**Springboard - Capstone Project**

Goals :

1. Take part in an active Kaggle competition.
2. Learn new concepts
3. Develop a data story
4. Apply machine learning methods
5. Gain data wrangling experience in python

**Project Ideas**

**Zillow Home Value Prediction (Zestimate)**

Reference : <https://www.kaggle.com/c/zillow-prize-1>

“Zestimates” are estimated home values based on 7.5 million statistical and machine learning models that analyze hundreds of data points on each property. And, by continually improving the median margin of error (from 14% at the onset to 5% today), Zillow has since become established as one of the largest, most trusted marketplaces for real estate information in the U.S. and a leading example of impactful machine learning.

When you have an existing model that you think is pretty good (like the Zestimate) but not perfect, modeling errors can be a very powerful way to find areas to improve your existing model. Models that are able to reliably predict error for out of sample sales have identified additional signal in the source data that is not currently exploited.

Home sale prices have a right skewed distribution and are also strongly heteroscedastic, so we need to use a relative error metric instead of an absolute metric to ensure valuation models are not biased towards expensive homes. A relative error metric like the percentage error or log ratio error avoids these problems. While we report Zestimate errors in terms of percentages on Zillow.com because we believe that to be a more intuitive metric for consumers, we do not advocate using percentage error to evaluate models in Zillow Prize, as it may lead to biased models The log error is free of this bias problem and when using the natural logarithm, errors close to 1 approximate percentage errors quite closely.

Initial observations from EDA

1. The goal is to develop an algorithm that makes predictions about the future sale prices of homes.
2. Zillow has modeled the competition to predict the log error instead of Zestimate or Actual Sale Price.
3. Dataset has high rate of missing values
4. There is no definite pattern/correlation identified

The project is interesting especially since the value to be predicted is the log error than the actual zestimate or sale price. This provides opportunity to explore the advantage and concepts behind using log error.

The project offers opportunity to learn regression methods and metrics and to possibly use unsupervised learning methods like clustering to improve results.

**Instacart – Market Basket Analysis**

Reference : <https://www.kaggle.com/c/instacart-market-basket-analysis>

Instacart is challenging the Kaggle community to use anonymized data on customer orders over time to predict which previously purchased products will be in a user’s next order.

Instacart’s data science team plays a big part in providing this delightful shopping experience. Currently they use transactional data to develop models that predict which products a user will buy again, try for the first time, or add to their cart next during a session.

Initial observations from EDA

1. The goal is to develop an algorithm that predicts the products that might be purchased in the next order

The project is interesting since this is my first attempt at Market Basket Analysis and the project presents opportunity to learn the underlying concepts and provides a basis for understanding recommendation systems.