John Nguyen’s Deloitte AI Academy Apprenticeship Capstone Project Proposal Template 2023

Project Name: Using 2021-2022 Fortune 100 data for revenue and expenses to predict 2023 Tech vs Non-Tech industry health

Project Statement:

By acquiring the revenue and expenses for both Tech and Non-Tech sectors, I will predict the health (margin) of a sector for 2023. This project matters because consulting firms need to adjust the way they approach their sales/pursuit efforts in the tech and non-tech sectors during a volatile market.

Business Understanding:

- What problem are you trying to solve, or what question are you trying to answer?
  - Which sector (Tech & Non-Tech) has better ‘health’? (Health = Profit Margin)
  - How can a firm like Deloitte focus their efforts to sell the most work during a volatile market?
- What industry/realm/domain does this apply to?
  - Tech & Non-Tech Markets
  - Consulting (via effective firm pursuit/sales efforts)
- What is the motivation behind your project?
  - As we’re navigating a volatile market, I want to do a data science project while in AI Academy that can help my firm’s performance in the marketplace & increase our project sales revenue.

Data Understanding:

- What data will you collect?

  Fortune 1000 companies in 2021 and 2022 | Kaggle

  - 2021 Revenues
  - 2022 Revenues
  - 2021 Expenses (By using Profit-Revenue to find Expenses)
  - 2022 Expenses (By using Profit-Revenue to find Expenses)
- Is there a plan for how to get the data (API request, direct download, etc.)?
  - Direct downloading excel from Kaggle

- What are the features you’ll be using in your model?
    - 2021 Revenues
    - 2022 Revenues
    - 2021 Expenses (By using Profit-Revenue to find Expenses)
    - 2022 Expenses (By using Profit-Revenue to find Expenses)
    - Predicting 2023 Profit Margin (-> answers our question about ‘health’)
    - Predicting 2023 Revenue (by using 2021 & 2022 Revenue)
    - Predicting 2023 Expenses (by using 2021 & 2022 Expenses)

Data Preparation:

- What kind of preprocessing steps do you foresee (encoding, matrix transformations, etc.)?
    - Finding the independent variable “Expenses” by using Profit-Revenue to find 2021 & 2022 Expenses
    - Shaving off 900 records from the “Fortune 1000” data set to get “Fortune 100”
    - Manually classifying ‘Fortune 100’ companies to create the classifier “Tech” or “Non-Tech
- What are some of the cleaning/pre-processing challenges for this data?
    - Because we’re pulling “2021” and “2022” datasets from the same Kaggle link, there should be less room for error
    - All the above: (Finding Expenses, Reducing the sample size to 100, and manually classifying “Tech” or “Non-Tech”)

Modeling:

- What modeling techniques are most appropriate for your problem?
    - Linear Regression
- What is your target variable? (Remember - we require that you answer/solve a supervised problem for the capstone, thus you will need a target)
    - 2023 Profit Margin
- Is this a regression or classification problem?
    - Regression

Evaluation:

- What metrics will you use to determine success (MAE, RMSE, etc.)?
    - MAE to reduce error
    - Success = Accuracy/Precision of finding the 2023 (Revenue, Expenses, Margin)

Tools/Methodologies:

- What modeling algorithms are you planning to use (i.e., decision trees, random forests, etc.)?
    - Scikit Learn: Linear Interpolation Graph
    - Time Series Forecasting: Multivariate Forecast
    - Regression Analysis: Multiple Linear Regression