Business Characteristics and their Effect on Company's Annual Pay

OSMBA 5067 Spring 2021 – Data Translation Challenge Milestone #4

Lynna Tran

Introduction

Project Problem

The project problem I am interested in solving using machine learning algorithms is whether a company's annual pay can be determined by business characteristics such as employee size, number of firms, and types of business owners. It is a business-related problem because I am comparing the characteristics of businesses and its annual pay in order to test if a relationship exists between types of businesses and their pay.

Motivation

My motivation for doing this research is that in my own career, I am an employee of a company that has about three hundred employees which is considered a medium-sized company. In today's business climate, it seems that large, well-known, and international corporations attract the best employees in their fields with offers of high pay, job satisfaction and is seen as the end goal for many careers. I am curious to learn if a big company such as Amazon, Apple, and Google actually do provide better pay, compared to other sized companies or if it is just brand recognition that appeals to workers. The answer to my research problem will help small to medium sized companies recruit top talent in competition with the bigger companies.

Literature Review

During my research into prior literature about my project problem, I learned that the relationship between employer size and wage has been studied often in the last 30 years. It seems like a popular topic because economists are interested in what drives employee's wages.

In *Employer-Size Wage Effect* by Brown, Charles, and James Medoff in 1989, they studied the relationship between employee size and wage. What they concluded is that the relationship is not necessarily between size and wages but that large firms are usually older firms. They hypothesize that in reality, the relationship might really be based between an employer age and wages. The study by Heyman, F in *Firm Size or Firm Age? The Effect on Wages using Matched Employer-Employee Data*, expanded on the question posed by Brown, Charles, and James Medoff. They found that there was a positive employer age and wage relationship. They contribute the relationship to systematic differences and not from business characteristics. Employee tenures was a factor between young and old firms along with individual differences in human capital. Longer companies had more employees stay longer and were often hiring higher positions which increased wages.

The Relationship between Owner Characteristics, Company Size, and the Work-Family Culture and Policies of Women-Owned Businesses by Adkins, Cheryl & Samaras, and Steven & Gilfilla, they specifically looked at the relationships between women-owned businesses and whether it has an effect on company size and work-family culture. They found that the size of businesses was associated with family-friendly policies, but not associated with work-family culture or necessarily related to higher pays. I thought the research negatively portrayed women in business, based on the idea that women must want better work-family culture, mainly because women are usually more empathic to family benefits. It was assumed that all women's main priority is family and ignored that women might have other ambitions as a business leader. It would have been better to compare these businesses to men owned companies and their work-family culture instead of just focusing solely on women.

Experiment

The dataset was dense with information because it came from a credible source such as the Census data which collects large amounts of information from businesses and then analyzes it in several different reports. There was a lot of features to choose from, especially between categorical and numerical values.

Dataset Source

For the public dataset used in the data translation project, I used the Annual Business Survey on Business Characteristics in 2017 and 2018. This survey is conducted through the Census Bureau. The Census Bureau releases a survey to sample businesses across the United States that have business licenses and pay business taxes. The dataset is a combination of responses to the survey, data from the economic census, and administrative records data. The Annual Business Survey on Business Characteristics is just a subset of the Annual Business Survey with the survey itself encompassing a range of topics such as business incomes and business diversity, then divided into topics for reporting.

The dataset by the Census Bureau is conducted on a company or firm basis rather than an establishment basis. A company or firm is defined by the Census as a business consisting of one or more domestic establishment under its ownership or control. For confidentiality reasons, the businesses are grouped together by certain characteristics that are shared. That way, important business characteristics cannot be traced back to an individual business.

Variables

The categorical features are often defining characteristics about the business while the numerical features specify information about the company. The categorical values used in my research are employee size, owners' sex, and race. Employee size is given in range of employees such as this group of companies had an employee size between 1-49 employees. Owner sex is defining if an owner is female or male. If there is more than one owner, an owner sex can be equally male, and female owned or if it is majority owned by males or females. Owner race is also categorized by what race demographic the owner belongs to. Figure 1 shows the breakdown of number of businesses per owner characteristics.

The numerical features are given when the company characteristic can be defined in numerical terms. In my research, I used year of the survey, the number of owners, firms, and overall annual pay. The year of the survey could be either 2017 or 2018 as this is when the Annual Survey was delivered. We are looking at the number of owners for each type of business, such as whether it is owned by one individual, a family, or even a group of shareholders. The number of firms is defined as how many firms does a company operate like in cases when a business has more than one location. The annual pay of a company is the most important feature used in my research project because it is also my target value. The previous features are used as my dependent values to see if their effect on annual pay, my experiment's independent value. Figure 2 shows the annual pay of companies by their categorical employee size.

The target value was broken up into three categories: low, medium, and high pay. They are defined as 0, 1, and 2, respectively in my dataset. I was able to sort these values by dividing the range of annual into thirds and placing a category based on where the business' annual pay falls into.

