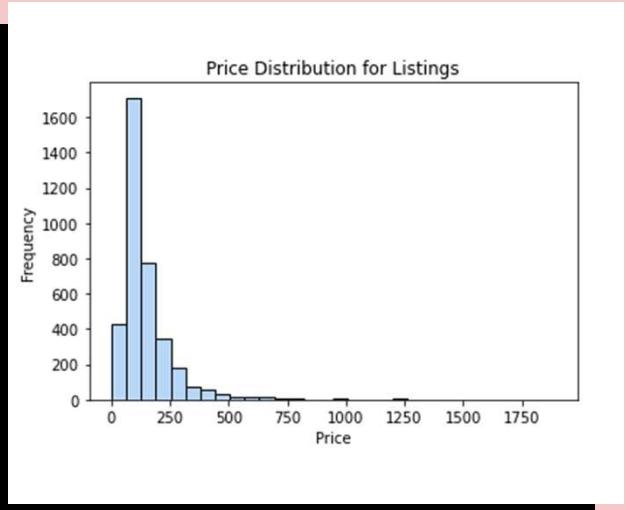
EDA – LISTING NEIGHBORHOODS

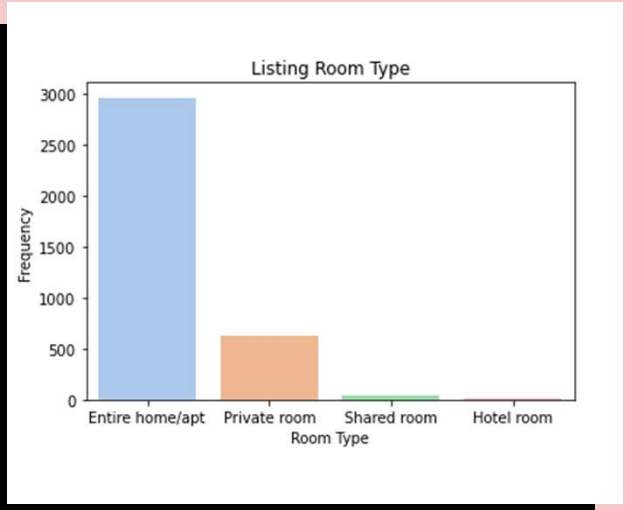


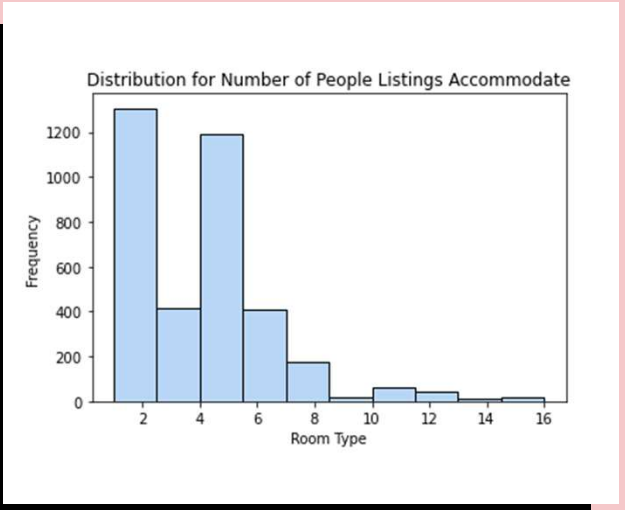
EDA – LISTINGS BY NEIGHBORHOOD & POPULARITY



