Best Machine Learning Models for Labor Data

DATS 6103 - Introduction to Data Mining

Spring 2020

Ignatios Draklellis

Cate Lee

Curtis Nguyen

Sanhanat Satetasakdasiri

Outline

Section 1: Introduction

Section 1.1 Project Objectives

Our analysis looks at labor demographic survey data in the United States from the 2019 Current Population Survey (CPS) of the United States Census. The objective of our research is to use several machine learning models to predict four key variables that we identified as wages, gender, employment, and geography. For each of these variables we construct five different machine learning models that are as follows: logistic regression, logistic regression with cross-validation, K Nearest Neighbors (KNN), Linear Support Vector Classifier (SVC), and Decision Trees. After interpreting the results of each we looked at the accuracy, precision, recall rate, F1 score, and runtime performance measurements to determine which model had the best fit for predicting the variable in question. We also examined the area under the curve (AUC) and receiver operating characteristics (ROC) to check the classification model's performance metric. We also present summary and descriptive statistics of the relevant variables, as well as several data visualization plots.

Section 1.2 About the Data

We used Current Population Survey data from the United States Census Bureau and Bureau of Labor Statistics from the year 2019. Our data came as a uniform Stata extract from the Center for Economic Policy and Research (CEPR)[1] website. The data set was downloaded as a .dta file with a total of 55,871 observations and 162 variables and primarily contains information on labor demographics for the US adult working population. However, we will choose only 21 variables with 55,871 observations in this analysis. The variables are as follows:

[1] http://ceprdata.org/cps-uniform-data-extracts/cps-outgoing-rotation-group/cps-org-data/?fbclid=IwAR2KaqSqeteagah3ojtcc3-NcgZxBSSRAc4n_prGpBD5KSDcPbHSYYFmM-k

- age - age (Numeric)
- female - sex (0 = male, 1 = female)
- wbho - Race (white, Hispanic, Black, Other)
- forborn - Foreign born (0 = foreign born, 1 = US born)
- citizen - US citizen (0 = No-US citizen, 1 = US citizen)
- vet - Veteran (0 = No-veteran, 1 = veteran)
- married - Married (0 = Never married, 1 = married)
- marstat - Marital status (Married, Never Married, Divorced, Widowed, Separated)
- ownchild - Number of children (Numeric)
- empl - Employed (0 = employed, 1 = unemployed)
- unem - Unemployed (0 = employed, 1 = unemployed)
- nilf - Not in labor force (0 = Not in labor force, 1 = in labor force)
- uncov - Union coverage (0 - non-Union coverage, 1 = Union coverage)
- state – US state (50 states)
- educ - Education level (HS, Some college, College, Advanced, LTHS)
- centcity - Central city (0 = no Central city, 1 = Central city)
- suburb - suburbs (0 = no suburbs area, 1 = suburbs area)
- rural - rural (0 = no rural area, 1 = rural area)
- hourslw - Hours last week, all jobs (Numeric)
- rw - Real hourly wage, 2019$ (Numeric)
- multjobn - Number of jobs (Numeric)

