FeatureEngineering

November 26, 2019

In [1]:
```
import google
import numpy as np
import pandas as pd
pd.options.display.max_columns = None
# from google.colab import files
# from google.colab import drive
# drive.mount('/gdrive')
# uploaded = files.upload()
```

In [2]:
```
df = pd.read_csv("FinalDraftCleanedMergedCheckPointData.csv")
df.drop(columns = "Unnamed: 0", inplace = True)
```

Feature Engineering

This portion will mainly focus on preparing the features to be prepared in a way that will be meaningful when building our regression model. We will focus on the following:

1. One-Hot Encoding all categorical variables of interest
2. Engineering new features from exiting features
3. Checking the quality of our Features

Once we complete the feature engineering we will be ready for running the model

1 One-Hot Encoding

We will strip the categoricals and get the variables of interest here

In [3]:
```
df["HBR Ranked"] = (df["HBR CEO Rank"] > 0).astype(int)
df["GD Ranked"] = (df["GD CEO Rank"] > 0).astype(int)
df["GD & HBR Ranked"] = df["GD Ranked"] * df["HBR Ranked"]
```

In [4]:
```
#5 categories:
# < 250 million
# 250 ~ 1000 million
# 1001 ~ 5000 million
# 5001 ~ 50000 million
# 50 Billion+
```

```
df["MarketSize: < 250 million"] = (df['Market Value (M)'] <= 250).astype(int)
df["MarketSize: 250 ~ 1,000 million"] = (df['Market Value (M)'] > 250).astype(int) * (df
df["MarketSize: 1,001 ~ 5,000 million"] = (df['Market Value (M)'] > 1000).astype(int) *
df["MarketSize: 5,001 ~ 50,000 million"] = (df['Market Value (M)'] > 5000).astype(int) *
df["MarketSize: 50 Billion +"] = (df['Market Value (M)'] > 50000).astype(int)
```

```
In [5]: #Breaking it up to the sectors that will be one-hot encoded
        #Light cleaning during the process
        import re
        sectors = pd.get_dummies(df["Sector"])
        bad_names = sectors.columns.values
        good = []
        for name in bad_names:
            good.append(str.join("&",name.split("&amp;")))
        sectors.rename(columns = {bad_names[i]:good[i] for i in range(len(good))}, inplace=True)
        df = df.join(sectors)
```

2 Engineering New Features

We will mainly try to group each data into: > 1.geographical regions > 2.geographical division

