Introduction and Predictive Task:

Income is dependent on various factors. Some factors like education level and hours worked are on the surface more merit-based (though there is an argument to be made that some of these factors are largely influenced by underlying socioeconomic inequality). However, miscellaneous factors like the country resided in and given race may unfairly impact income beyond a level explained by random variation. Economists constantly debate over what factors explain the vast income inequality within the domestic United States and in the international context and what can be done to bridge the gap.

My goal is to create a predictive model that can accurately conclude whether a given demographic will have an income above or below $50,000 dollars. This model will allow economists to gauge whether more merit-based or more situational-based factors can explain income inequality. Understanding what largely determines income within a social group can help political scientists and other social analysts when examining long standing issues. Furthermore, I believe that on an individual level, understanding which factors we have control over are important can guide certain decisions like career choice and whether it's worth it to work for more hours per week.

Dataset Source:

Data from this project is sourced from the UCI Machine Learning Repository, providing various information about numerous factors that may impact income. The dataset will need some cleaning with categories like FinalWeight needed to be cut out because they serve no clear purpose in my analysis. Furthermore, I need to clean the income target bracket as there seems to be a typo that creates extra categories. I will need to combine categories like ">50K" & ">50K."

One issue that may erupt is whether factors like career and capital-gains are dependent on underlying socioeconomic factors like country or race. Also, since the target variable is categorical, we will not know the extent of inequality between groups. Regardless, I think we will get a model that can make accurate predictions based on the multitude of factors involved. However, the task will be in screening for correlation between the most important factors selected.

Also, another interesting point is the effect of national origin on the data. Since countries have different GDP levels and levels of economic development, inequality in the education and expertise of immigrants will inevitably skew highly. It may be further necessary to analyze groups originating from a single nation like the US to better understand what other factors drive inequality.

Data Summary:

The age range is 17 to 90. There are 9 working class categories. I assume that '?' indicates off-the-record employment. There are a number of education categories that reflect various levels. Interestingly, it even reflects the year that people dropped out of high school. There are 7 marital status categories that reflect the different legal arrangements undertaken. There are multiple career categories that reflect a good portion of white collar and blue collar jobs. The race categories are identical to those listed on the US Census. There are two sexes. The capital gains range is from $0 to $99,999. The capital loss is $0 to $4356. The hours worked per week range from 1 to 99 hours. There are a number of countries of origin including developed and developing nations. The target has two categories, ">50K" and "<=50K"

Data Description:

Pie Chart Representing Distribution of Those Making above 50K and Below 50K: Approximately 75% of the population has an income below 50K while 25% has one above it. This indicates a high degree of inequality. Still, we do not know the actual distribution of incomes since the data is categorical.

Histogram of Age: Age seems to be skewed moderately left with more people concentrated on the younger side. This is indicated by the graph and the fact that the mean age of 38.6 is greater than the median of 37.

Capital Gains Boxplot: This is very revealing. The capital gains realized by the majority of the population is virtually zero. I believe this is because only the extremely wealthy in developed countries saw realized capital gains during the 90s. Furthermore, this indicates that capital gains will not reveal much about inequality among all the social groups except for millionaires and the ultra wealthy.

Capital Loss Distribution Plot: Since only a very select demographic was invested in the stock market, it makes sense that capital-loss is largely irrelevant with most not realizing anything. This is very similar to capital gains

Histogram of Working Classes/Types: The majority of people seem to be privately employed. This checks out given that the vast majority of our economy is in private markets. We might see some disparities with those self-employed having higher distributions of >50K incomes. However, private employment makes up a significant portion of the workforce and includes a wide array of jobs from IT to C-Suite. This specific category might not reveal anything significantly discriminating because of its wide range of jobs, indicating that this might not be the best predictor in the model.

Barplot of Education Types: There seems to be a healthy mix of education levels. This could make education a very valuable predictor. The diversity of education levels makes sense, as various people in developed and developing countries pursue different degrees of higher education depending on their preferences.

Marital Status Chart and Countplot: There is also a healthy diversity of marriage types that could provide more utility for our predictive model. However, marriage status may be dependent on social situations.

Charts of Sex and Relationship Statuses: There is a skewed division between males and females. The two largest categories are husband and not-in-family. There is a great diversity of relationships to examine. Also the number of wives is very low compared to the ratio of females-to-men.

Boxplot of Hours Worked per Week: 50% of the working hours are concentrated between 40 and 45 hours per week. I believe the majority of working class people will be in this range, indicating that <50K income is concentrated in this range.

