MGT 6203 Group Project Proposal

Systematic Approach to Identify Factors to Optimize Income

TEAM INFORMATION

Team #: 23

Team Members:

1. **Ngoc Le, ngoc8:** I worked as an automotive QA, a Bachelor of Computer Science degree and data analytics certificate in R at Cornell University. In the previous course, I built and evaluated predictive models to determine the chance that a particular passenger may have survived based on the RMS Titanic dataset.
2. **Carlos Moncada, cgonzalez303:** I am an industrial engineer from Florida State University, currently working as a Data Scientist at a convenience store chain. Some of my key projects include developing an optimization model to recommend pricing strategies and utilizing a linear regression model to predict demand elasticity.
3. **Usha Sharma, usharma43:** Principal Engineer at Cisco, San Jose. I have a Bachelor of Engineering Degree from PEC, India, and courses towards MS Computer Science from Santa Clara University. I have worked on individual Analytics project for ISYE6501 and CSE 6040.
4. **Hal Zhang, hzhang957:** Medical doctor, biomedical engineering degree at Duke, medical school at Baylor College of Medicine, internal medicine residency at Baylor College of Medicine, individual project for CSE 6040, group project in college
5. **Brandon K Ritchie, britchie7:** Data Analyst for AgReserves which is an Agriculture Investment company headquartered in Salt Lake City. Some notable projects I have done include developing a longitudinal model to determine when to send cattle at our feedlots to be slaughtered, creating a commodity price forecast model for corn and soybeans, and developing a pipeline to track irrigation efficiency at our pistachio and almond orchards.

OBJECTIVE/PROBLEM

Project Title: Systematic approach to identify factors to optimize income

Background Information on chosen project topic:

Many have faced wage variation. According to the U.S. Bureau of Labor Statistics (BLS), the top-earning ten percent of athletes and sports competitors made more than $239,200, while the lowest ten percent earned less than $28,510. That is less than the $46,310 median wages for all occupations in May 2022 and nowhere close to what top athletes earn.

Wages also vary drastically between males and females. In 1982, women's median earnings were only 65 percent of men's. Although the gender pay gap has closed, women's salaries remain 80 percent under men's salaries since 2002 according to Pew Research. The 2021-2022 data from Basketball Reference shows the average male NBA player earns $5.3 million annually, compared with WNBA players who make an average of $130,000. In other words, a WNBA player earns less than 44 times what an average NBA player makes.

Even more disconcerting is the persistence of a racialized gender pay gap. Compared to white men, the depth of the under-payment of Black women and women of color is astounding. They lose out nearly one million in earnings over a career according to the National Women's Law Center.

Problem Statement:

Given there is so much inconsistency in income even in jobs of the same field as stated above, we want to investigate possible factors that lead to this income variation. The results can help us determine potential controllable versus uncontrollable factors. This analysis can not only help us personally to make informed decisions to improve our income but also help governments and businesses to try to form policy to decrease income variation.

Primary Research Question (RQ):

What are the main factors responsible for income variation?

Supporting Research Questions:

1. What are controllable factors (education, working hours, industry) that affect income?
2. What are uncontrollable factors (age, gender, citizenship) that affect income?

Business Justification:

There are policies that can be changed from a governmental perspective that could help with income variation:

- Try to eliminate discrimination based on certain demographics (being a citizen, being of a certain race, gender)
- Encourage more focused education given certain industries pay more

Businesses can use this data to pay their employees fairly which will help them retain their employees

DATASET/PLAN FOR DATA

Data Sources : Income prediction dataset (US 20th Century Data). | Kaggle

Data Description and Key Variables:

The dataset, constructed from a randomly chosen population, consists of roughly 200,000 individuals in the training set and approximately 100,000 individuals in the test set. It encompasses a wide range of attributes that detail demographic, employment, financial, and household information. For a comprehensive description of each attribute, please refer to Appendix 1, which contains screenshots of the dataset.

Dependent Variables:

- **Wage per Hour:** The wage earned by the individual per hour of work.
- **Income Above Limit:** Indicates whether an individual's income exceeds $50,000.