Regions and Divisions are determined by U.S. Census Bureau

2.1 Formatting

We have to format the city from *HQ_Location*. We do this by extracting the current formating using regex and string operations and then mapping it to the correct full state name.

```
In [6]: #Converting Improper state label format into standard full state names
        strip_state = lambda city: city.split(", ")[1] if str(city) != "nan" else np.nan
        pd.get_dummies(df["HQ Location"].apply(strip_state)).columns.values

Out[6]: array(['Ala.', 'Ariz.', 'Ark.', 'Calif.', 'Colo.', 'Conn.', 'D.C.',
               'Del.', 'Fla.', 'Ga.', 'Hawaii', 'Idaho', 'Ill.', 'Ind.', 'Iowa',
               'Kans.', 'Ky.', 'La.', 'Maine', 'Mass.', 'Md.', 'Mich.', 'Minn.',
               'Miss.', 'Mo.', 'N.C.', 'N.D.', 'N.H.', 'N.J.', 'N.Y.', 'Neb.',
               'Nev.', 'Ohio', 'Okla.', 'Ore.', 'Pa.', 'Puerto Rico', 'R.I.',
               'S.C.', 'Tenn.', 'Texas', 'Utah', 'Va.', 'Wash.', 'Wis.'],
              dtype=object)
```

```
In [7]: #Map of the state names
        standard =  {
                    'Ala.':"Alabama",
                    'Ariz.':"Arizona",
                    'Ark.':"Arkansas",
                    'Calif.' : "California",
                    'Colo.':"Colorado",
```

```
            'Conn.':"Connecticut",
            'D.C.': "District of Columbia",
            'Del.':"Delaware",
            'Fla.':"Florida",
            'Ga.' : "Georgia",
            'Hawaii':"Hawaii",
            'Idaho' :"Idaho",
            'Ill.' : "Illinois",
            'Ind.' : "Indiana",
            'Iowa' : "Iowa",
            'Kans.':"Kansas",
            'Ky.' : "Kentucky",
            'La.' : "Louisiana",
            'Maine' : "Maine",
            'Mass.' : "Massachusetts",
            'Md.' : "Maryland",
            'Mich.' : "Michigan",
            'Minn.' : "Minnesota",
            'Miss.' : "Mississippi",
            'Mo.' : "Missouri",
            'N.C.' : "North Carolina",
            'N.D.' : "North Dakota",
            'N.H.' : "New Hampshire",
            'N.J.' : "New Jersey",
            'N.Y.' : "New York",
            'Neb.' : "Nebraska",
            'Nev.' : "Nevada",
            'Ohio' : "Ohio",
            'Okla.' : "Oklahoma",
            'Ore.' : "Oregon",
            'Pa.' : "Pennsylvania",
            'Puerto Rico' : "Puerto Rico",
            'R.I.' : "Rhode Island",
            'S.C.' : "South Carolina",
            'Tenn.' : "Tennessee",
            'Texas' : "Texas",
            'Utah' : "Utah",
            'Va.' : "Virginia",
            'Wash.' : "Washington",
            'Wis.' : "Wisconsin"
}
#Mapping the incorrect name
standard_state = lambda city: standard[city] if str(city) != "nan" else np.nan
df["State"] = df["HQ Location"].apply(strip_state).apply(standard_state)
```

2.2 U.S. Census Region/Division

We load in U.S. Official Census Table for regions and divsions mapping. We then use the table to create maps to map our formatted state names to the respective regions/divisions.

```
In [9]: #Using U.S. Census region/division grouping table
        uploaded = files.upload()
```

2.2.1 Region Mapping

```
In [10]: #Identifying the region information and converting using the the map
         regions = pd.read_csv("us census bureau regions and divisions.csv")
         region_map = {regions["State"].values[i] : regions["Region"].values[i] for i in range(5
         region_map["nan"] = np.nan
         region_map["Puerto Rico"] = np.nan
         df["Region"] = df["State"].map(region_map)

         #One-Hot Encoding Region Info
         df = df.join(pd.get_dummies(df["Region"]))
```

2.2.2 Division Mapping

```
In [11]: #Identifying the division information and converting using the the map
         div_map = {regions["State"].values[i] : regions["Division"].values[i] for i in range(51
         div_map["nan"] = np.nan
         div_map["Puerto Rico"] = np.nan
         df["Division"] = df["State"].map(div_map)

         #One-Hot Encoding Division Info
         df = df.join(pd.get_dummies(df["Division"]))
```

3 Feature Quality Check

We want to make sure that the feature we choose are numeric now and also have explanatory power. We use our domain knowledge and decide whether or not each feature is removeable or not

```
In [12]: features_of_interest = ['Company', 'Market Value (M)', 'rank_change1000', 'employees',
                'CEO', 'Sector',
                'HQ Location', 'Employees',
                'Revenues ($M)', 'Revenues ($M)Growth', 'Profits ($M)',
                'Profits ($M)Growth', 'Assets ($M)', 'Assets ($M)Growth',
                'Total Stockholder Equity ($M)', 'Total Stockholder Equity ($M)Growth',
                'Profit as % of Revenues', 'Profits as % of Assets',
                'Profits as % of Stockholder Equity', 'Earnings Per Share ($)',
                'Total Return to Investors (5 year, annualized)',
                'Total Return to Investors (10 year, annualized)', 'Market Cap (M)',
                'No_Directors', 'Median_age', 'Board_Independance',
```

```
          'Median_Tenure', 'Median_pay', 'women_on_board',
          'GD_Approval',
          'HBR Ranked',
          'GD Ranked', 'GD & HBR Ranked', 'MarketSize: < 250 million',
          'MarketSize: 250 ~ 1,000 million', 'MarketSize: 1,001 ~ 5,000 million',
          'MarketSize: 5,001 ~ 50,000 million', 'MarketSize: 50 Billion +',
          'Aerospace & Defense', 'Apparel', 'Business Services', 'Chemicals',
          'Energy', 'Engineering & Construction', 'Financials',
          'Food & Drug Stores', 'Food, Beverages & Tobacco', 'Health Care',
          'Hotels, Restaurants & Leisure', 'Household Products', 'Industrials',
          'Materials', 'Media', 'Motor Vehicles & Parts', 'Retailing',
          'Technology', 'Telecommunications', 'Transportation', 'Wholesalers',
          'State', 'Region', 'Division', 'Midwest', 'Northeast', 'South', 'West',
          'East North Central', 'East South Central', 'Middle Atlantic',
          'Mountain', 'New England', 'Pacific', 'South Atlantic',
          'West North Central', 'West South Central']
```

