Fortune for Whom?

Examining the relationship between a Fortune-ranked company and its likelihood to invest in green building

What's changed…

- Scaled the scope back considerably
    - Allowed just two proxies for sustainability and gender
        - Use of LEED for sustainability
        - Percentage of women on board of directors
- Focused question to a classification, setting use of LEED as a binary response and gender equity as two possible features

The Data

From Hoovers

- Fortune 500 Ranking
- Company Name
- Contact Prefix
- Revenue (US Dollars, million)
- Revenue Growth (%)
- Net Income (US Dollars, million)
- Total Employees
- Employee Growth (%)
- Total Assets (US Dollars, million)
- Market Value (US Dollars, million)
- Year of Founding
- Primary Industry

From USGBC

- Whether a company uses LEED (meaning registered or certified projects owned or administered by a company)

External + Manually Sourced

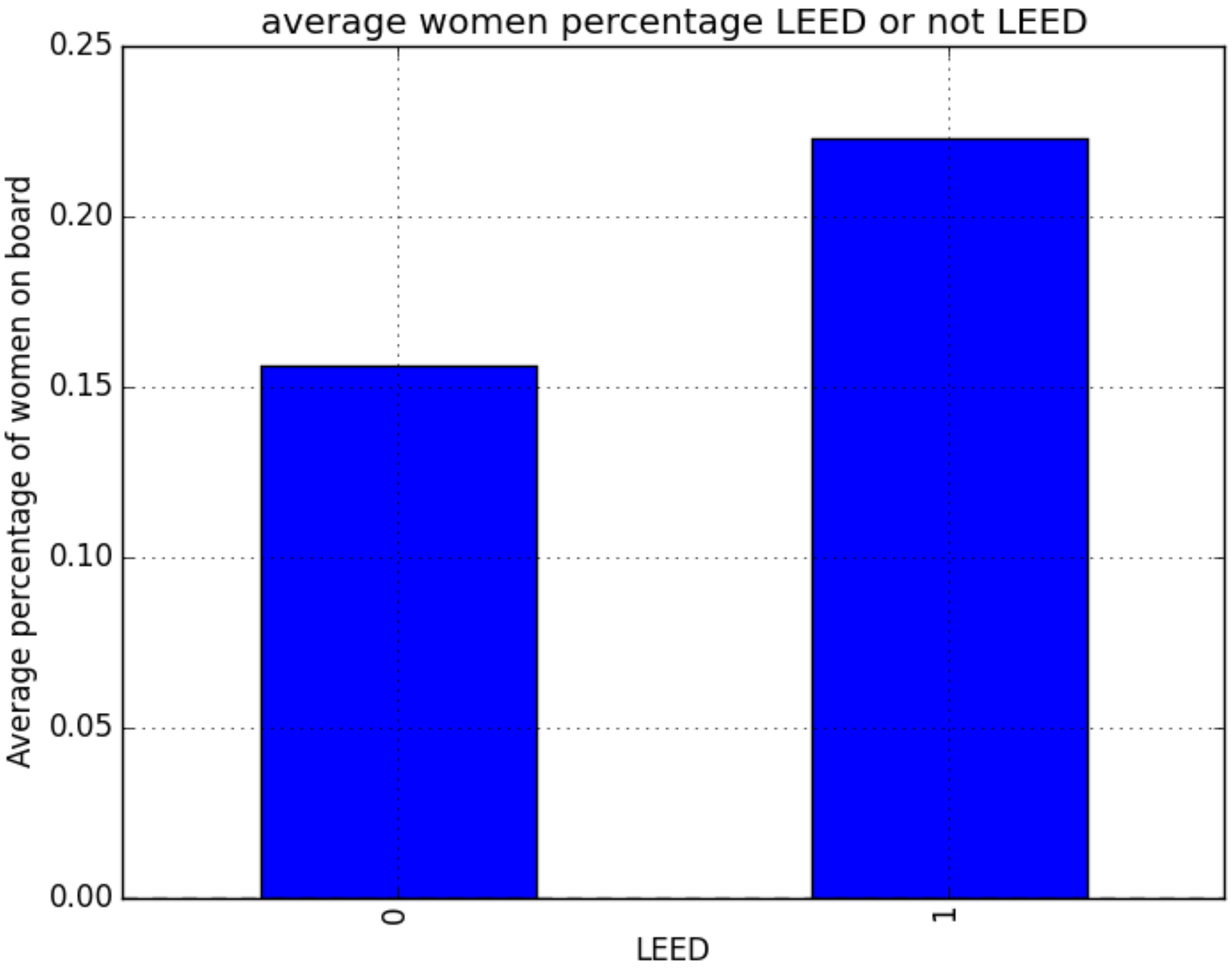
- Total number on board of directors
- Number of women on board of directors
- Number of men on board of directors
- Percentage of women on board of directors

