Part 2: Problem Statement, EDA and Dataset (My answers are in dark blue/italicised font)

Overview

In this section you will update us on your project, including the project you have chosen, your problem statement, an extensive outline of EDA and modeling to date, the goal of your predictive model, and the data you will use to explore that model.

Your data must be fully in hand by this point OR you must have a solid, achievable plan to do so that has been communicated to your local instructor.

Requirements

We expect a formatted and complete Jupyter Notebook by end of class on June 20th, 2018 which accomplishes the following:

- Identifies which of the three proposals you outlined in your lightning talk you have chosen
- Articulates the main goal of your project (your problem statement)
- Outlines your proposed methods and models
- Defines the risks & assumptions of your data
- Revises initial goals & success criteria, as needed
- Documents your data source
- Performs & summarizes preliminary EDA of your data

Time Series stock prediction of consumer goods stocks; in specific retail/fashion/luxury goods. Further exploring the effect of the recent 'musical chairs' of creative directors and how it affects their houses stock price.

Formulating your Problem Statement

Your problem statement should the guiding principle for your project. You can think about this as a "SMART" goal.

- Specific:
    - What precisely do you plan to do?
        - *First to build a stock predictor for consumer goods companies, retail groups and other luxury/fashion companies.*
        - *Then to do a snapshot of specific fashion comglomerates that are hiring these high priced and world renowned creative diectors in order to highlight:*
            - *the financial impact of their entrance and exits of the brand*
        - *Lastely illustrate a picture of how the $2.4 trillion retail sector is divided up based on an array of financial and social factors. The goal of this is to give more color to the industry as a whole and all the brands trying to grab their slice of the pie. Providing a more in-tune and micro perspective on an sector largely dominated by congolmerates.*
    - What type of model will you need to develop?
        - *Time Series predictive modeling*
        - *Clustering model of the retail industry and in specific popular popular and cult fashion brands*
            - *Features will include*
                - *financial data not included previosuly in the stock prediction time series model like:*

            - *sales*
            - *revenues*
            - *instagram followers*
            - *where the brands are sold*
                - *The barneys factor*
            - *Certain factors based off references or endoresement from influencers.*
        - *Home-made "Network Analysis" detailing the flight of the designers/creative directors given that there are so few competent and largely successfull ones. Where they come from, whats their background, work history and such:*
            - *Did they prove to be add financial value for the brand while they were there*
            - *How was there effects on instagram follower count*
- Measurable:
    - What metrics will you be using to assess performance?
        - *In order to measure the time series stock prediction:*
            - *Confidence intervals, standard deviation/variance, quantiles, sensitivity table/analysis, MASE (Mean absolute scaled error), MAPE (Mean absolute percentage error)*
    - MSE? Accuracy? Precision? AUC?
- Achievable:
    - Is your project appropriately scoped?
        - *There is a very clear scalable aspect to my project meaning there are a lot of extra components I can add to the core of the project. That core being a time series stock predictor for consumer goods companies and retail brands.*
    - Is it too aggressive? Too easy?
        - *It may be easy in terms of pulling the data(hitting an API), but overall I believe its falls fair between too agressive and too easy given the fact I broke it into many pieces hedging the intensity based off how much I can get done (Note: If your project is too big, break it up into smaller pieces. Sometimes a good project is the simply one part of a larger, longer-term agenda.)*
- Relevant:
    - Does anyone care about this?
        - *Time series prediction is a very valuable tool which is used heavily in forecasting and projections for many industries. A stock predictor for one of the hottest sectors of the econonomy over the last few*

years is bound to peak someones interest. Also for someone who has an interest in fashion it offers a cool perpsective into how the industry operates.

- Why should people be interested in your results?
  - *This question is mainly answered above but a very interesting aspect to this project is the fact that im trying to quantify a someone difficult indicator to quantify: artistic genius. A goal of this project is to highlight the finanical and social impact on a fashion house/brand from the entry and exit of their extremely sought after and expensive creative directors.*
- What value will the completion of your project be adding?
  - *Answered in the the two above questions*

- Time-bound
  - What's your deadline?
    - *Only deadline I have for now is the project deadline, most of this stuff is all work I can get done. Especially the time series stock predictor.*

- BAD: I will model emergency room visits.
- GOOD: I will build a regression model to predict the number of daily emergency room visits for St. Someone's Hospital. Model performance will be guided by RMSE, and the model should at least improve upon baseline by 10%. Baseline is defined as the monthly average of visits over the last 10 years.

- BAD: I will investigate the aftermarket pricing of sneakers.
- GOOD: Specific image and text features of sports sneakers are predictive of determinding wether they will sell for more or less in the aftermarket. The guiding metric will be area under the ROC curve.

- BAD: I will explore the link between obesity and blood pressure.
- GOOD: I will quantify the association between obesity and blood pressure through regression modeling.
- BETTER: As obesity increases, how does blood pressure change?

- BAD: I will predict that sources of news are liberal or conservative.

- GOOD: I will look at text features to undersatnd how news can be classified as liberal or conservative.
- BETTER: Specific text feature frequencies can determine the broader category of news sources using classification. I will describe what makes each class charactitaristically unique, describe what is both certain and uncertain using precision and recall as success metrics. Then I will conclude with a description of "why" my model describes potential to predict these two categories.