Most Important Independent variables:

- **Age:** The individual's age.
- **Gender:** The gender of the individual typically categorized as male or female.
- **Education:** The highest level of education attained by the individual.

- **Class:** The social class of the individual, inferred from income, occupation, or other socio-economic indicators.
- **Education Institute:** Information about whether the individual is currently attending an educational institution.
- **Employment Stat:** A numeric or binary indicator representing the employment status of the individual.
- **Is Labor Union:** A binary indicator of whether the individual is a member of a labor union.
- **Working Week per Year:** The number of weeks per year the individual works.
- **Industry Code:** A numeric code representing the industry in which the individual is employed.
- **Industry Code Main:** A textual description of the main industry category.
- **Occupation Code:** A numeric code representing the individual's occupation.
- **Occupation Code Main:** A textual description of the main occupation category.
- **Total Employed:** The total number of people employed in the individual's household.
- **Old Residence State:** The state of the individual's previous residence.
- **Importance of Record: Re**presents the weight of an instance in the dataset.

Less Important Independent Variables:

- **ID:** A unique identifier for each individual in the dataset. Independent.
- **Marital Status:** The marital status of the individual, such as single, married, divorced, etc.
- **Race:** The race or ethnicity of the individual.
- **Is Hispanic:** A binary indicator of whether the individual identifies as Hispanic.
- **Employment Commitment:** Details about the individual's employment status, such as full-time, part-time, unemployed, etc.
- **Unemployment Reason:** The reason for unemployment, if applicable.
- **Household Stat:** A description of the individual's household status, such as head of household, spouse, child, etc.
- **Household Summary:** A summary description of the household composition.
- **Under 18 Family:** Information about family members under the age of 18.
- **Veterans Admin Questionnaire:** Responses to a specific questionnaire for veterans.
- V**et Benefit:** Information on veteran benefits the individual may be receiving.
- **Tax Status:** The tax filing status of the individual, such as single, married filing jointly, etc.
- **Gains:** Financial gains reported by the individual, excluding wages/salary (e.g., from investments).
- **Losses:** Financial losses reported by the individual.
- **Stocks Status:** Information on whether the individual owns stocks or other financial instruments.
- **Citizenship:** The citizenship status of the individual.
- **Mig Year:** Information on the year of migration, if applicable.
- **Country of Birth Own:** The individual's country of birth.
- **Country of Birth Father:** The father's country of birth.
- **Country of Birth Mother:** The mother's country of birth.
- **Migration Code Change in MSA:** Code for change in Metropolitan Statistical Area (MSA) for the individual, indicating if they've moved between MSAs.
- **Migration Prev Sunbelt:** Indicates whether the individual previously lived in the Sunbelt region of the United States.
- **Migration Code Move Within Reg:** Code for the individual's movement within a region.
- **Migration Code Change in Reg:** Code for change in the individual's region of residence.
- **Residence 1 Year Ago:** Indicates the individual's residence status one year prior to the survey.

- **Old Residence Reg:** The region of the individual's previous residence.

APPROACH/METHODOLOGY

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?))

We will initially scatter plot the data to see if there are linear relationships between the predictor and response variables then we will fit linear regression models to predict income using wage as the response variable and controllable and uncontrollable factors as the predictors. We will then look at the p-values of the coefficients of the variables to see if they are significant, if they are not we will remove them and create another model. We will compare the initial linear regression with the model with less parameters using the testing data set to see which one more appropriately predicts income. Ultimately this will allow us to see which predictor variables affect income the most.

We will use the predictors above that are significant for a classification model, the response variable will be high or low income. We will use the training data set to create a classifier and test the accuracy of the model with the testing data set. The goal is to be able to predict someone's salary (high or low) given the relevant predictor variables.