Section 2: Pre-Processing

Let's start with describing the original data as follows:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 55871 entries, 0 to 55870
Data columns (total 21 columns):
age          55871 non-null int64
female       55871 non-null int64
wbho         55871 non-null object
forborn      55871 non-null int64
citizen      55871 non-null int64
vet          54765 non-null float64
married      55871 non-null int64
marstat      55871 non-null object
ownchild     33693 non-null float64
empl         55657 non-null float64
unem         55657 non-null float64
nilf         55657 non-null float64
uncov        26467 non-null float64
multjobn     33130 non-null float64
state        55871 non-null object
centcity     55871 non-null int64
suburb       55871 non-null int64
rural        55871 non-null int64
educ         55871 non-null object
hourslw      32091 non-null float64
rw           29522 non-null float64
dtypes: float64(9), int64(8), object(4)
memory usage: 8.1+ MB
```

Moving onto summary statistics of the original data as follows:

	age	female	forborn	citizen	vet	\
count	55871.000000	55871.000000	55871.000000	55871.000000	54765.000000	
mean	48.343559		0.519447	0.135652	0.934367	0.078609
std	18.820813		0.499626	0.342421	0.247642	0.269129
min	16.000000		0.000000	0.000000	0.000000	0.000000
25%	32.000000		0.000000	0.000000	1.000000	0.000000
50%	49.000000		1.000000	0.000000	1.000000	0.000000
75%	63.000000		1.000000	0.000000	1.000000	0.000000
max	85.000000		1.000000	1.000000	1.000000	1.000000

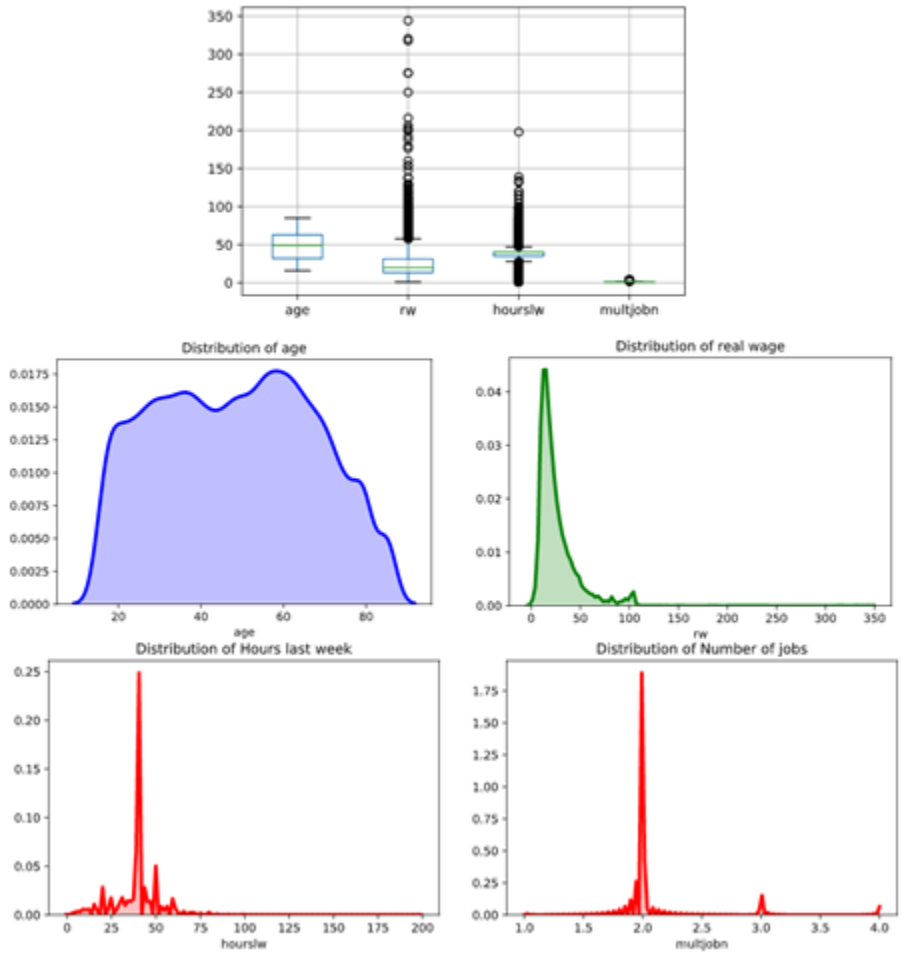
	married	ownchild	empl	unem	nilf	\
count	55871.000000	33693.000000	55657.000000	55657.000000	55657.000000	
mean	0.528682	0.719437	0.595253	0.023501	0.381246	
std	0.499181	1.089311	0.490847	0.151490	0.485697	
min	0.000000	0.000000	0.000000	0.000000	0.000000	
25%	0.000000	0.000000	0.000000	0.000000	0.000000	
50%	1.000000	0.000000	1.000000	0.000000	0.000000	
75%	1.000000	1.000000	1.000000	0.000000	1.000000	
max	1.000000	8.000000	1.000000	1.000000	1.000000	

	uncov	multjobn	centcity	suburb	rural	\
count	26467.000000	33130.000000	55871.000000	55871.000000	55871.000000	
mean	0.015415	1.058044	0.241198	0.404056	0.185391	
std	0.123200	0.260803	0.427814	0.490713	0.388618	
min	0.000000	1.000000	0.000000	0.000000	0.000000	
25%	0.000000	1.000000	0.000000	0.000000	0.000000	
50%	0.000000	1.000000	0.000000	0.000000	0.000000	
75%	0.000000		1.000000	0.000000	1.000000	0.000000
max	1.000000	4.000000	1.000000	1.000000	1.000000	

	hourslw	rw
count	32091.000000	29522.000000
mean	38.447135	25.880490
std	12.891361	19.789828
min	1.000000	1.000000
25%	35.000000	13.461500
50%	40.000000	19.830000
75%	40.000000	31.250000
max	198.000000	344.166660

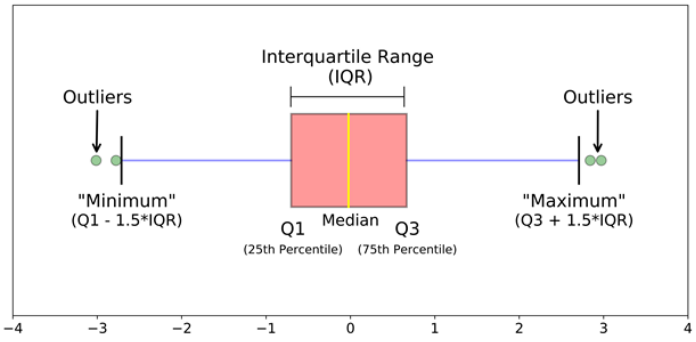
Section 2.1 Clean outliers

We start to identify outliers on numeric variables, including age, real wage (rw), hours worked last week (hourlw), and number of jobs (multjobn), using a box plot visualization and histogram as follow:



According to the plots, these four variables were not normal distributions because of the outliers, so we need to drop the outliers out of the dataset. Because some classifiers, such as logistic regression, require properties of a normal distribution. To make sure the outliers didn't skew our data we used the interquartile range (IQR) that is calculated as the difference between the 75th and 25th percentiles. It is represented by the formula IQR = Q3 − Q1. The rule of thumb is that anything

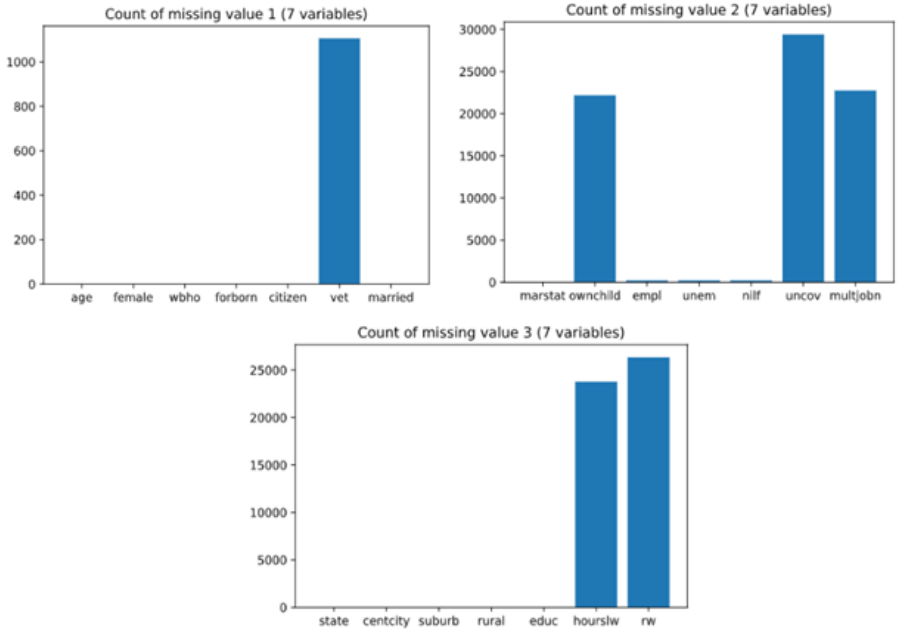
outside of the range of (Q1 - 1.5 IQR) and (Q3 + 1.5 IQR) is an outlier, and can thus be removed.



Section 2.2 Clean missing values (NA)

Next, we will check for missing values in each variable of the dataset.

age	0
female	0
wbho	0
forborn	0
citizen	0
vet	1106
married	0
marstat	0
ownchild	22178
empl	214
unem	214
nilf	214
uncov	29404
multjobn	22741
state	0
centcity	0
suburb	0
rural	0
educ	0
hourslw	23780
rw	26349
dtype: int64	



We then remove the missing values by using python's *drop.na()* function, and outliers in four numeric variables by using IQR scores. Then we will have a clean dataset to conduct the remainder of our analysis.

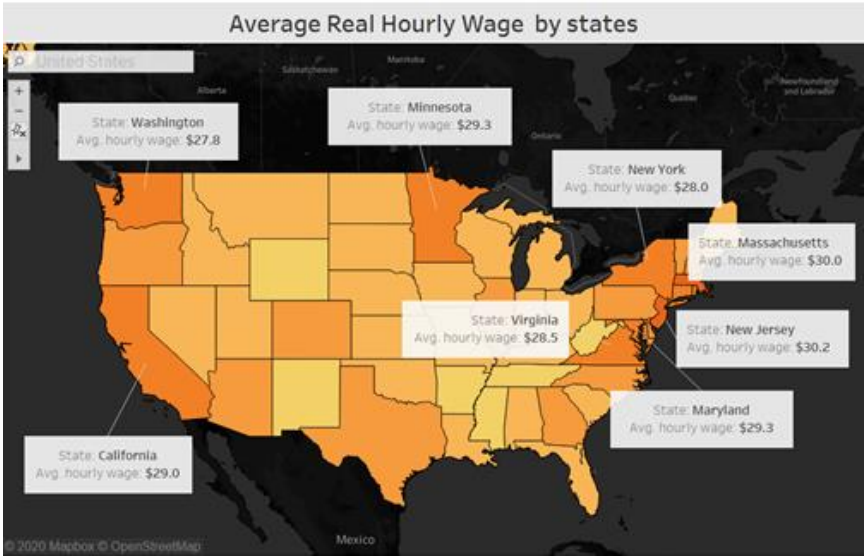
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10418 entries, 0 to 10417
Data columns (total 21 columns):
age         10418 non-null int64
female      10418 non-null int64
wbho        10418 non-null object
forborn     10418 non-null int64
citizen     10418 non-null int64
vet         10418 non-null float64
married     10418 non-null int64
marstat     10418 non-null object
ownchild    10418 non-null float64
empl        10418 non-null float64
unem        10418 non-null float64
nilf        10418 non-null float64
uncov       10418 non-null float64
multjobn    10418 non-null float64
state       10418 non-null object
centcity    10418 non-null int64
suburb      10418 non-null int64
rural       10418 non-null int64
educ        10418 non-null object
hourslw     10418 non-null float64
rw          10418 non-null float64
dtypes: float64(9), int64(8), object(4)
memory usage: 1.5+ MB
```

