MGT 6203 Group Project Proposal Template

TEAM INFORMATION (1 point)

Team #: 19

Team Members:

1. Luke Burdette; lburdette7. Bachelors in computer science with minor in Data Science. Previously experience with data science projects in both Python and R involving biostatistics, demographics, and manufacturing.
2. Reagan Buske; rbuske3. Current data scientist in the Ag financial services industry with a bachelor's degree in data science. Have worked on a large variety of analytics projects in undergrad and at work.
3. Zoe Chapman; zchapman9. I have a bachelor's degree in Physics & Astronomy. I have worked on a variety of research projects in astronomical data analytics, and I have worked professionally as a data analyst in the healthcare sector.
4. Christopher Romain Baker; cbaker92. I am a data analyst who graduated with a bachelor's degree in Mathematics, and have worked individually and in groups on various projects within the financial and travel industries.
5. Taylor Buske; tbuske3. Current business intelligence analyst in financial services with a bachelor's degree in statistics and a minor in general business. Have worked on general data visualization, summarization, as well as forecasting in my job.

OBJECTIVE/PROBLEM (5 points)

Project Title: Labor Union Membership Prediction from Census Data

Background Information on chosen project topic:

This dataset provides census data leading through the 20th century, providing various information. This includes data on income, race, economic status, stock market investments, and whether they are a member of a labor union. A labor union is an organization of workers who may collectively seek higher wages, safer working conditions, better benefits, and/or increased employee leverage.

Problem Statement (clear and concise statement explaining purpose of your analysis and investigation):

This project's objective is to build a logistic regression model to predict the likelihood of someone belonging to a labor union.

State your Primary Research Question (RQ):

What are the most important factors that indicate whether someone is a member of a labor union?

Add some possible Supporting Research Questions (2-4 RQs that support problem statement):

1. Are the most influential predictors for someone's labor status easy to intuit, or surprising?
2. Is there a strong correlation between income and financial security in predicting labor status?
3. Can some of these influential predictors determine the desire of future employees to form or join a pre-existing labor union?
4. Is there a pattern in the data for the rows that have null values for the labor union column?

Business Justification: (Why is this problem interesting to solve from a business viewpoint? Try to quantify the financial, marketing or operational aspects and implications of this problem, as if you were running a company, non-profit organization, city or government that is encountering this problem.)

Businesses may use this information to gauge whether the circumstances around a set of employees make them likely to form a union. Labor Unions have specific merits and detriments that employers may need to consider. From a business perspective, unions can increase motivation, productivity, and retention of workers. On the other hand, unions can lead to higher costs and a less direct relationship with employees. If a business does not yet have employees in a union, they could use our model to predict if a union is likely to form based on the characteristics of their employees. This information could help them better budget and plan for a potential future union forming.

DATASET/PLAN FOR DATA (4 points)

Data Sources (links, attachments, etc.):

https://www.kaggle.com/datasets/kamaumunyori/income-prediction-dataset-us-20th-century-data/data

Data Description (describe each of your data sources, include screenshots of a few rows of data):

The dataset will include age, gender, level of education, class, university enrollment status, marital status, race, part-time or full-time employment, unemployment reason, whether self-employed, wage per hour, industry, weeks worked per year, number of employees working for them if self-employed, household statistics (homeowner, has children, etc.), tax status, stock gains and losses and dividends, citizenship, migration year, country of birth, parents' countries of birth, migration info, whether they have lived at one location more than one year, and whether their income is above or below 50k annually.

These will ultimately be used to predict whether someone belongs to a labor union or not.

ID	age	gender	education	class
ID_TZ0002	21	Male	12th grade no diploma	Federal g
ID_TZ0005	45	Male	Bachelors degree(BA AB BS)	Private
ID_TZ0006	53	Male	High school graduate	Private
ID_TZ0010	30	Male	High school graduate	Local gove

...

residence_1_year_ago	income_above_limit	is_labor_union
Same	Below limit	No
Same	Below limit	No
	Below limit	No
	Below limit	No

Key Variables: (which ones will be considered independent and dependent? Are you going to create new variables? What variables do you hypothesize beforehand to be most important?)

Our dependent variable is "Is part of labor union". The rest of the variables provided in the dataset will be independent variables. We will use a model to determine which are the most significant.

We plan to create new variables such as a college education indicator (yes or no), age groups, from the US vs. not from the US, as well as fixing any variables with null values that do not make sense.