Pre-treated Data

- Converted all binaries to 1/0= LEED (y=1, n=0), and chairman/CEO (Man=0, woman=1)
- Converted revenue and market value to all integer
- Left revenue growth and employee growth as a float
- Employee growth and revenue growth: Replaced null values with 0
- Market value: replaced all null values with average of the data from same sample (ie Fortune 100 average for the null values in the Fortune 100)

Data Dictionary

LEED: translated from 'Use_LEED': binary of whether a company is a project admin or owner of a LEED project (y=1, n=0)
Fortune_Ranking: Ranking within the 2014 Fortune 100
Total_BOD: total number on board of directors
Men_BOD: total number of men on board of directors
Women_BOD: total number of women on board of directors
Pwbod: renamed from 'Per_women_BOD': percent of women on the board of directors
Company_Name: object: name of company
CEO_Prefix: Mr., Ms., Dr., prefix of the CEO or Chairman listed for the company. This was converted into a binary based on gender in the source file
Rev: Revenue (US Dollars, million)
Revenue_Growth: revenue growth (%)
Net_Income: Net income (US Dollars, million)
Total_Employees: total number of employees
Employee_Growth: employee growth (%)
Total_Assets: total company assets (US Dollars, million)
Market_Value: total market value (US Dollars, million)
Year_Found: year company was founded
Prim_Ind: object: primary industry of the company


average women percentage LEED or not LEED
Average percentage of women on board
0.25
0.20
0.15
0.10
0.05
0.00
0
1
LEED

The Model

Started with KNN

- started with a KNN model
- issue getting any movement in the accuracy score even after manipulating features included
- realized the weakness of KNN being the scale of variables

Moved to LogReg

- Also had issues with seeing reaction in model to manipulating outcome in accuracy

Landed on RandomForest

- Found that in fact there was a higher AUC score when all features were included rather than just the top 5.

rfclf.feature_importances_
zip(rfclf.feature_importances_, feature_cols)

Fortune 100 possible feature importance ranking

(0.33904329604128142, u'Total_Employees'),

(0.10538148998910525, 'Fortune_Ranking'),

(0.1027924649481555, u'Market_Value ')

(0.09906472255053636, u'Total_Assets '),

(0.08614533293924298, u'Rev'),

(0.085220910020266896, u'Net_Income'),

(0.071456392568709837, u'Employee_Growth'),

(0.05367953491580657, u'Revenue_Growth'),

(0.051753832800620791, u'pwbod'),

(0.0054620232262745146, u'CEO_Prefix'),

On Fortune 100

With all possible features included
LogReg Accuracy score: 0.8
AUC: 0.8492063492063491 9

With top 5 selected features:
LogReg Accuracy score: 0.8
AUC: 0.8015873015873016 3

What next?

- want to do some more isolated variations to find the highest AUC score combination of features
- just finished the Fortune 200 and will test the model
- then finish the 500 and test model again
- train test split on the full 500 to see if it increases the model accuracy of the random forest outcome

Ultimately would like to continue to add features and test relationship between sustainability, equity and revenue

Applications

- Plan to present this as a starter conversation to our business development team