Capstone Project Proposal

1. What is the problem you want to solve?

I want to predict the house prices at King County, Seattle, USA, based on multiple features using regression analysis.

2. Who is your client and why do they care about this problem? In other words, what will your client do or decide based on your analysis that they would not have otherwise?

My client are both people who want to buy houses, and real estate agents who want to sell houses at King County. They will have general knowledge about house prices according to the features of houses.

3. What data are you going to use for this? How will you acquire this data?

I will use “House Sales in King County, USA” data from Kaggle. This dataset contains house sale prices for King County, which includes Seattle. It includes homes sold between May 2014 and May 2015. It has 19 house features plus the price and the id columns, along with 21613 observations. It's a great dataset for evaluating simple regression models.

Data Source: https://www.kaggle.com/harlfoxem/housesalesprediction

4. In brief, outline your approach to solving this problem (knowing that this might change later).

I will approach this machine learning project by following the steps below:

a. Create a repository in Github
b. Get the data from Kaggle.
c. Analyze the data to determine the data quality using Jupyter Notebook.
d. Prepare the data and make EDA:

(1) Clean data (remove duplicates, deal with missing values, correct errors, normalization, data type conversions, etc.)

(2) Manipulate data

(3) Visualize data to help detect relevant relationships between variables and perform exploratory analysis.

e. Feature engineering and selection
f. Split the data as training and test data
g. Choose a machine learning algorithm
h. Train the model
i. Evaluate the model
j. Parameter Tuning
k. Make predictions.
l. Prepare a report

5. What are your deliverables? Typically, this would include code, along with a paper and/or a slide deck.

My deliverables will be a paper, a powerpoint presentation summarizing my paper, and the code associated with my project.