**Three Ideas for Capstone Project**

1. **Rideshare Revenue Optimization in Chicago**

In Chicago, Transportation Network Providers (rideshare companies) have been required to report all trips to the City of Chicago since November of 2018. In the current data set, each row is an individual trip with 21 attributes and there are currently 45.3 million trips. The attributes for each trip include the following data:

* Trip ID, start time, end time, duration, miles, fare, tip, additional charges, Trip total.
* Pickup census tract, drop off census tract, pick up community area, drop off community area, pick up centroid longitude, pick up centroid latitude, drop off centroid longitude, drop off centroid latitude.

This data set is available at:

<https://data.cityofchicago.org/Transportation/Transportation-Network-Providers-Trips/m6dm-c72p>

A data set which maps the Census Tracts to Community Areas is available at:

<https://data.cityofchicago.org/Facilities-Geographic-Boundaries/Boundaries-Census-Tracts-2010/5jrd-6zik>

From this data set, two models are proposed:

* A machine learning model to predict the revenue for trips from one location to another location at various times of the day. This model would allow an individual driver to identify the most lucrative trips and expected earnings during his or her hours of work so that those trips can be selected from the incoming ride requests at specific times of the day.
* An optimization model for a driver to follow in order to maximize their revenue if they are planning to work for a given number of hours starting at a certain time of the day. This would allow a driver to “think ahead” when selecting which ride request to accept as they will likely wish to accept another ride request in the area where they perform the drop off from the initial ride request (and so on). Essentially, this model would identify ideal sequences of trips that a driver should try to take in order to maximize revenue.

1. **AirBnB Availability and Pricing in Toronto**

An organization called “Inside AirBnb” scrapes AirBnB listing data for various cities around the world to determine the impact that AirBnb has on the local rental market.

For the City of Toronto, listing scrapes have been performed once per month from April 9, 2018 to May 6, 2019 with the exception of June 2018. Two files are created for each of these scrapes:

* A “listing” file with the listing information for an individual rental unit.
* A file showing the availability of a given listing for the 365 days following the scrape.

The data sets are available at:

<http://insideairbnb.com/get-the-data.html>

“listings.csv.gz” and “calendar.csv.gz”

Using the data sets, a machine learning model could be developed to predict availability of rentals as follows:

* The rentals in Toronto could be split up into segments based on type, location, minimum stay, etc. for the first eight scrapes.
* A machine learning model could be applied to each segment over the first eight scrapes to develop predictions for the future availability of units at the remaining four scrape dates based on price.
* Individual rental units could then be compared against predictions and trends for the segment to determine if they are overpriced or underpriced based on their availability. A more optimal price could be established to achieve earlier booking of the unit.

1. **Predicting Peer to Peer Lending**

The Lending Club matches borrowers with investors and publishes a dataset of all loans issued through its platform since 2007. The dataset contains a substantial number of attributes for each loan that can be used to profile the borrower. A dataset of “rejected” loan requests is also published but it contains substantially less attributes.

The datasets are available at:

<https://www.lendingclub.com/info/download-data.action>

Using the data set for loans granted, a model could be developed to predict the profiles of borrowers who are successful and make predictions. The rejection data, which has much less attributes, could be used to gain insight into which factors potentially result in rejection.