Daniel Mejia

DSC-680

Third Project – Race Hiring

**Business Problem**

The business problem would be to keep consistency in the hiring protocol and to detect any probabilistic percentage changes beyond a threshold in the company or team. Currently, the percentage are only being applied to the whole company but with the right data, we can apply the data to teams and watch trend of hiring from managers detecting inconsistency hiring.

**Background/History**:

According to Greet, the industry leadership roles are exclusively for white man. The data gathered by him concluded the underrepresentation of Hispanic and Blacks. He points that about 67% of companies are made of less than 5% of black employee (1). Also, he points out companies withholding the information to release the data of hiring to the public. Race and gender both affect leadership positions, however, Greet focuses on race and not the gender inequality. The tech industry is represented by white man mostly and shows a major problem.

**Data Explanation**:

The data source will be Kaggle and for now one documents have been found further exploration for use and implementation is required. Job category prediction from the Dataset. The file Reveal\_EEO1\_for\_2016.csv contains all the data to perform the analysis.

The data field were mostly in categorical type, this was not optimal for AI and machine learning models. Categorical fields cannot really be handled by most models properly and need conversion to numerical values for processing. The process was made more difficult because the data was mostly categorical.

The data set contains the following features for flight prediction:

**~~Index~~** – \*\* ***Removed*** \*\*

**Company** – Name of the workplace employee conducts business

**~~Year~~**– Year data was recorded **\*\*Unused\*\***

**Race** – ethnicity of the person employed

**gender** – male or female

**job\_category** – Type of job being performed

**Count** – amount of people in position

In our process to create the value to find the chances and outcome of employment according to position. A new variable had to be ***created***, probability

**Probability** – percentage of employee in the current company position based on race and gender.

**Methods**

Standard EDA will be used to explore and analyze the data. Also, a step further will be done by getting data important information to make recommendations. The data was explore using the provided function available to data frames such as describe, and similar methods. The standard deviation and mean will be measure for groups to understand anomalies and percentage above or below a threshold.

**Assumptions**

My assumption aligns with Greet; I assume most leadership roles being held by white man. However, I also wanted to know the second race to hold the throne. Hispanic were my second choice when it came to leaderships positions.

**Analysis**

Data was mostly in categorical values and meaning must be obtained to get the values and answer required if these exists in the data. Trying my typical approach of graphing data to see pattern, I found the following to yield almost nothing to the user.

Chart, scatter chart

Description automatically generatedThere might be a pattern in this graph, but it might be hidden as I could not determine a pattern or algorithm to fit this pattern. I decided to approach the problem in another manner and create the probability of the data depending on the company. Currently, I am having trouble aligning the data to fit the dataframe.

After creating the probability for each job category from the count of employees. I was able to obtain data to facilitate with the analysis. I found most position to be hold by white and the second race to hold leadership positions are Asian.

A picture containing chart

Description automatically generated

Reading the above graph would require the following keys to be used for analysis. Yellow is woman and purple man.

Text

Description automatically generated

**Limitations**

The data was a problem since the fields were categorical and provided little to no value to the outcome of the field. However, the little data I required was enough to create a classification model. The data was not large to affect my computer performance, I was able to run the code with almost no delay.

**Challenges**

The problem came from the categorical values in the data. I have work with classification system coming from continuous values but not the other way around in the data. The models were giving me problems and errors mixing values. I solve the problems and learn new problem-solving skills. Once the data was graph and transform into the last form, plugin the data for model consumption also yield problems.

**Future Uses/Additional Applications**

The project is not a representation of people length at the companies, and these might result in a lack of displaying why most of the leaderships positions might be held by a certain race. We need more data to analyze the length of employees and those who hold powerful positions to study a correlation.

**Recommendations**

This study must be taken with a grain of salt since is not up to date. The file is from 2016, I hope now in 2022 there has been a change in companies leaderships position and changes have been done to fix the problem of races in these positions. Companies should be inclusive to all races; one must feel like we could obtain these positions if we work hard enough for the company.