In [15]:
```
df[features_of_interest].head().fillna(0).to_csv('dataproject_interest.csv')
df[features_of_interest].head().fillna(0)
```

Out[15]:
```
             Company  Market Value (M)  rank_change1000  employees  \
0            walmart          279880.3              0.0  2200000.0
1         exxonmobil          342172.0              0.0    71000.0
2              apple          895667.4              1.0   132000.0
3  berkshirehathaway          493870.3             -1.0   389000.0
4          amazoncom          874709.5              3.0   647500.0

                 CEO      Sector        HQ Location  Employees  \
0  C. Douglas McMillon   Retailing  Bentonville, Ark.  2200000.0
1      Darren W. Woods      Energy      Irving, Texas    71000.0
2      Timothy D. Cook  Technology  Cupertino, Calif.   132000.0
3     Warren E. Buffett  Financials         Omaha, Neb.   389000.0
4      Jeffrey P. Bezos   Retailing     Seattle, Wash.   647500.0

   Revenues ($M)  Revenues ($M)Growth  Profits ($M)  Profits ($M)Growth  \
0       514405.0                  2.8        6670.0               -32.4
1       290212.0                 18.8       20840.0                 5.7
2       265595.0                 15.9       59531.0                23.1
3       247837.0                  2.4        4021.0               -91.1
4       232887.0                 30.9       10073.0               232.1

   Assets ($M)  Assets ($M)Growth  Total Stockholder Equity ($M)  \
0     219295.0                0.0                        72496.0
1     346196.0                0.0                       191794.0
2     365725.0                0.0                       107147.0
3     707794.0                0.0                       348703.0
4     162648.0                0.0                        43549.0
```

	Total Stockholder Equity ($M)Growth	Profit as % of Revenues
0	0.0	1.3
1	0.0	7.2
2	0.0	22.4
3	0.0	1.6
4	0.0	4.3

	Profits as % of Assets	Profits as % of Stockholder Equity
0	3.0	9.2
1	6.0	10.9
2	16.3	55.6
3	0.6	1.2
4	6.2	23.1

	Earnings Per Share ($)	Total Return to Investors (5 year, annualized)
0	2.26	6.1
1	4.88	-4.3
2	11.91	16.6
3	2446.00	11.5
4	20.14	30.4

	Total Return to Investors (10 year, annualized)	Market Cap (M)
0	7.8	0.0
1	1.5	344980.0
2	30.8	666252.0
3	12.2	335798.0
4	40.2	293398.0

	No_Directors	Median_age	Board_Independance	Median_Tenure	Median_pay
0	0.0	0.0	0.00	0.0	0.0
1	12.0	65.0	0.92	5.5	360513.0
2	8.0	64.0	0.88	4.5	317829.0
3	12.0	66.5	0.67	12.0	2700.0
4	10.0	64.0	0.80	11.5	0.0

	women_on_board	GD_Approval	HBR Ranked	GD Ranked	GD & HBR Ranked
0	0.00	0.00	0	0	0
1	0.17	0.00	0	0	0
2	0.25	0.92	1	1	1
3	0.25	0.00	0	0	0
4	0.30	0.00	0	0	0