The variables we predict will be most influential are: Wage per hour, Working weeks per year, Main industry code, Main occupation code, Employment commitment, Number employed, Country of birth own.

APPROACH/METHODOLOGY (8 points)

Planned Approach (In paragraph(s), describe the approach you will take and what are the models you will try to use? Mention any data transformations that would need to happen. How do you plan to compare your models? How do you plan to train and optimize your model hyper-parameters?))

A portion of our training set does not include complete data for various fields, including our dependent variable. We will first address the gaps in our dependent variable, determining whether this data can be collected from other sources, or whether it must be removed. If the data is to be pruned from our dataset, we must determine if the associated independent variables that have been pruned have an underlying distribution. Essentially, we are determining if the missing data is missing for a reason, and thus should not be modeled. If this need arises, we will add a new categorical variable is_union_missing and create interaction terms between this new variable and all other independent variables, thus creating two unique datasets (and models): one in which our labor data is missing and the other in which it is complete.

Following this, we will check the completeness of our independent variables. If a field has less than 5% of its values missing, we will consider imputation. Otherwise, we will need to use a different methodology to determine if there is a pattern in the missing data. We will next assess the imbalance in our dataset that exists for our response variable. Our dataset is heavily imbalanced towards individuals who are not a part of a labor union. As such, we will need to look towards whether there should be some level of oversampling or undersampling to address this discrepancy.

Given that our response variable will be of the form “Yes”/”No,” it seems natural to begin with a logistic regression. We will fine-tune the proper cutoff for our probability, p, and use a confusion matrix to determine how accurate we are. Given the dataset's nature and how there are fewer positive labor union responses, we may need to tweak the model to increase the number of True Positives. Naturally, this will mean a lower threshold for a positive response and risk more False Positives.

Several logit models will be created using varying combinations of the independent variables to determine their net impact on our dependent variable. We will then examine the models, factor by factor, to draw conclusions and determine whether there are some factors with very little impact or very high impact.

Anticipated Conclusions/Hypothesis (what results do you expect, how will your approach lead you to determining the final conclusion of your analysis) Note: At the end of the project, you do not have to be correct or have acceptable accuracy, the purpose is to walk us through an analysis that gives the reader insight into the conclusion regarding your objective/problem statement

We anticipate that certain features will be more influential in predicting if someone is in a labor union than others. Some of those features include Wage per hour, Working weeks per year, Main industry code, Main occupation code, Employment commitment, Number employed, Country of birth own. If we can create a statistically significant logistic regression model that has high accuracy, we will conclude that our model can predict labor union membership. A business can then use that model to predict if their employees are likely to form a union or not.

What business decisions will be impacted by the results of your analysis? What could be some benefits?

It is unlawful to discourage union formation or other union activities via discrimination regarding the hiring of new employees or any other employment condition. However, the likelihood of current employees forming a union may still be of interest to company shareholders and leadership. The results of this analysis may provide valuable insights for employers from a budgeting and financial planning perspective.

According to the US Bureau of Labor Statistics, unions can potentially result in higher labor costs for a company, and thus higher production costs overall. Furthermore, if wage demands aren't met in a timely manner, unionized workers are legally allowed to strike with low risk of termination. Strikes can lead to lost production time, harm public image, and even tarnish relationships with suppliers, all leading to lower margins. Additionally, union workers may have a bolstered willingness and ability to file a lawsuit against their employers. This can lead to higher risk of arbitration or litigation against the company, accompanied by harsh legal expenses. A company's ability to predict the formation of a union may help account for these potential costs in advance while granting time to adjust accordingly.

PROJECT TIMELINE/PLANNING (2 points)

Project Timeline/Mention key dates you hope to achieve certain milestones by:

The two most important dates that we are driving towards are 3/17 (progress report due) and 4/21 (final paper due).

We can expect to have feedback on our proposal around the beginning of March. At that time, we will regroup and make any changes to our desired problem to solve or questions to ask. By March 8, we will ideally have answers to our questions about the completeness of the dataset and our questions about the labor union variable (looking at if there may be any bias or interactions).

Following this, we will begin to create our logistic regression models and should be started with that by the time we write our progress report. Our progress report should be finished by March 16, so we have it turned in on time. We will continue to refine our model, and once we receive feedback from the progress report, we will take necessary actions from that.

Given the timeline and length of the final report, we hope to finish most of our modeling by the end of March. Will we then have plenty of time in April to write a strong final report and turn it in by April 20.

Appendix (any preliminary figures or charts that you would like to include):