After preprocessing our dataset it contains 10,418 observations in total. We have cleaned the data and can now continue with an Exploratory Data Analysis (EDA).

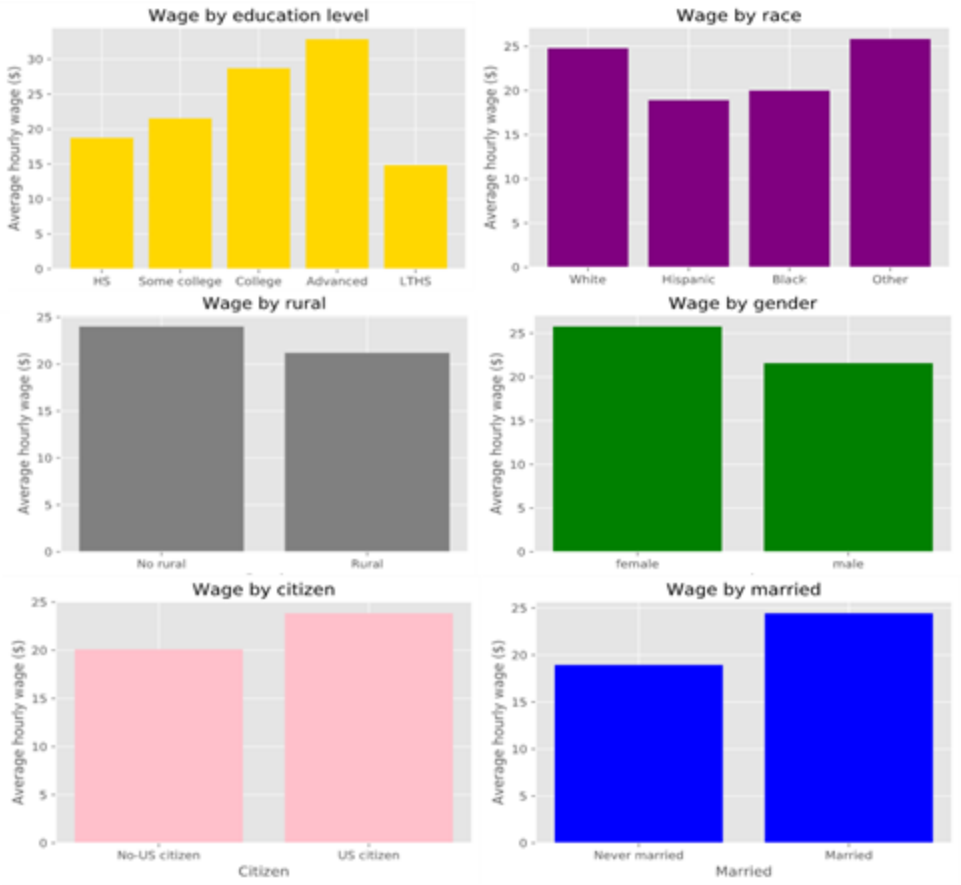
Section 3: Exploratory Data Analysis

Section 3.1 Wage and relevant variables

Let's start with average wages by US states. We look at New Jersey, Massachusetts, Maryland, Minnesota, and California in particular, and all have high hourly wages, while state's like Mississippi, Wyoming, West Virginia, and New Mexico have low hourly wages, according to the below map.

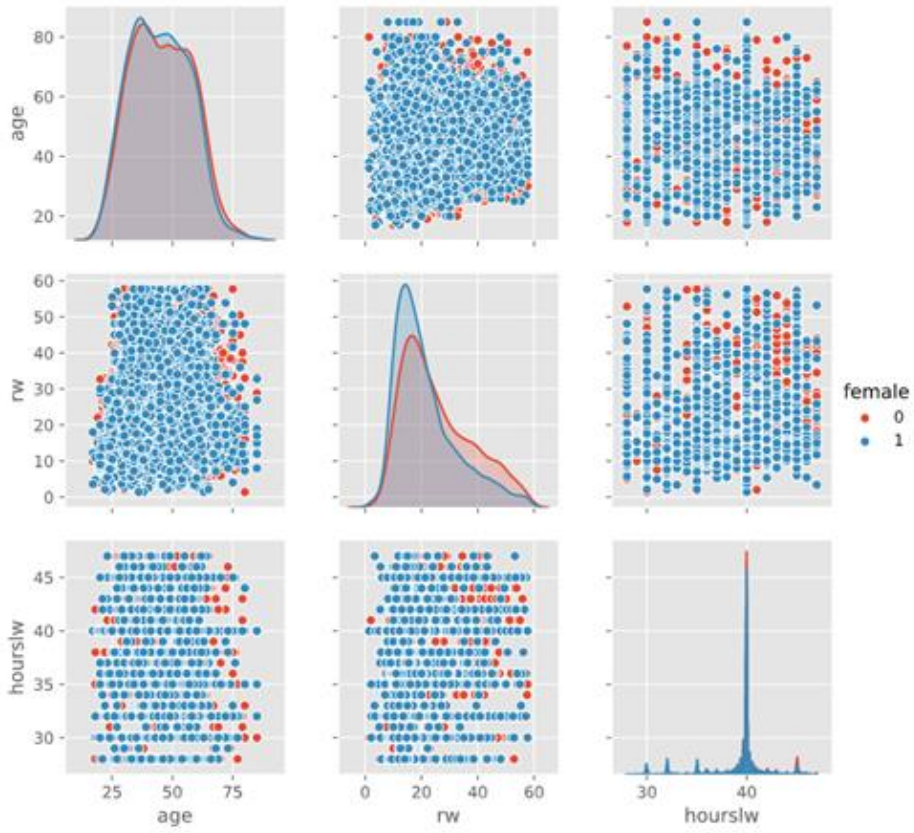


Next, we will compare the wage with relevant categorical variables, including education level, race, geography, gender, citizen, and married. The bar charts are plotted below with each variable. According to the plot, we found that advanced educational labor has the highest average hourly wage compared to the other levels, and white labor has the highest average hourly wage compared to the other races. Moving onto gender, marital, and citizens, we found that females have a higher average hourly wage than males. American citizens have higher average hourly wage than Non- American, and individuals who are married have a higher average hourly wage than non-married individuals. Lastly, labor in rural areas has a lower average wage than labor in non-rural areas.



