Capstone Project Proposal

Home in the Outback: Predicting Housing Prices in Australia.

Business Understanding

- What problem are you trying to solve, or what question are you trying to answer?
    - How can you determine the true value of a housing property in Australia?
- What industry/realm/domain does this apply to?
    - This applies to the real estate industry.
- What is the motivation behind your project? (Saying you needed to do a capstone project for flatiron is not an appropriate motivation)
    - Australians are having a difficult time finding affordable properties to purchase or rent due to rising real estate prices and underquoting in the industry. Solving this, or at least taking the steps towards it can help mitigate the challenges in buying housing.

Data Understanding

- What data will you collect?
    - A sample of 1000 housing listings for Australian homes.
- Is there a plan for how to get the data (API request, direct download, etc.)?
    - Direct download from Kaggle.
- What are the features you'll be using in your model?
    - Features: building size, land size, bedrooms, bathrooms, parking spaces, property type, product depth.
    - Target: housing price

Data Preparation

- What kind of preprocessing steps do you foresee (encoding, matrix transformations, etc.)?
    - Property type and product depth need to be encoded, while the rest of the features require scaling.
- What are some of the cleaning/pre-processing challenges for this data?
    - Building and land size are strings that need to be converted into numeric data. It also needs unit conversion into meters squared.
    - Pricing is also in a string format, and needs to be converted into numeric data.
    - Missing data needs to be imputed, whether through a simple median or a median based on a categorical variable (such as property type)

Modeling

- What modeling techniques are most appropriate for your problem?

    - Multiple linear regression is a good statistical model to begin with, along with decision trees and gradient boosting models.
- What is your target variable? (remember - we require that you answer/solve a supervised problem for the capstone, thus you will need a target)
    - Housing listing price.
- Is this a regression or classification problem?
    - Regression.

Evaluation

- What metrics will you use to determine success (MAE, RMSE, Accuracy, Precision etc.)?
    - RMSE; in this context, it punishes models that have larger price errors more than those that don't. This helps us optimize models.

Tools/Methodologies

- What modeling algorithms are you planning to use (i.e., decision trees, random forests, etc.)?
    - Median (baseline)
    - Multiple linear regression
    - Decision tree
    - Gradient boosting