Method

To solve my research problem, I will be implementing machine learning algorithms to find whether business characteristics affect annual pay of a company. I will be using the KN Neighbors and Naïve Bayes method to test my dataset.

I chose to use the KN Neighbors method on my dataset because of the multitude of features that was present. KN Neighbors' advantage is exploring large amounts of data points to navigate the closest target value, so I thought that was useful in my research due to the ranges of values used to characterize businesses and that annual pay.

For the Naïve Bayes models, I used the gaussian and categorical techniques. I ran these two models because there are two types of values in my dataset. They can both be used because Naïve Bayes is a probability function, so I am able to run these two models separately and then multiple the probabilities together.

Implementation

The data is already on a firm-level, so it has been grouped by category of business. When we are thinking about the results of this research, we have to remember that we are talking about groups of businesses. So, in a row of data, we might read it as ‘companies that are female-owned and have a range of 10-49 employees average about 1 million dollars in annual wages’. The data is already consolidated by the Census Bureau so we must believe that they have done their due diligence in aggregating the data already.

The categorical values were converted to separate binary values in order to run the KNN models. We want these variables to only have two elements, True or False in order to set up the data points that will be used to calculate the distances between them. It would have been difficult to use categorical values in a KNN model because the algorithm will not know how to calculate the distance between data points that have several elements. For example, the gender of the owner was separated into six separate binary values for whether the business has a women owner or not, male owner or not, and so forth.

For the Naïve Bayes models, I was able to leverage off that Naïve Bayes has a few different techniques for different types of variables. I used the Categorical Naïve Bayes on the categorical values because it is suitable for features that are already categorically distributed. The Categorical classifier is able to produce probability on the occurrence of each category. For the numerical values, I implemented the Gaussian Naïve Bayes model. This technique worked best because the variables were already normal distributed as they were counting the numbers of owners, firms, and annual pay.

Results

The result from the KN Neighbors has an error rate of 3% with a K-value of 1. Using the K-value of 1 seems most promising as the error rate is at 3.49% which is the same as when the K-value is at 2. When the K-value is 3, the error rate increases slightly to 3.5%. The error rate increases again at K-value of 4 to 3.6% but drops at 5 K-value to 3.4%. Even though the error rate seems to decrease at 5 K-value, sing a small K-value is best suited for my research because the target value only has 3 elements available, so it is advantageous to keep the data points close together. A large K-value might be misleading since it is encompassing a wider range of data points. Figure 3 shows the K-value by error rate.

The result from the Naïve Bayes model has an error rate of 1%. In their separate models, the Categorical Naïve Bayes models has an error rate of 45% with a probability of predicting the target value at 25%. For the Gaussian model, the error rate is 7% and a probability of less than 1%. Since these two models are done on separate values, I am able to multiply the probabilities together in order to get a probability of the entire dataset predicting the correct target value, which is the range of annual pay. The probability of the whole dataset is less than 1%. That shows that our categorical values used in our dataset seem to have a higher relationship to business’ annual pay then its numerical values. The data that we gain from the business characteristics such as owner demographics and employee size have a bigger impact than number of owners and firms.

Conclusion

Both of my models that I used in my research resulted in low error rates, with the KN Neighbors at 3% and the Naïve Bayes model at 1%. This shows that using a classification and a probability machine learning models is well-suited for my research because it allows us to look at certain features of business and accurately predict the pay of the companies. It shows that business characteristics do have an impact on the annual pay and follows a predictable trend in how much a company’s annual pay will be. That gives me some insight that big companies are able to provide better wages than smaller companies.

If time and resources were unlimited, there will be several things I would implement in future work. I would expand on my dataset by adding more features and elements in my target value. I only used seven features in my research so if I would expand upon them, I can minimize the error term and reflect the realities of businesses better. There will also be more elements in the target value instead of just low, medium, and high pay to understand more fully the intricacies of how company’s wages can be defined.

References

Adkins, Cheryl & Samaras, Steven & Gilfillan, Sally & McWee, Wayne. (2013). The relationship between owner characteristics, company size, and the work-family culture and policies of women-owned businesses. *Journal of Small Business Management*, *51*.

Ady Milman. (2003). Hourly employee retention in small and medium attractions: the Central Florida example. *International Journal of Hospitality Management*, *22*(1), 17-35.

Brown, Charles, and James Medoff. (1989). The employer size-wage effect. *Journal of Political Economy*, *97*(5), 1027–1059.

Heyman, F. (2007). Firm size or firm age? The effect on wages using matched employer–employee data. *LABOUR*, *21*, 237-263.

J.J. Ramsden, Gy. Kiss-Haypál. (2000). Company size distribution in different countries.
Physica A: Statistical Mechanics and its Applications, *277*, 1–2