Next, we looked at some scatter plots and histograms to explore relevant numeric variables, including real wage (rw), age, and hour last week (hourslw), in terms of gender. According to the histograms, the distributions of these three variables look relatively normal. However, according to the scatter plots, it is likely that there was no clear relationship between real wage and other numeric variables.

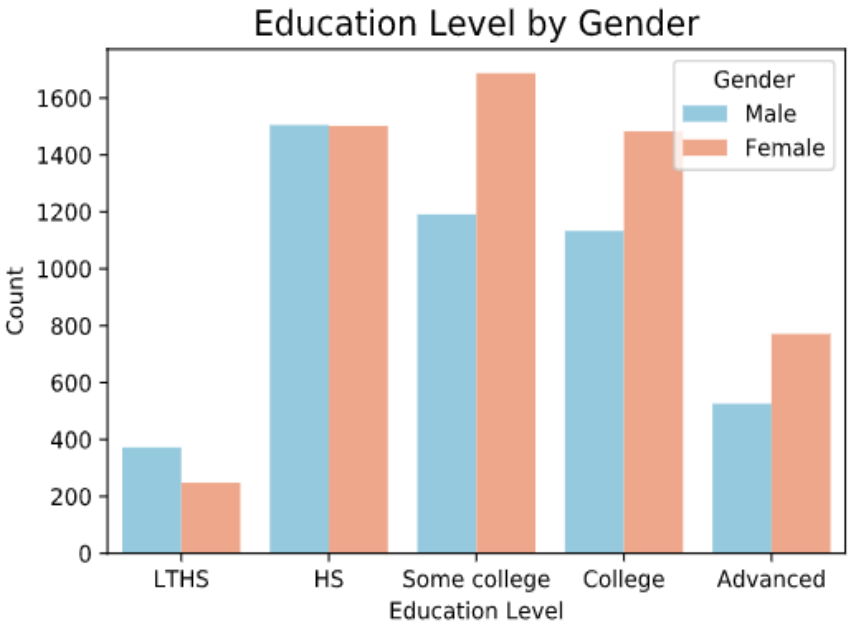
Although real wage is a numeric variable, we converted it to a binary variable in order to run the classifiers, which require a binary dependent variable. We divided wages into 2 groups by using the mean ($23.5 per hours). We consider low wage labor as the people who earn wages lower than the mean, while we consider high



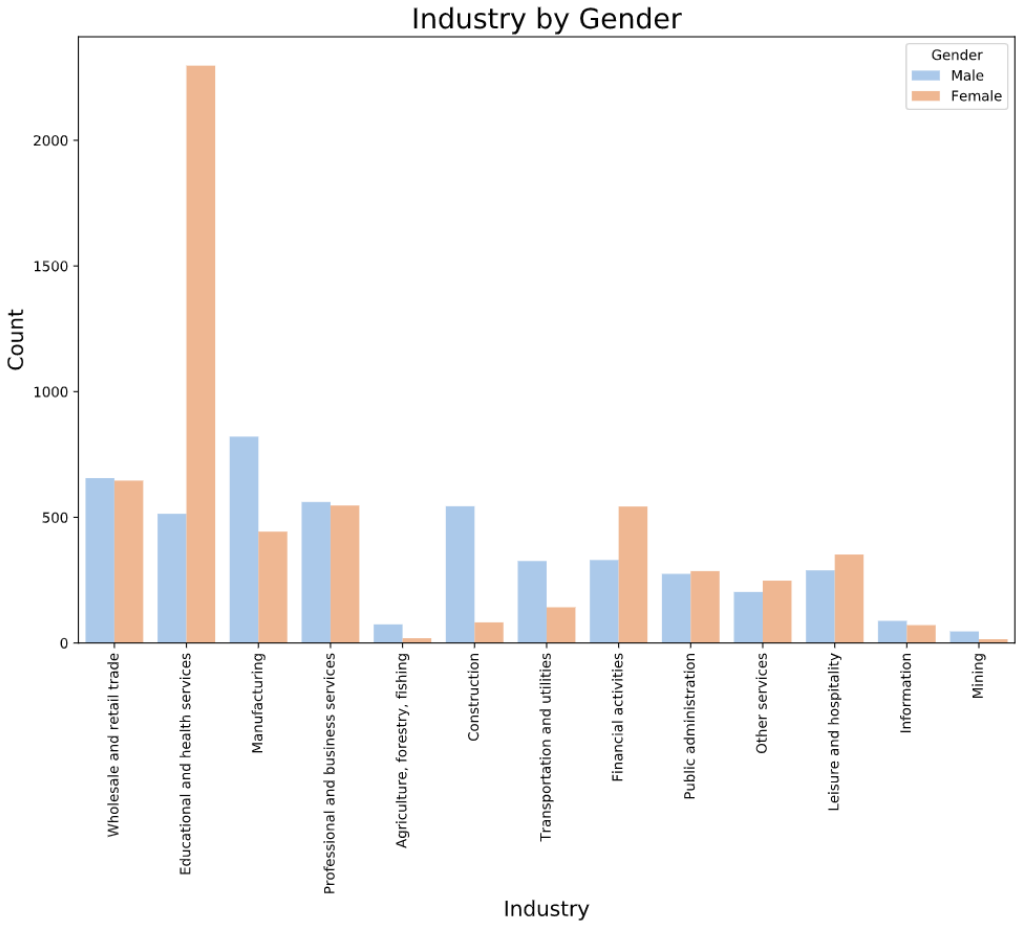
wage labor the individuals who earn wages higher than the mean. We can see that the number of low wage labor is greater than high wage labor.

Section 3.2 Gender and relevant variables

For the gender model, we considered how other attributes in the data set might be used to predict a participant's gender. Let's begin by examining the relationship between gender and some of the other variables in our dataset.