For data cleaning, we will filter the data to limit the scope and remove any column that doesn't contain data relevant to the prediction of the models. The outliers and missing data will be assessed, removed, or replaced depending on their impacts. For example, if the two averages of the variable wage are hugely different, we will replace missing values with the average value. Otherwise, we will exclude those observations that have missing values.
Data transformations may be needed to resolve any violation of critical model assumptions and improve the models' performance metrics, such as increasing the R-squared and making the distributions more normal.

Anticipated Conclusions/Hypothesis (what results do you expect, how will you 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 hypothesize that there will be controllable and uncontrollable factors that affect income. Out of the controllable factors, we think higher amounts of education will increase average income. Out of the uncontrollable factors, we think that an increase in age up to a certain amount (50 years old) will increase salary but then will likely lead to a decrease in salary when the person gets older. We also think that males and U.S. citizens are like to have a higher salary.

One issue we predict we might run into is that many of the predictor variables we are comparing could be correlated with each other and therefore would need to eliminate some of the predictor variables to make reasonable conclusions or perform a PCA analysis. For example, age and amount of education could be correlated with each other. We will test this using correlation matrices and VIF.

What business decisions will be impacted by the results of your analysis? What could be some benefits?
For an individual: Decision making related to optimizing income and to enhancing skillsets which will give best return on education investment.

For a business owner: Based on the skillsets of the workforce and salary expenditure, business owners can make appropriate hiring decisions. Businesses can also create some training programs for the employees.

For a government planner: To bring down the disparity in income or to help low-income population. Government policies can be created if the income variation is found to be due to uncontrollable factors.

PROJECT TIMELINE/PLANNING

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

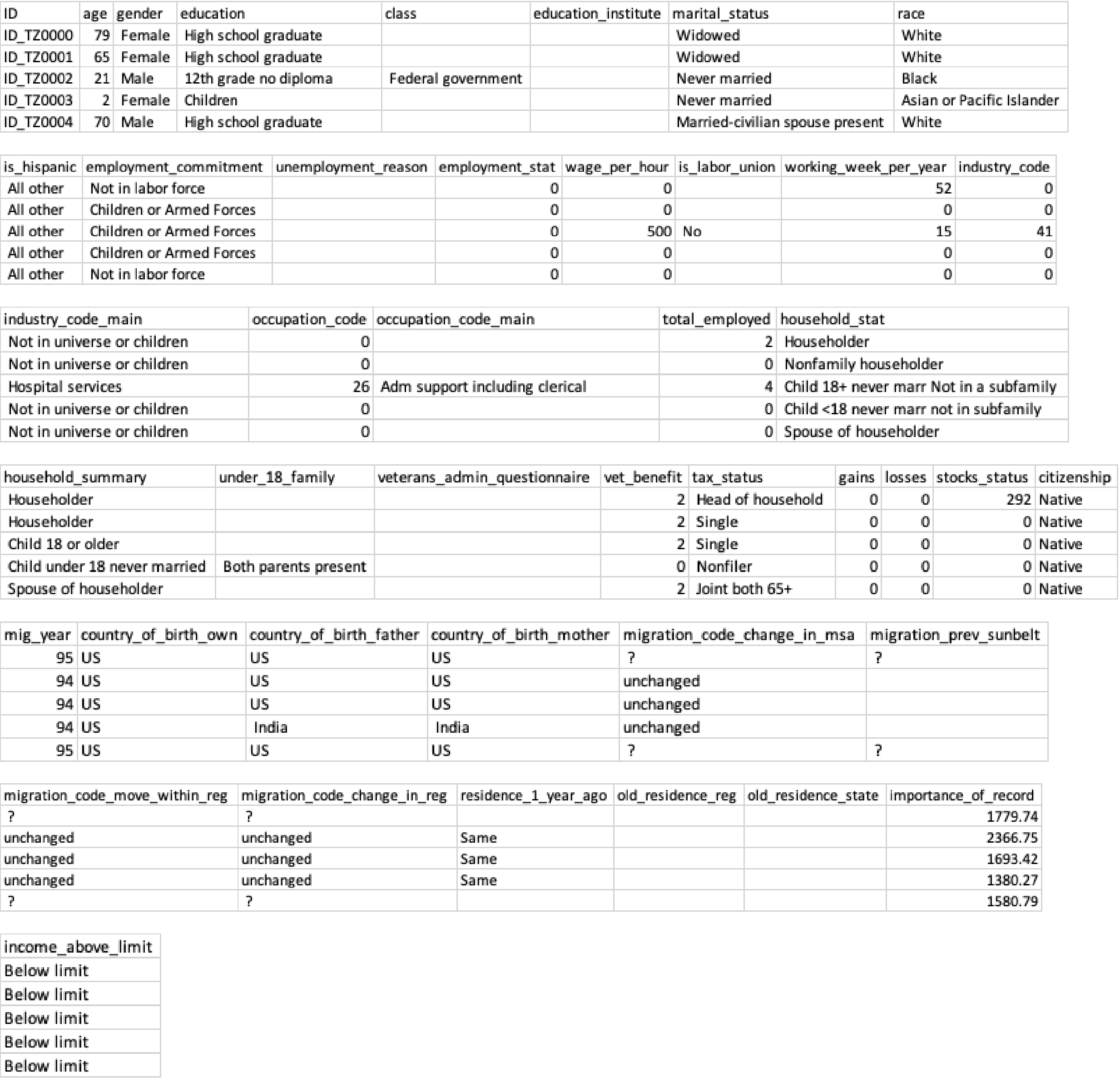
2/19- finish and submit project proposal