**Implementation Plan**

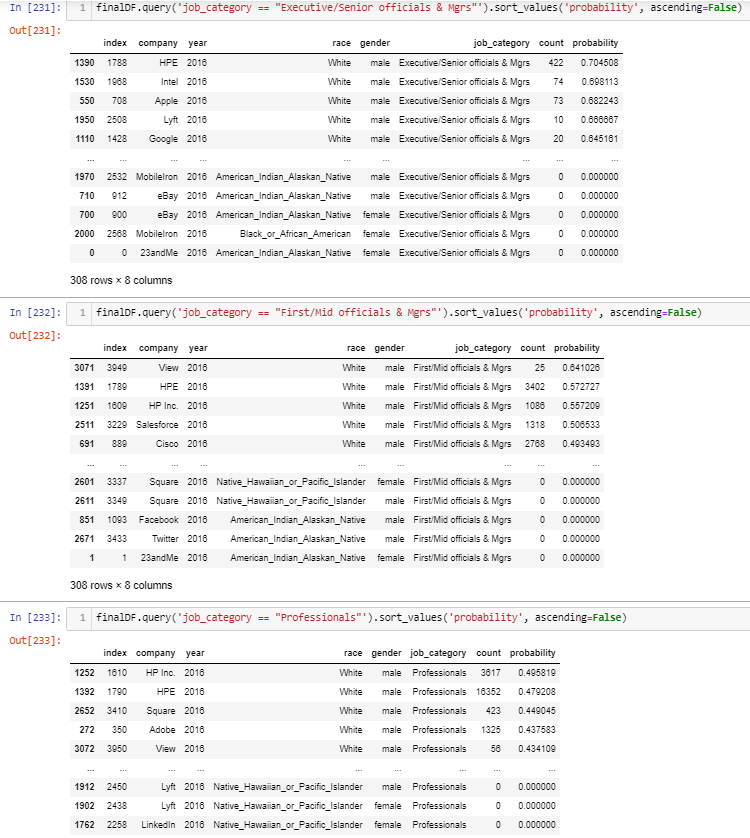
Companies should always be assessing the percentage of races within the leadership positions and making sure to promote people of these underrepresented races who have earned and work hard to lead. A model like this one which can be used to find the percentage of employees in these positions could be beneficial to a company. The minimal data should allow the model to run extremely well for monitoring of percentages. Then, an alert system could be implemented if the levels of certain positions get below the desired threshold.

**Ethical Assessment**

The model created could also cause damage to a company reputation if the levels are below the standard for the company. It could have the consequences of putting the company on a bad light, by pointing the fact that their hiring of Latino and other ethnicities is way below company’s standards.

**Conclusion**

After the analysis, I found most leaderships positions to be hold by white man. The probability was undeniable. Also, the correlation between leadership positions had a strong percentage with people race. In the picture below, we can see the probability of percentage compared to other races. HPE, Intel, and Apple closeness to 70% white male for executive and seniors’ positions should be a warning sign to stay away if you are not within this race.



The model were classifications since I wanted a model to yield the percentage of people in leadership positions. This are exact numbers and not to be guess. The models yield the following results.

Text

Description automatically generated

**Questions**:

1) How much will it cost to implement the system?

The implementation of such system should not be expensive since the data is coming from the company and any team can be assigned to gather the data for pipeline implementation. This would be HR responsibility in my terms of the job.

2) How difficult is to implement the pipeline to the model?

The pipeline should be easy because of the minimal data to make the model work and anything else could just be added to a different model creating modules which handle different part of the company.

3) Reducing limitation on the model?

Adding more data which will add more problems to the company and merging this data with time prolong and historical analysis could yield even more answers in team building and race promotion within the company.

4) What is the accuracy and if required how to increase it?

The accuracy is bound to the hiring and people being promoted to leadership positions. The models should always be accurate at determining these probabilities.

5) Is the business model being solve by the model?

The model answers the basic question of the people with the leadership position and yes it does yield the answer required to determine if changes are required within the organization and promotion to leadership positions.

6) Scalability and maintenance requirements to keep it functional?

Since the data is minimal maintenance won’t be a problem for the model to keep up, unless a more complicated system is required to identify more in depth the companies problem with promotion.

7) How reliable is the system?

The system will not fail as the numbers yield only come from the hiring of the company and cannot be distorted to be untrue. We are getting the real value and people at the leadership positions at the time of the analysis and data provided.

8) How will the company make profit?

The company will make profit by attracting the best candidates for the job from all races instead of limiting the leadership to one race and not taking advantage of different mentalities in the power positions.

9) What are the benefits of implementing the system?

The benefit would be to keep a company where everyone feels inclusive and can talk to someone of their own race. Also, these employees should be able to find someone in a leadership position to be able to help them get to the same positions.

10) Are there any issue with the model and how can they be fixed?

The model can only give exact number, we could change it for a probabilistic model for guessing instead exact data. I also must refine my understanding of ML models to make the model work better at guessing these probabilities.

**Reference**:

1. Greet, S. (n.d.). *Racial Diversity In Tech By The Numbers*. Beamjobs.Com. https://www.beamjobs.com/diversity/racial-diversity-in-tech