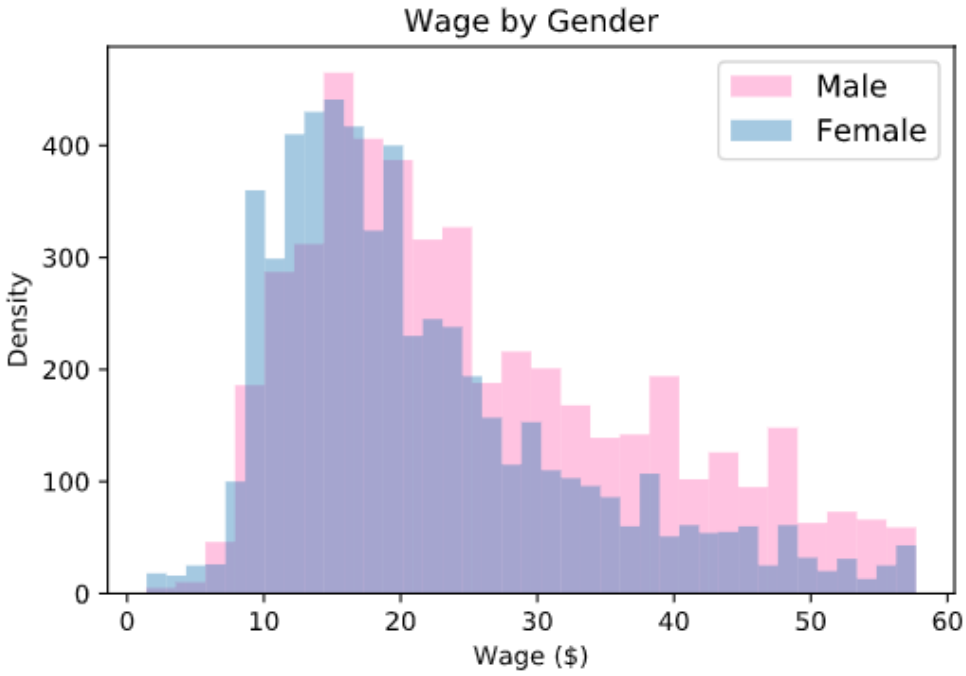


We can see that in our sample, females had slightly higher educational attainment than males at every education level. The most common education levels for females were ‘Some College’ or ‘College’. Now let’s look at the relationship between gender and industry.



We can see that there are large gender imbalances in some industries. Notably, ‘Education and health services’ has far more women than men. Industries like Construction and Manufacturing have more men than women in this sample.

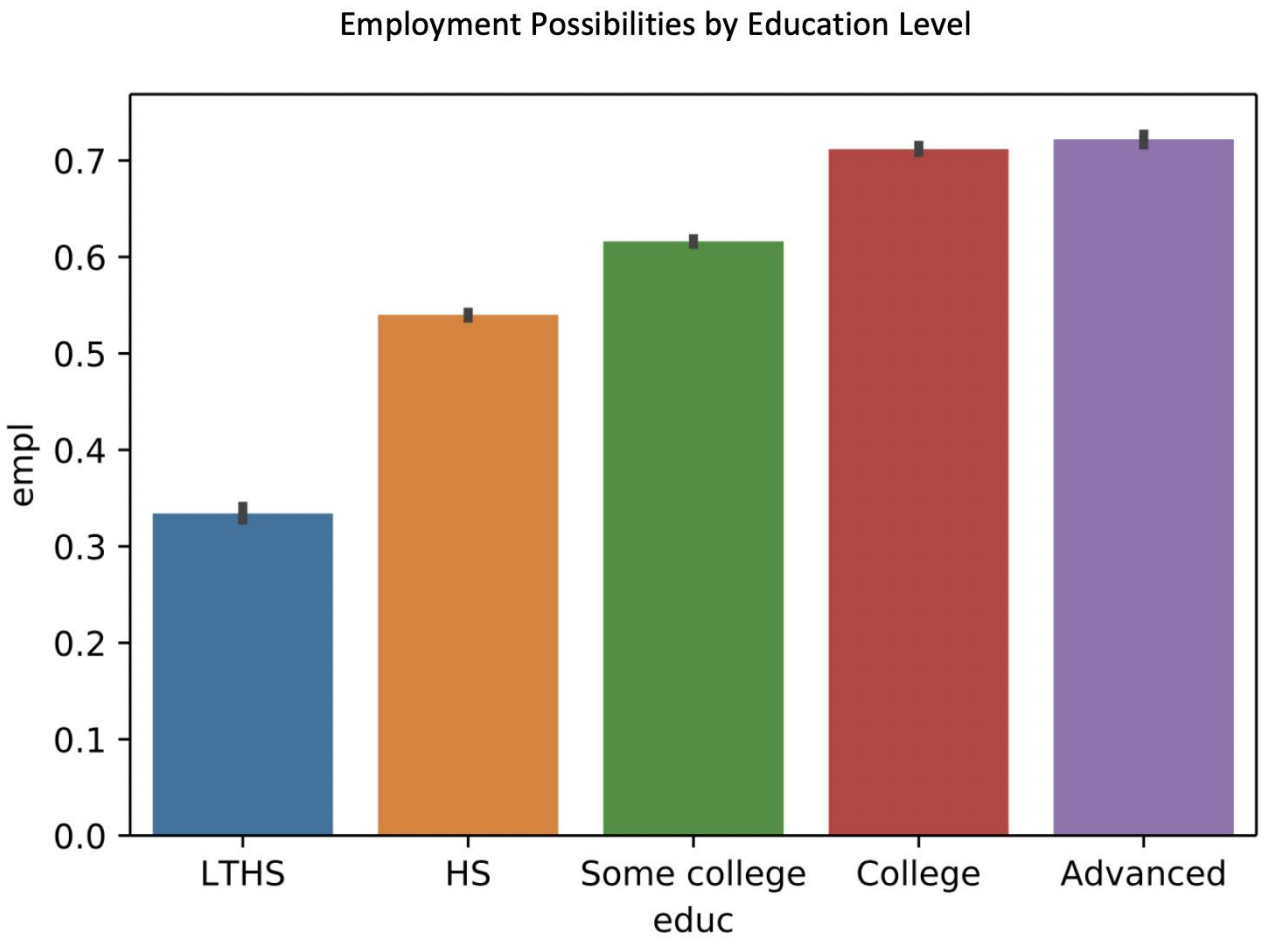
Let’s examine if there is any evidence of a gender wage gap in our sample.