Barplot of Native Countries: There is a very healthy mix of countries, with not one country having a much larger count of samples than the others except for Guam and Portugal which should not have this huge of a representation given their population sizes.

Initial Visualisations:

Job Types for Each Income Bracket: These two bar graphs represent the counts of each working type in a sample that has only those earning below 50K and those earning above 50K. As expected, privately employed dominate both graphs as there are a wide range of jobs in that sector. Interestingly, the share of local government and self-employed people in the population of those making above 50K rises significantly.

Capital Gains for Hours Worked Per Week Depending On Income: This line plot relation hours per week and capital gains is interesting. While for the segment of those making greater than 50K, the capital gains spikes seem to be random and uniform regardless of hours worked per week, the spikes in the lower income bracket spike up at higher work hours. I believe that this indicates that these people are working hourly-wage jobs that directly increase their income with more hours, allowing them to invest more. As such, I believe that they are much closer to 50K than those not working as much. Afterwards, they are likely promoted to managerial roles which are more contract-based and don't reward hourly efforts, which makes the graph make sense.

Violinplot of Age Distribution Split by Income: There is a great deviation in age between those making less than 50K and those making more than 50K per year. This checks out as older generations have had more time to

store wealth and had less constraints due to the lower cost of living, meaning that less money had to go into rental payments and such.

Capital Loss Violin Plot Split by Income: There is a higher spread of capital loss for those making above 50K, but not by significantly more. Notably there is a spike with higher capital loss in this bracket. This may be attributed to investing schemes by the wealthier that result in loss.

Bar Graph of Marital Status: This graph makes a bar graph of the different marriage types split by income status. Those having civil marriages have a higher proportion of >50K income to <50K income, making it seem like relationship stability directly impacts wealth. Divorced couples and marriages with absent, separated, or widowed couples have a much higher proportion of those with less than 50K. Same for never married, as they likely chose to remain unmarried due to financial difficulty.

Occupation Bar Graphs: This graph of specific occupations shows that those with managerial, administrative, craft repair, and other jobs tend to make up the highest proportions of those making less than 50K. However, some of these jobs are also well represented in the more than 50K population such as executive, craft repair jobs, sales. Professional specialized jobs make up a noticeably higher proportion in the more than 50K income bracket. Sales and executive jobs being spread out evenly makes sense as they may be more commissions based and tied to performance as opposed to professionalized jobs which all give a set rate above 50K.

Proportion of High Income Residents for Each Immigrant Population: I made a dataframe of the proportion of those of each national origin making more than 50K per year compared to the entire population surveyed of each origin. There is a wide disparity with countries like Mexico and the DR having only 5% of those surveyed in the high income bracket while those of origin from India, France, and Taiwan have massive proportions of 40-42%. This makes sense as immigrants from the latter countries have taken managerial and technical roles while immigrants from generally poorer Latin American and Caribbean nations have primarily worked in lower-paying jobs less advanced in the value chain.

Racial Proportions Array: I made an array dividing the number of people in each race making >50K by the total population of each race. Evidently Asian groups and White groups have a significantly greater proportion making above 50K, indicating that race is a greater factor in predicting income than it should be. However, race is indirectly tied to both nation of origin and social strata within the US, due to the nation's long history of enforcing discrimination and disenfranchisement against non-White minorities. Thus race might not be a completely independent predictor.

Furthermore, I think it might be necessary to do this analysis for the racial makeup of each district in order to get a more accurate overview of racial inequality. For example, in the South there might be significant difference between African Americans and White Americans in the make-up of people making greater than 50K per year due to the long-standing effects of segregation and discrimination. However, in an already wealthy ethnically diverse area like Beverly Hills, racial inequality may be radically different, with African Americans and Hispanic upper classes earning high incomes.

Models and Methods: To predict income, I used multiple different categorical models to see which one performs the best in predicting these scores and accounting for the variation in my data. I decided to utilize an 80-20 train-test split for each model. The categorical models I used were Random Forest, Logistic Regression, and Decision Tree.

I evaluated the success of each of my models by comparing its performance metrics, against the baseline. To get my baseline value, I divided the proportion of those making >50K by the total population. That amounts to 0.24, or 24%. Thus, if one were to constantly guess that the samples were high income, he would most likely be right 24% of the time. Conversely, if one were to guess that the samples were low income, he would have an accuracy of 76%. This high value will be our baseline. In order to score the models, I evaluated their scores/accuracies dividing the number guessed correctly by the total number of test samples.