	MarketSize: < 250 million	MarketSize: 250 ~ 1,000 million
0	0	0
1	0	0
2	0	0
3	0	0
4	0	0

```
   MarketSize: 1,001 ~ 5,000 million  MarketSize: 5,001 ~ 50,000 million  \
0                                  0                                    0
1                                  0                                    0
2                                  0                                    0
3                                  0                                    0
4                                  0                                    0

   MarketSize: 50 Billion +  Aerospace & Defense  Apparel  Business Services  \
0                         1                    0        0                  0
1                         1                    0        0                  0
2                         1                    0        0                  0
3                         1                    0        0                  0
4                         1                    0        0                  0

   Chemicals  Energy  Engineering & Construction  Financials  \
0          0       0                           0           0
1          0       1                           0           0
2          0       0                           0           0
3          0       0                           0           1
4          0       0                           0           0

   Food & Drug Stores  Food, Beverages & Tobacco  Health Care  \
0                   0                          0            0
1                   0                          0            0
2                   0                          0            0
3                   0                          0            0
4                   0                          0            0

   Hotels, Restaurants & Leisure  Household Products  Industrials  Materials  \
0                              0                   0            0          0
1                              0                   0            0          0
2                              0                   0            0          0
3                              0                   0            0          0
4                              0                   0            0          0

   Media  Motor Vehicles & Parts  Retailing  Technology  Telecommunications  \
0      0                       0          1           0                   0
1      0                       0          0           0                   0
2      0                       0          0           1                   0
3      0                       0          0           0                   0
4      0                       0          1           0                   0

   Transportation  Wholesalers       State   Region            Division  \
0               0            0    Arkansas    South  West South Central
1               0            0       Texas    South  West South Central
2               0            0  California     West             Pacific
3               0            0    Nebraska  Midwest  West North Central
```

```
4              0            0  Washington     West            Pacific

   Midwest  Northeast  South  West  East North Central  East South Central  \
0        0          0      1     0                   0                   0
1        0          0      1     0                   0                   0
2        0          0      0     1                   0                   0
3        1          0      0     0                   0                   0
4        0          0      0     1                   0                   0

   Middle Atlantic  Mountain  New England  Pacific  South Atlantic  \
0                0         0            0        0               0
1                0         0            0        0               0
2                0         0            0        1               0
3                0         0            0        0               0
4                0         0            0        1               0

   West North Central  West South Central
0                   0                   1
1                   0                   1
2                   0                   0
3                   1                   0
4                   0                   0
```

4 Conclusion for this Stage of Feature Engineering

Fortunately, much of the data we have does not require further engineering due to the rigor of the scraping previously done. It is possible that upon further analysis we will recgonize the need for some other derived variable based off of the ones we have, but otherwise, one-hot encoding the geographical location according to the Cenesus Bureau regions, the sectors the companies belonged to, and whether a CEO was ranked or not was much of the general adjustment we needed after cleaning the variables to be their respective numeric or string variables.

As it stands, our adjustment of our project proposal remains more or less the same as it was since the data cleaning checkpoint. Perhaps the only difference would be the features we have elected to concentrate on, as seem by the list "features of interest." We also recognize that a question of reverse causality can be answered with the data we have, as the overall performance of a company could have an effect on whether a CEO is recognized and thereby ranked or not, and thus we intend to investigate this possibility in addition to our original proposal

At this point, we anticipate analyzing the data to answer question of a CEO's influence on a company's deliverables not just in aggregate, but also segmented by different areas, ergo the one-hot encoding of different variables. We are curious to see if the effect of a ranked CEO may have more sway in a Technology versus Retail center, or on the coast versus south part of the US. Fortunately, with the variety of variables we have, we can also see if different forms of growth differ between different segments. It could be that A successful CEO improves revenue growth in the technology sector, but not profit, and vice-versa for a company in the Retail sector. These are questions we are now posed to answer with the data we have obtained, cleaned, and engineered.