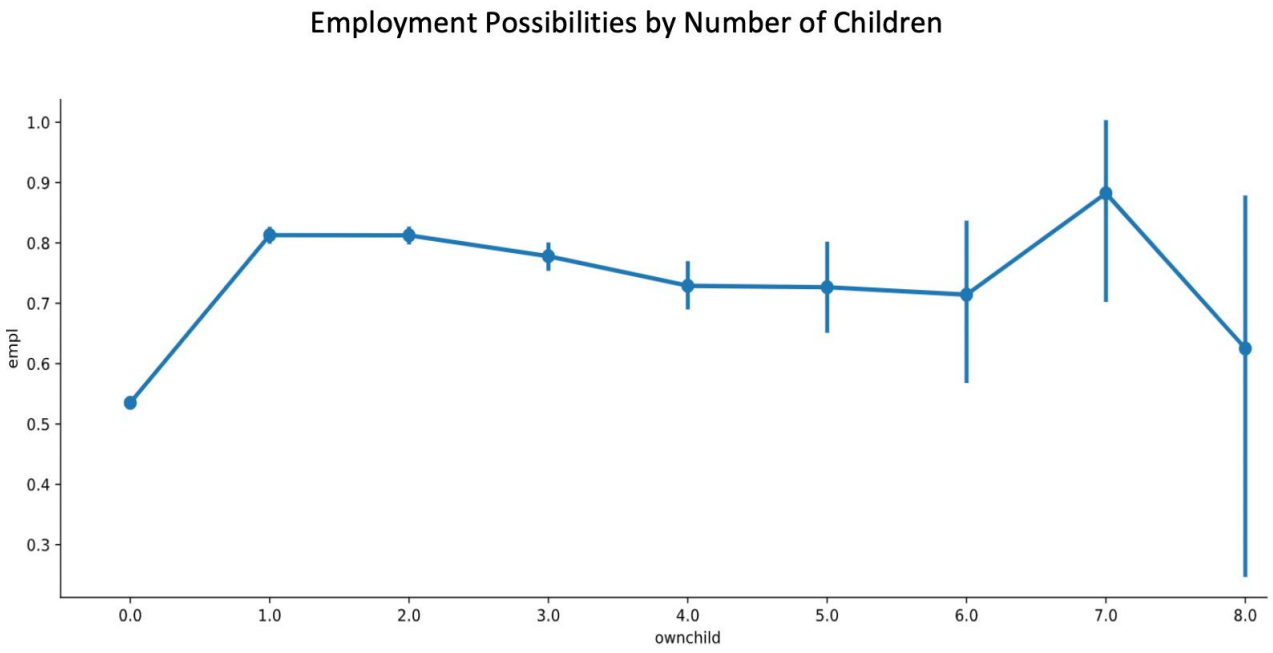
We can see that in our sample women had slightly lower wages overall, as there is only a small portion of their distribution above about $30 per hour, and a greater portion around $10 to $12 per hour.

Section 3.3 Employment and relevant variables

In order to predict a person’s employment status, we identified and incorporated relevant variables in our employment model. In the process of selecting suitable explanatory variables, we evaluated different logistic regression models using employment status as a binary and dependent variable. Ultimately, the selected independent variables are: age, gender, citizenship, marital status, education, race, number of children, and veteran status. These variables were found statistically significant in the regression models, hence deemed important in explaining an individual’s employment status. The following graphics demonstrate the relationships between several individual explanatory variables and the chance of being employed.



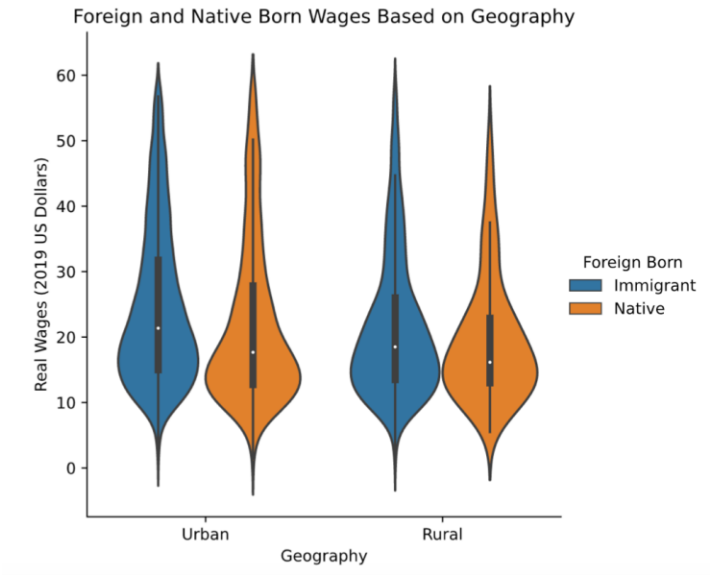
The bar chart illustrates the chance of being employed ranked by education levels. People with “less than high school” (LTHS) level of education have the lowest chance of being employed while people with advanced degrees have the highest chance of having a job. The graph shows a pattern that the more educated a person is, the more likely that person is employed.

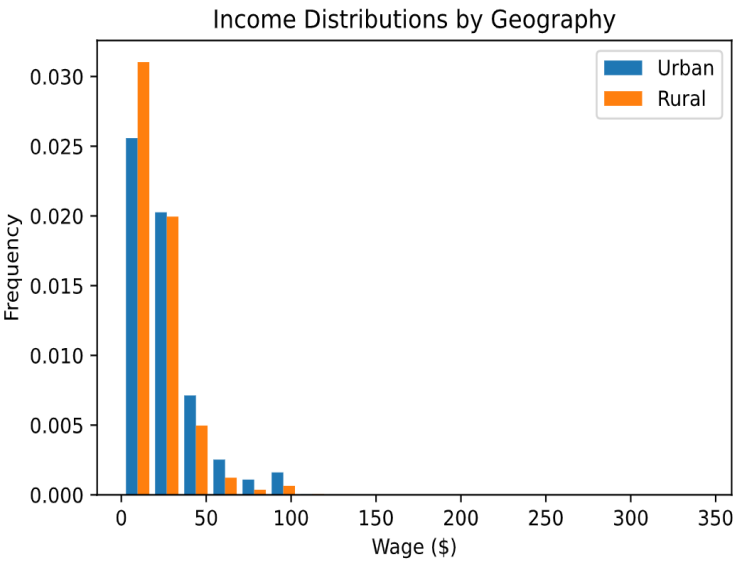


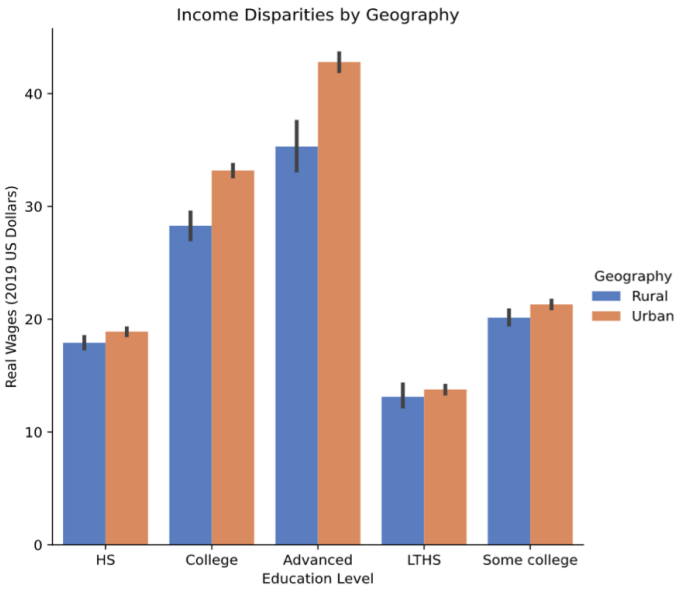
This graph highlights some interesting facts from our dataset, people with no children have the lowest chance of being employed while people with seven children surprisingly have the highest chance of being employed. This piece of information demonstrated in the graph is probably due to the fact there are not as many people with seven children as those that have three or fewer kids.