3/15- finish cleaning up data set, choose which models to analyze the data, finish 80% of data analysis, submit the progress report

4/14- finish the final report

Appendix:

Appendix 1: Screenshots of dataset



ID	age	gender	education	class	education_institute	marital_status	race
ID_TZ0000	79	Female	High school graduate			Widowed	White
ID_TZ0001	65	Female	High school graduate			Widowed	White
ID_TZ0002	21	Male	12th grade no diploma	Federal government		Never married	Black
ID_TZ0003	2	Female	Children			Never married	Asian or Pacific Islander
ID_TZ0004	70	Male	High school graduate			Married-civilian spouse present	White

is_hispanic	employment_commitment	unemployment_reason	employment_stat	wage_per_hour	is_labor_union	working_week_per_year	industry_code
All other	Not in labor force		0	0		52	0
All other	Children or Armed Forces		0	0		0	0
All other	Children or Armed Forces		0	500	No	15	41
All other	Children or Armed Forces		0	0		0	0
All other	Not in labor force		0	0		0	0

industry_code_main	occupation_code	occupation_code_main	total_employed	household_stat
Not in universe or children	0		2	Householder
Not in universe or children	0		0	Nonfamily householder
Hospital services	26	Adm support including clerical	4	Child 18+ never marr Not in a subfamily
Not in universe or children	0		0	Child <18 never marr not in subfamily
Not in universe or children	0		0	Spouse of householder

household_summary	under_18_family	veterans_admin_questionnaire	vet_benefit	tax_status	gains	losses	stocks_status	citizenship
Householder			2	Head of household	0	0	292	Native
Householder			2	Single	0	0	0	Native
Child 18 or older			2	Single	0	0	0	Native
Child under 18 never married	Both parents present		0	Nonfiler	0	0	0	Native
Spouse of householder			2	Joint both 65+	0	0	0	Native

mig_year	country_of_birth_own	country_of_birth_father	country_of_birth_mother	migration_code_change_in_msa	migration_prev_sunbelt
95	US	US	US	?	?
94	US	US	US	unchanged	
94	US	US	US	unchanged	
94	US	India	India	unchanged	
95	US	US	US	?	?

migration_code_move_within_reg	migration_code_change_in_reg	residence_1_year_ago	old_residence_reg	old_residence_state	importance_of_record
?	?				1779.74
unchanged	unchanged	Same			2366.75
unchanged	unchanged	Same			1693.42
unchanged	unchanged	Same			1380.27
?	?				1580.79

income_above_limit
Below limit
Below limit
Below limit
Below limit
Below limit