Decision Tree: I used a Decision Tree Classifier due to its ability to model non-linear relationships. Firstly, it distinguishes its predictions for each round based on predictors that it targets as the most impactful. This makes it very easy and intuitive to look into the decision making process and gauge which factors are relevant.

First I constructed a graph of Decision Tree Depth vs. Accuracy in order to find which depth was the optimal one to pursue. After that I constructed my decision tree with that level of depth and then displayed it. I also built a confusion matrix from predictions based on the test data to gauge its accuracy levels which I then scored according to the criteria mentioned earlier. I also extracted the most important variables.

Random Forest: I used the Random Forest Classifier model because it tends to improve the performance of decision trees by processing multiple of them. I built my forest and fit it onto the training data. Then I use GridSearchCV to scan for the best parameters/Decision Tree to then apply to my test data. I then extracted the best predictors and scored my model using a confusion matrix similar to the decision tree.

Logistic Regression: I used logistic regression because it applies the linear predictive method to classification problems, making it an interesting alternative to the Decision Tree and Random Forest. Its application was similar to Random Forest. I fit it onto the training data and then made a confusion matrix and scored it. I then extracted the most important variables.

Results and Interpretation:

Results of Decision Tree: Since test MSE is lowest at the 9th level, I would ideally have stuck with that for my classifier. However, it was impractical to fit 9 levels in my visualizations. I went with 6 as that is when the MSE stops decreasing at an extremely fast rate.

Train Score: 0.8548

Test Score: 0.8494

5 Most Important Factors: Married with Civilian Spouse, Capital Gain, Bachelors Education, Professional Speciality, Capital Loss

Results of Random Forest:

Train Score: 0.9740

Test Score: 0.8420

5 Most Important Factors: Married with Civilian Spouse, Capital Gain, hours-per-week, Relationship not in Family, Age

Results of Logistic Regression: Married Civilian Spouse, Bachelors, Capital Gain, Masters, Executive Managerial

Train Score: 0.8516

Test Score: 0.8486

From Consensus Analysis: Married with civilian spouse, capital gain, Bachelors seem to be the generally agreed upon deciding predictors. Two of them correspond with my analysis. Capital gain may be an accurate distinguisher in that all those with capital gains are likely high income individuals but those with capital gains make up a small portion of the data set. As such, it may not be distinguishing for a major part of the data.

Some of the least impactful factors were being Asian Pacific/Chinese and unmarried. The first doesn't make much sense as there is a significant skew in favor of Asians and Chinese in the data. Maybe there is some collinearity that makes the model reject this.

Conclusion and Next Steps:

In my analysis, all the models were significantly better than the baseline predictor, demonstrating their usefulness and effectiveness. In regards to test scores, they ranked in order from best to worst: Decision Tree, Logistic Regression, Random Forest. However, the differences are marginal at best.

Finding 1: The success of all the models indicate that regardless of the classifications being used, income can largely be attributed to certain factors.

Finding 2: Marriage, capital gains, and education are highly impactful due to their strong underlying correlation.

Finding 3: The relevance of certain factors like relationship, hours per week and executive vary based on the model indicating that while they're not as strong as the previous three, they are still strongly relevant.

Finding 4: The strong diminishment of test score in relation to train score indicates that some variation may be unpredictable for a demographic despite accounting for them in the train set. This shows that social scientists will

always be forced to reckon with the fact that income inequality may be dependent on randomness that you can't account for.

In conclusion, classification models were highly effective in explaining income inequality. The emphasis on marriage status in the model was very interesting and indicates that social stability is necessary for growth in income. The variation in findings like hours per week show that they don't account for as much as the core three do, which can spur more sociological research into these factors.

Improvements Needed and Further Research:

1) I think income needs to be represented quantitatively so that we may better gauge the extent of inequality. For example, capital gains was extremely misrepresented because only an extremely small percentage of even the above 50K group had capital gains. If the numbers were more quantitative, you would see this predictor be more useful to determine groups above 100K.

2) Distinguishing between different classes: I believe it is necessary to do so, because there is much inequality among different social groups dependent upon the community they come from and the background conditions they're subject to. Accounting for those differences may allow us to see the impact of factors we do have control over like hours worked and education (to a certain extent)

3) I want to see more factors like income taxes involved that more directly predict income. Furthermore, I want to analyze more indirect factors like police concentration and the urban/rural divide to give us a better picture of income inequality in America.

4) I am interested in adjusting my findings in relation to the cost of living of regions in America. Analyzing relative income may shed light on other factors being of importance for quality of life as opposed to just making a higher gross salary. Maybe an education in an urban expensive area is not as conducive to quality of life as hours worked on a farm in a cheaper area?