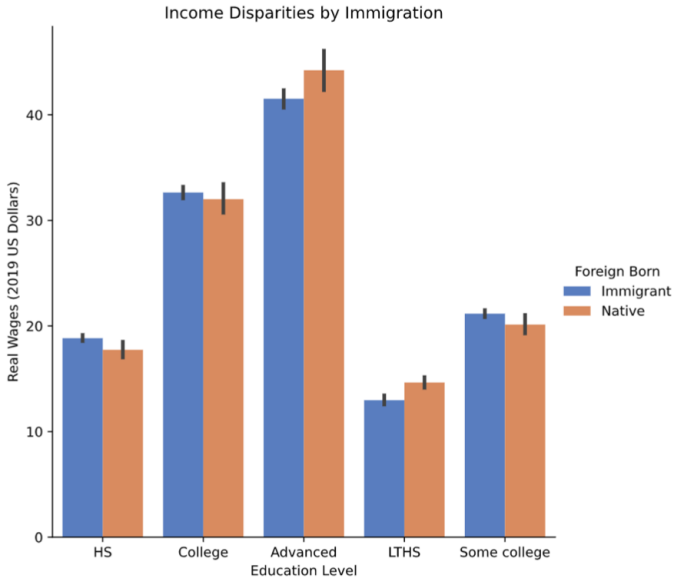
Section 3.4 Geography and relevant variables

For this portion we looked at several relevant variables that were useful in predicting an individual's geography. Geography is a binary variable that is divided into urban and rural. The relevant covariates we used are: age, wage, immigrant status, labor force presence, hours worked last week, and number of jobs. Below are some relevant plots of the underlying data.









We can see from the distributional plots above that urban residents earn higher wages than their rural counterparts in the United States. The bar plots also confirm that urban residents earn higher wages than rural dwellers. Using these variables we constructed the five aforementioned models to predict an individual's geography. The results of each are listed in the following section.

Section 4: Methodology (Model selection)

Our process for building and evaluating each of the four models is as follows:

- Run multiple linear regression to select relevant variables
- Check for multicollinearity with VIF
- Make a train-test split in 4:1 ratio
- Run the following models: Logistic Regression, Logistic Regression with CV, K-Nearest Neighbor, Linear SVC, Decision Tree
- Choose the best model based on our research question, variables, accuracy score, precision score, recall score, the speed of the model

Using multiple linear regression helped us to identify which variables would be important in each model. By following the same procedures in building the models, we could more easily compare the results of the four models. Once the models were built, we evaluated them on the same criteria to make a holistic choice for the best-fitting model.

Section 5: Modeling & Evaluation

Section 5.1 Wage prediction model results

We start to build a wage model with multiple linear regression. I run with 7 independent variables, including Age, gender, citizen, married, education, race, and rural. Then, all of them were significant. So, we move onto running six classifiers, including Logistic regression, Logistic regression with CV, K - Neighbors Regressor, SVC, and Decision Tree Classifier.

To determine variables, we put these 7 independent variables to be xtarget, and wage (binary variable)" in ytarget, then make a train-test split in 4:1 ratio in order to run the classifiers. The results show as follow:

	Logistic Regression	Logistic Regression with CV	K Nearest Neighbor (K = 8)	SVC (Linear)	Decision Tree (max depth = 5)
Accuracy	.633	.637	.586	.599	.643
Precision	.621	.624	.565	.359	.632
Recall	.633	.637	.586	.599	.643
F1 Score	.617	.616	.563	.449	.626
Run Time	3.61s	45.5s	2.54s	20.4s	185ms

Logistic Regression

Normally, the logistic regression was used to predict binary variables. In this case, the wage model performed well, with high accuracy, precision and recall scores. The scores look higher than K - Neighbors Regressor and linear kernel SVC. However, the logistic regression model has the run time slower than K - Neighbors Regressor and Decision Tree Classifier.

Logistic Regression with CV

Adding cross validation to the logistic regression slightly improved accuracy, precision and recall scores. It indicated that we got the wage model that performed better. However, this CV model took the longest time, compared to the others, to run the data. The model was increased run time from 3.61s (without CV) to 45.5s (with CV).

K - Neighbors Regressor

We tried to run the KNN model with different k-values from k=1 to k=11, we then found the model by using k=8 that performed the highest accuracy score. T 8-KNN model was the second top rank of run time. It was faster than Logistic Regression and SVC but slower than Decision Tree Classifier. However, the accuracy, precision and recall scores are not pretty good, compared to the others.

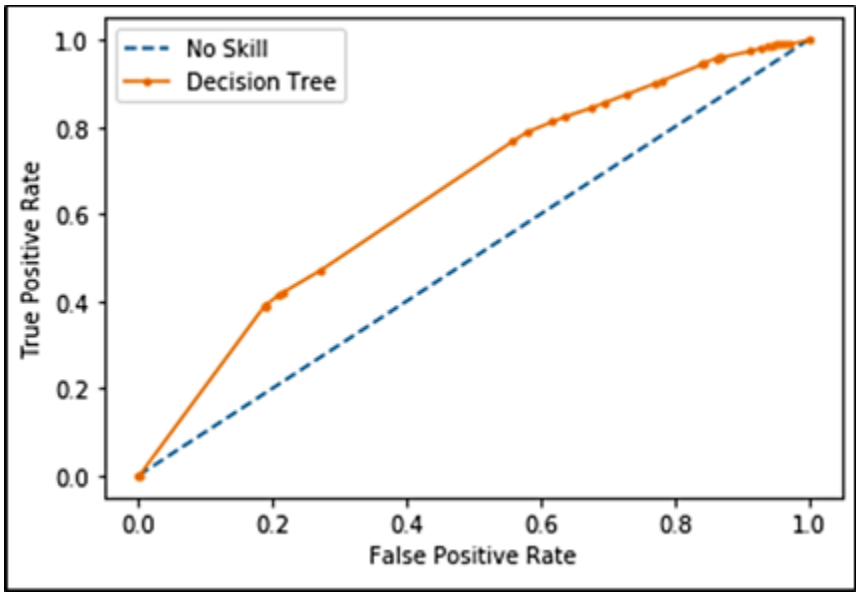
SVC

The linear kernel produced the best results for the support vector machine model for wage. The scores fell in this model, particularly precision score. Moreover, the run time was slow but faster than Logistic Regression with CV.

Decision Tree Classifier

To find out the best results for this model, we tested the model with different max depth and different parameters, including Gini, Entropy, and without both of them. We then got running the model with max depth of 5 produced the best results. The model performed well, with high accuracy, precision and recall scores, higher than the others. In addition, the run time of the model was the fastest with 185 microseconds per loop.

Model Choice- Decision Tree Classifier with max depth of 5



Taking into account all factors, the Decision Tree (max depth = 5) is the best model to predict wage because there are many supports. First of all, the model performed high accuracy, precision, recall and f1 scores, comparing with the other classifiers. Second, for the speed of the model running, the decision tree was the fastest with 185 microseconds per loop. AUC of the model was 0.647. It means there was about a 60% chance that the model will be able to distinguish between positive class and negative class. The ROC curve shows the trade-off of true positive rate and false positive rate, we can see that the true positive rate is steeper in the model. It indicates that it has a bit of skill.