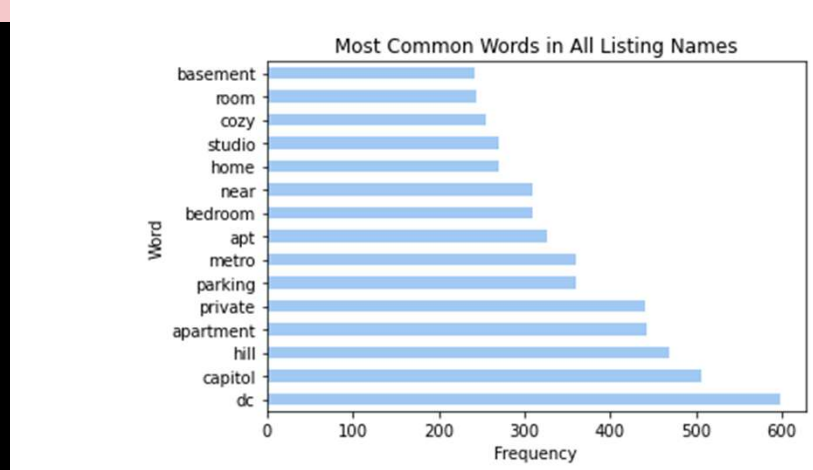
EDA – LISTING CHARACTERISTICS

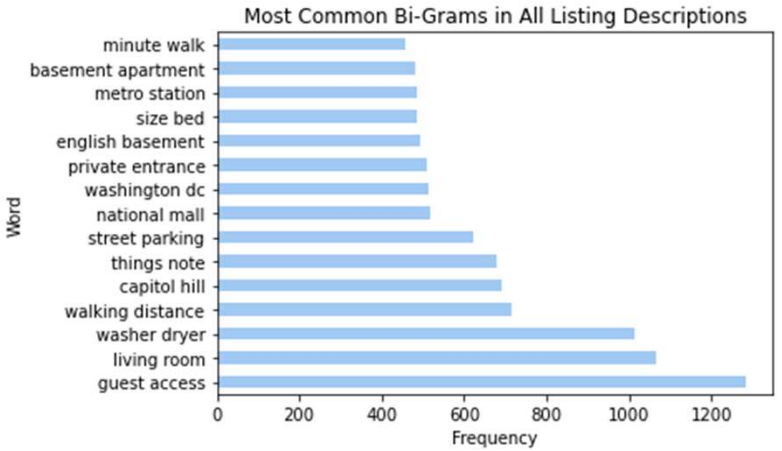


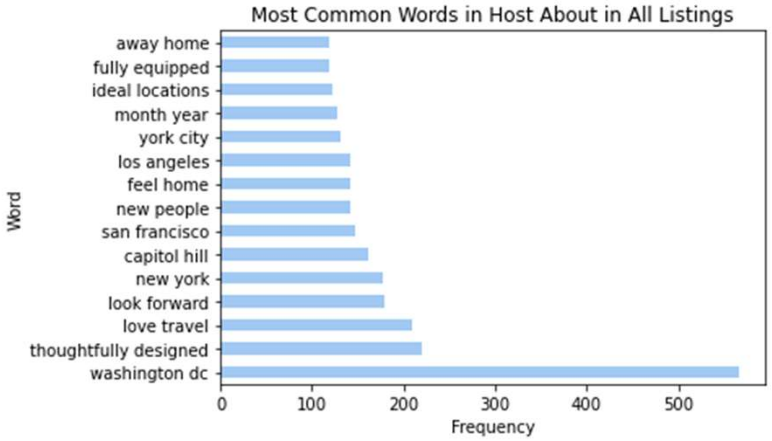


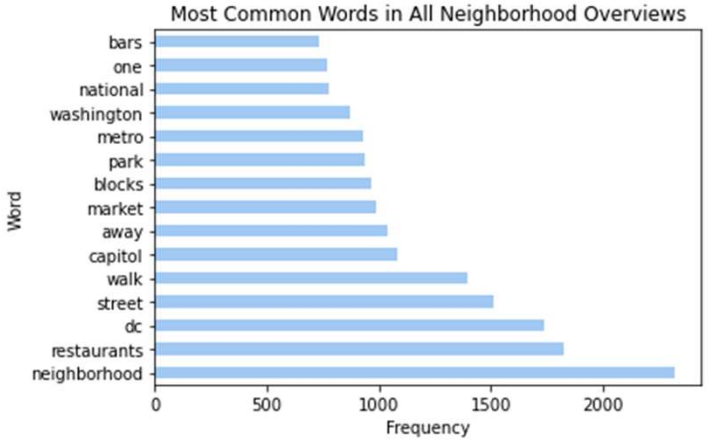


EDA – LISTING TEXT









MODELING PROCESS

I tested lots of models to find a **best predictive model** and a **best interpretive model** for the app. The metrics I used to evaluate both models were the **accuracy and precision scores** in comparison to the baseline.

The **baseline score is 77%,** indicating unbalanced classes in the dataset.

All types of classification models were tested for the best predictive model (KNN, SVC, Logistic Regression, Ensemble Methods, and Neural networks). Once I determined that the Extra Trees Classifier was performing the best in accuracy and precision, I played with feature extraction, hyper tuning parameters, and other sampling methods to improve the model even further.

The best interpretive model needed to be the best Logistic Regression model possible since this is the coefficients are the most interpretable. I also played with feature extraction, hyper tuning parameters, and thresholds to get the best model.

