I. Problem

Airbnb's vast user base, seasonality, and broad reach (both across the US and internationally) make discerning signal from noise very difficult. Using sample data from user sessions, I'm going to predict which users are most likely to result in a (1) a booking request, (2) repeat booking user, (3) user who doesn't book in this session, but may in the future, and (4) user who never books.

II. Client and Use Case

The client is Airbnb. With the data that I provide, Airbnb will be able to more accurately predict, trend, and forecast future bookings. Airbnb will be able to customize the user experience based on my data to both maximize revenue from current "high-value" customers while attempting to keep or sell "lower-value" or "no-value" customers. Ideally this would increase the percent of sessions that end with a booking and, perhaps, increase the average revenue per booking.

III. Data Source

The data will be pulled from: http://databits.io/challenges/airbnb-user-pathways-challenge

The data provided is a .txt delimiter separated data file as well as a data dictionary:

- Number of records in data: 7756
- Date span of the data: ['2014-05-05', '2015-04-23']
- Number of unique users in data: 630
- Number of unique sessions in data: 7756
- Percent of sessions with search: 15.9%
- Percent of sessions with sent message: 16.5%
- Percent of sessions with booking request: 1.9%

IV. Approach Outline

???

V. Deliverables

Deliverables for this project will be:

- Code
- Paper
- Slide deck

I. Problem

Why is the English Premier League the "most popular in the world"? James P. Curley and Oliver Roeder explore a few reasons why it might be – and generally can't find a reasonable "data" reason why. It's a fascinating question as we continue to see soccer grow globally, how can various leagues throughout Europe (and other growing soccer regions – i.e. China) capture the world's audience?

II. Client and Use Case

The client is any soccer league in the world besides the Premier League (i.e. German Bundesliga). With the data that I provide, the Bundesliga may be able to promote more popularity of their league around the world. Are there not enough "popular" players in the league? Is the competitive balance stilted, thus generating less interest? Is there not enough commentary on social media surrounding the leagues big games? What about their smaller teams? If the Bundesliga can understand why the Premier League is pulling in billions with their foreign television deals, they will be able to follow a path paved by the English and challenge them for the top.

III. Data Source

The data will be pulled from a variety of sources:

- Soccer data:
  - https://github.com/jalapic/engsoccerdata
  - http://www.jokecamp.com/blog/guide-to-football-and-soccer-data-and-apis/
  - https://github.com/openfootball/
  - http://openfootball.github.io/
  - https://www.washingtonpost.com/news/fancy-stats/wp/2014/09/12/the-fastest-and-slowest-attacking-teams-in-the-premier-league/
- Social media
  - https://dumps.wikimedia.org/other/pagecounts-raw/
  - http://tinyletter.com/data-is-plural/letters/data-is-plural-2015-11-25-edition
  - http://projects.fivethirtyeight.com/reddit-ngram/?keyword=premier%20league.bundesliga.liga&start=20071015&end=20150831&smoothing=10

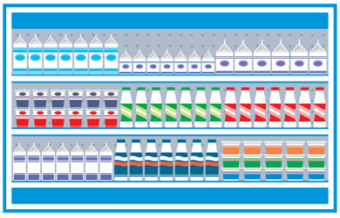
IV. Approach Outline

???

V. Deliverables

Deliverables for this project will be:

- Code
- Paper
- Slide deck



I. Problem

How can we predict repeat purchasers in online retail? With enough user purchase history, it's possible to predict which shoppers will purchase what product and provide them with customized recommendations (everyone is doing it these days). But, what are the ingredients that make a loyal customer? Can we identify that loyal customer prior to the initial purchase? This is the problem posed in the Acquire Valued Shoppers challenge (from Kaggle).

II. Client and Use Case

The client is any online retailer. With the data that I provide, the online retailer will be able to predict, forecast, and promote their loyal and valued shoppers. Knowing who is likely to be a repeat purchaser before their initial purchase is extremely useful. They will be able to customize recommendations, provide more discounts / offers, or introduce other programs designed to increase the number of loyal customers and make sure that their loyal customers are purchasing more from the retailer.

III. Data Source

The data will be pulled from: https://www.kaggle.com/c/acquire-valued-shoppers-challenge/data

The data provided is five separate .csv files:

- transactions.csv - contains transaction history for all customers for a period of at least 1 year prior to their offered incentive
- trainHistory.csv - contains the incentive offered to each customer and information about the behavioral response to the offer
- testHistory.csv - contains the incentive offered to each customer but does not include their response (you are predicting the repeater column for each id in this file)
- offers.csv - contains information about the offers

IV. Approach Outline

???

V. Deliverables

Deliverables for this project will be:

- Code
- Paper
- Slide deck