Capstone Project Proposal Template

Notes:

- This should take no more than one hour to complete – the clearer you are about the business problem you're working to solve with your ML-driven solution, the easier your proposal will be to complete
- This will be uploaded to your repo, which will be a part of your final submission
- Due date for submission is June 23, 2023.

Instructions:

1. Download this document as a Word Doc
2. Answer each question using a few sentences, at most
3. Save your completed proposal as a PDF
    a. File should be saved in the following format:
    b. GROUP NUMBER_DATE OF SUBMISSION (example: GROUP 8_MAY 2)
4. Create a project GitHub repo (if you have yet to do so)
5. Add your instructor as a collaborator to your project repo
6. Add your Deloitte mentor and VT Advisor (when assigned) as a collaborator
7. Push your proposal PDF (created in Step 3) up to your repo
8. Copy the URL corresponding to the location of the PDF in your repo
9. Submit the copied URL using this link

Housing ROI Analysis by Zip Code

Business Understanding

- What problem are you trying to solve, or what question are you trying to answer?
    - The housing market tends to be more stable and profitable in certain areas than it is in others. We aim to determine which areas (by zip code) tend to yield the highest return on investment. What factors contribute the most/least to economic successes or failures in the housing market.
- What industry/realm/domain does this apply to?
    - This is mainly applying to the housing/financial(investments) industries
- What is the motivation behind your project? (Saying you needed to do a capstone project for flatiron is not an appropriate motivation)
    - The motivation behind this project came from an interest in the housing market and factors affect it the most. What factors make a house/property a good or bad investment.

Data Understanding

- What data will you collect?
    - Zillow Home Value Index (ZHVI): Zillow provides a comprehensive dataset with historical home values, rental prices, and other related metrics at the zip code level. This dataset can be used to analyze real estate price appreciation and rental income potential.
    - U.S. Census Bureau: The Census Bureau offers various datasets related to demographics, population growth, income levels, education, and more. These datasets can provide valuable insights into the socio-economic factors that influence real estate investment potential.
    - Bureau of Labor Statistics (BLS): BLS provides economic indicators such as employment rates, wage data, and industry-specific statistics. These datasets can help assess the economic stability and growth potential of different zip codes.
    - Federal Housing Finance Agency (FHFA): FHFA offers datasets related to housing market indicators, including historical home price indices, mortgage rates, and loan performance. These datasets can provide a broader perspective on real estate market trends and help identify potential high ROI zip codes.
    - OpenStreetMap (OSM): OSM is an open-source mapping platform that provides geospatial data, including amenities such as schools, hospitals, public transportation, and other points of interest. Integrating OSM data with real estate datasets can help analyze the impact of proximity to amenities on ROI.
- Is there a plan for how to get the data (API request, direct download, etc.)?
    - All the data sets listed above should be available by direct download.
- Are the features that will be used described clearly?
    - They are described clearly within the data set as to how they contribute to the overall set.

Data Preparation

- What kind of preprocessing steps do you foresee (encoding, matrix transformations, etc.
- Data cleaning, Feature Selection, Encoding variables, Scaling and Normalization.

What are some of the cleaning/pre-processing challenges for this data?

- The biggest challenge I foresee facing is cleaning the data and being able to incorporate them with one another.

Modeling

- What modeling techniques are most appropriate for your problem?
    - Time Series Analysis and/or classification 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)
    - Real-estate price appreciation
- Is this a regression or classification problem?
    - It is a classification problem because the goal is to classify zip codes into different categories based on their ROI potential,

Evaluation

- What metrics will you use to determine success (MAE, RMSE, etc.)?
    - Accuracy, precision and recall, and F1 score.

Tools/Methodologies

- What modeling algorithms are you planning to use (i.e., decision trees, random forests, etc.)?\
    - Logistic Regression or Random Forest