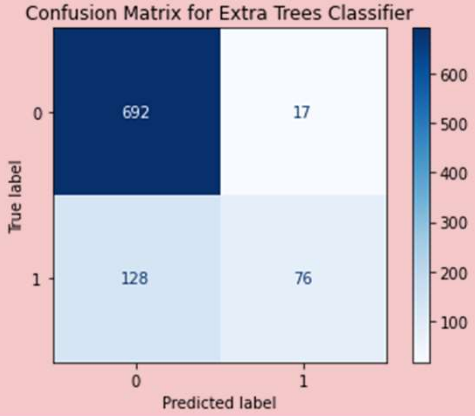
MODELING RESULTS

BEST PREDICTIVE MODEL

- EXTRA TREES CLASSIFIER & SAMPLING MINORITY CLASS

Testing accuracy score of 85%
Testing precision score of 82%

All features included, default model parameter used, oversampled popular class

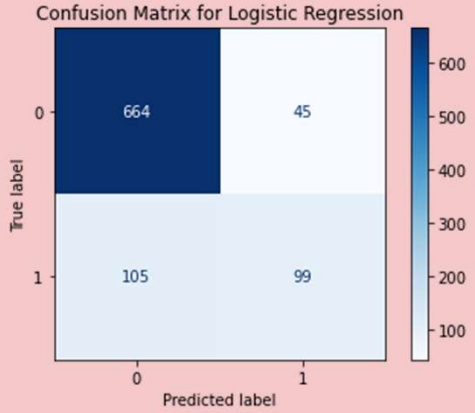


BEST INTERPRETIVE MODEL

- LOGISTIC REGRESSION

Testing accuracy score of 84%
Testing precision score of 69%

All features included and scaled, default parameters and thresholds used



MODEL INTERPRETATION

- The best predictive model is included in the app to make the best prediction on whether a listing is popular.
- The coefficients from the best interpretive model are also incorporated into the app. After taking the inverse log of the coefficients, any coefficient with a value greater than 1 will increase the odds of the listing being popular.
- The app will check if the listing has the positive coefficients as its features and will produce a recommendation to add the feature if it does not already have it.
- 5 largest coefficients: days since first review, being a super host, how many guests the listing accommodates, the neutral sentiment of the listing description, and dishes and silverware as an amenity.
- How to interpret: holding all else constant, if the host is a superhost, the odds that the listing is popular is 2.25 times as large as the odds that the listing is not popular.

CONCLUSION - FINDINGS

MODEL TYPE

The best Extra Trees Classifier model includes all 132 features (lots of binaries). Extra trees does well with sorting through lots of information and deciding what is important and what is noise.

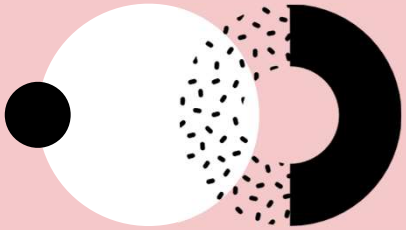
MODEL SCORES

The accuracy and precision scores capped out in the mid 80%s, which is a decent improvement from the baseline score of 77%. It may not be any higher due to the data itself. Differences in popular and not popular listings may not actually be that distinct.

MODEL FEATURES

The most important features in both models are features about the listing (ex: amenities, descriptions) and not DC neighborhoods venues. If the model predicts a listing as popular incorrectly, it is because a listing has most of the important features.

CONCLUSION – NEXT STEPS



There are limitations for these models. The popularity metric is not a perfect because longer stays will have fewer reviews over time.

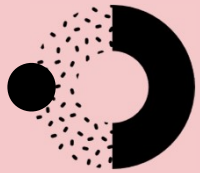
If there were more time, I would recommend 3 things:

- Gather more data for modeling.
- Test models using different review and rating thresholds to calculate the popularity of listings and seeing if these yielded any better results.
- Explore the possibility of predicting popularity rankings and not just a binary class.

STREAMLIT APP

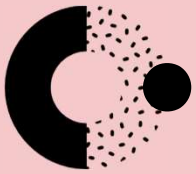
Host can use the DC Hosting Helper as a tool to evaluate their listing. The host can input information about their listing and the app will return a prediction on its popularity as well as a list of features that can be added to increase the odds of it popular.

This will help hosts to have more competitive listings and increase their bookings and profits. Let's take a look!

THANK YOU!

QUESTIONS?




REFERENCES

LISTING DATA	Inside Airbnb - http://insideairbnb.com/index.html
NEIGHBORHOOD VENUES	Foursquare API - https://developer.foursquare.com/
LISTING STATS	SmartAsset - https://smartasset.com/mortgage/where-do-airbnb-hosts-make-the-most-money
LISTING STATS	IPropertyManagement - https://ipropertymanagement.com/research/airbnb-statistics
SLIDES	Canva - https://www.canva.com/