Data Guidelines

What should you thinking about and looking for as you collect your capstone data?

- Source and format your data
    - Have a way to save data locally (e.g., SQL or CSV), especially if scraping from the web or collecting from an API.
        - *Using the Quandl API I pulled stock pricing history of the following:*
            - *PVH - NYSE*
            - *KORS - NYSE*
            - *RALPH LAUREN - NYSE*
            - *COACH - NYSE*
            - *LVMH - Euronext*
            - *DIOR - Euronext*
            - *BURBERRY - LSE*
            - *KERING - Euronext*
            - *HERMES - RMS*
        - *Need to figure out how to pull the following:*
            - *PRADA - Hong Kong*
            - *MONCLER - Italy/Milan*
            - *ADIDAS - Frankfurt*
        - *Need to put together a larger list of larger retail groups to pull from the API*
        - *Questions regarding the state of my data:*
            - *State of my data is fairly straightforward I just need to pick which stocks I want and then hit the API*
            - *Already have a function in line to combine all of those stocks into a single API*
    - Create a data dictionary to accompany your data.

        - *The data dictionary only contains straightforwad trading data about the openening and closing prices*
    - *As for the data which is specific to the creative directors and there accompanying houses:*
        - *Try to pull more financial data on the house which are hiring them to better reflect their impact on revenue and following (social media)*
        - *Using google trends and historical timeline of instagram followers*
        - *Could be room for an extra time series predicting instagram followers, mentions in general and future revenues once these creatives enter or exit.*
            - *"Hey we could see a xx% increase in your instagram followers if you hire someone with this esteem and popularity*
        - *Then using linkedin, wikipedia, and other research outlets I follow to get more info of the designers historical/work timeline.*
        - *Perform initial cleaning and munging.*
    - Organize your data relevant to your project goals.
        - *As stated above, hitting the API then getting them into a single csv*
        - *For the smaller research aspects concering the designers and such, using excel to organize those details.*
    - Write functions to automatically clean and munge data as necessary.
    - Take copious notes, for both others and yourself, describing your assumptions and approach.

EDA Guidelines

Think about the following as you perform your initial EDA.

- Identify the data types you are working with.
- Examine the distributions of your data, numerically and/or visually.
    - *Have not gotten far enough within the process to do this*
- Identify outliers.
    - *Have not gotten far enough within the process to do this*
- Identify missing data and look for patterns of missing data.
    - *Have not gotten far enough within the proces to do this.*
- Describe how your EDA will inform your modeling decisions and process.
    - *Getting my historical stock pricing data into into a readable CSV in order for the time series to be able to read it.*

- *My EDA on the clustering will be picking certain features I want and building that dataframe for the clustering model to take it.*

Necessary Deliverables / Submission

Materials must be presented in a Jupyter Notebook stored within a repository on your personal (*not* GA) GitHub. Please submit a link to this repository by the due date (submission link).

BONUS

- Create roadmap of your project with milestones.
- Write a blog post on what you learned from your EDA.

Version 1, my rough draft:

Predicting future stock prices for consumer goods/retail/luxury/fashion companies(brands)

a) Pull historical stock pricing from Quandl API (going back to Feb 2014)
   i) PVH - NYSE
   ii) KORS - NYSE
   iii) RALPH LAUREN - NYSE
   iv) COACH - NYSE
   v) LVMH - Euronext
   vi) DIOR - Euronext
   vii) BURBERRY - LSE
   viii) KERING - Euronext
   ix) HERMES - RMS
   x) PRADA - Hong Kong
   xi) MONCLER - Italy/Milan
   xii) ADIDAS - Frankfurt
   xiii) ***** need more to reflect the entire industry

2) Highlight impact of creative directors

a) Timeline of entrants and exits
b) Which creative heads are coming and going
c) Where are they coming from and who are they going to
d) Dates of their entry and exit, so you can use that timeframe as a snapshot to reflect their impact on stock price
e) Houses who are hiring these directors are all controlled by two or three conglomerates
    i) Try to pull more financial data on the house which are hiring them to better reflect their impact on revenue and following (social media)
    ii) Using google trends and historical timeline of instagram followers
    iii) Could be room for an extra time series predicting instagram followers, mentions in general and future revenues once these creatives enter or exit

3) Network Analysis on Designers themselves

a) Mapping the flight of these rarified creatives
    i) Virgil Abloh
    ii) Raf Simmons
    iii) Kim Jones
    iv) Christopher Bailey (maybe)
    v) Ricardo Tischi
    vi) Hiedi Slimane (maybe)
b) Origins aka where they are from
c) Where they went to school
d) Financial success of when they were at each brand vs the years before and after they entered and exited respectively.
    i) Could also do this for social media popularity
        (1) Likes and followers over time

4) Clustering model of the retail industry and in specific popular and cult fashion brands.

a) Features will include
    i) financial data not included previosuly in the stock prediction time series model like:
        (1) sales
        (2) revenues
        (3) instagram followers

(4) where the brands are sold
    (a) The barneys factor
(5) ASAP Rocky factor
(6) Kanye West factor (maybe)
(7) Luka Sabatt factor
(8) Virgil Abloh factor

b) The goal of this is to give more color to the industry as a whole and all the brands trying to grab their slice of the pie. Providing a more in-tune and micro perspective on an sector largely dominated by congolmerates.