Furthermore, there were some advantages for using decision tree models for prediction. First, the decision tree requires fewer data preprocessing from the user. Second, a decision tree is easy to interpret a complex model by using visualizations (like a tree) and explaining probabilities. Finally, the decision tree can easily capture Non-linear patterns, and doesn't require you to choose a kernel like in SVC.

Section 5.2 Gender prediction model results

	Logistic Regression	Logistic Regression with CV	K Nearest Neighbor (K=6)	SVC (linear)	Decision Tree (entropy, maxdepth=5)
Accuracy	.601	.603	.527	.566	.555
Precision	.602	.603	.583	.588	.601
Recall	.793	.792	.470	.686	.552
F1 Score	.684	.685	.521	.633	.575
Run Time	202 ms	2.96 s	54s	58.5 s	313 ms

Logistic Regression

The logistic regression was a natural choice for this variable because it's well suited to predicting binary variables. The model performed well, with high accuracy, precision and recall scores. It also had the shortest run time of all the models.

Logistic Regression with CV

Adding cross validation to the logistic regression decreased run time, as is expected because of the multiple samplings performed in cross validation. However, it added very slightly to the precision and recall of the model.

K - Neighbors Regressor

The KNN model had one of the slowest run times out of all the models, and somewhat lower precision, accuracy and recall scores. The recall score was particularly low, meaning this model had difficulty correctly identifying positives in the sample.

SVC

While the linear kernel produced the best results for the support vector machine model, the precision accuracy and recall scores all fell with this model. This model also had the longest run time of the five models.

Decision Tree Classifier

After testing the model with different parameters, using entropy with a max depth of 5 produced the best results. The model's accuracy, precision and recall scores were not superior to the other models, although the run time was very fast.

Model Choice-Logistic Regression with CV

Logistic regression with CV was the best choice of model for this variable, as it produced strong accuracy, precision and recall scores as well as a low run time. Logistic regression was a good fit for the binary variable, it allows a linear or non-linear decision boundary, and it has the advantage of giving probabilistic results. For a given observation the model produces a probability between 0 and 1 of the likelihood that participant is a female. Including cross validation improves accuracy slightly and gives more reliable results.

Section 5.3 Employment prediction model results

	Logistic Regression	Logistic Regression with CV	K Nearest Neighbor	SVC	Decision Tree
Accuracy	0.75	0.77	0.65	0.77	0.77
Precision	0.77	0.76	0.79	0.76	0.76

Recall	0.88	0.77	0.61	0.93	0.77
F1 Score	0.82	0.76	0.69	0.84	0.76
Run Time	1.42s	15.1s	1.28s	2min10s	183ms

Logistic Regression

The logistic regression model gives relatively high accuracy, precision, recall and F1 scores with a considerably quick runtime of only 1.42 sec.

Logistic Regression with CV

The logistic regression with cross-validation requires more runtime and reduces recall and F1 scores.

K - Nearest Neighbors

The KNN model has the quickest runtime but accuracy, recall and F1 scores are significantly lower compared to other models.

SVC

This model gives high accuracy, recall and F1 scores but runtime is extremely slow.

Decision Tree Classifier

The model provides relatively high accuracy and precision scores but lower recall and F1 scores compared to logistic regression and SVM.

Model Choice-Logistic Regression

Most models for employment provide similar and relatively high accuracy. While the logistic regression did not return the highest accuracy score, it was very close to the best accuracy score achieved by the tested models. The SVC model has high recall and F1 scores but its runtime is extremely slow. The logistic regression model has a high precision score, which measures the preciseness of values tested positive are actually positive. If a person is not employed (actual negative) is identified as employed, that would skew the prediction results. In addition, the model has a very high recall score, which demonstrates the rate of true positive compared to the actual positive values. If an employed person (actual positive) is predicted as unemployed (predicted negative), this will damage the prediction model greatly. The high precision and recall scores provided by the logistic regression imply such possibilities are less likely to happen. In addition, logistic regression allows a non-linear decision boundary and it is suitable for predicting a binary dependent variable such as employment status. All metrics considered, the logistic regression is the best predictive model for employment status.

Section 5.4 Geography prediction model results

	Logistic Regression	Logistic Regression with CV	K Nearest Neighbor	SVC	Decision Tree
Accuracy	.84	.84	.81	.84	.84
Precision	.71	.70	.75	.72	.71
Recall	.84	.84	.81	.84	.84
F1 Score	.77	.76	.77	.77	.77
Run Time	1.12s	13.1s	1.66s	22.4s	514 ms

Logistic Regression

By every performance measure the logit model compares relatively well and processes at one of the fastest run times.

Logistic Regression with CV

The addition of cross validation to the logit model increased the run time and did not have any beneficial implications for the other metrics.

K - Nearest Neighbors

The KNN returned an accuracy and recall rate that was lower than each of the other models. It's precision, F1 and run time were relatively good performance measures by comparison.

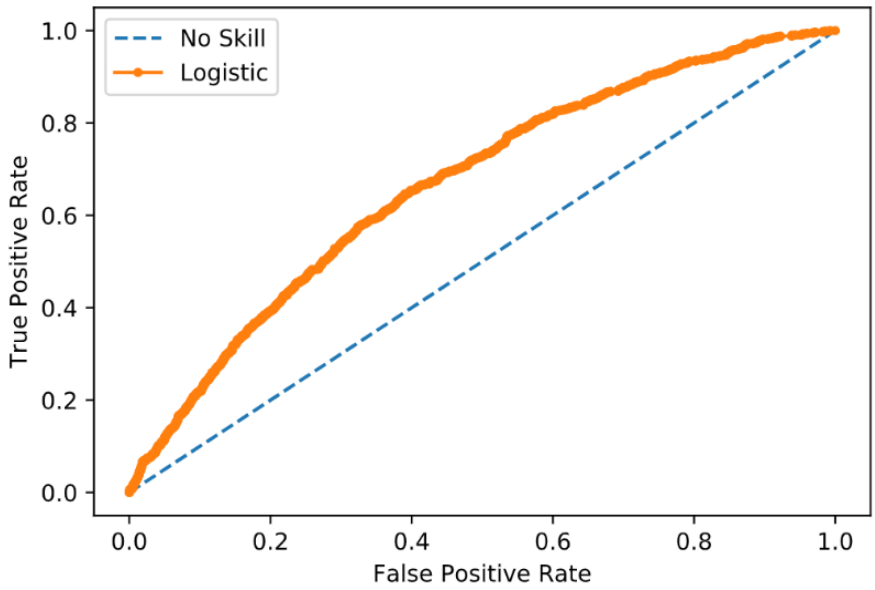
SVC

SVC retained all the same performance metrics as the logit model but required a longer runtime than the simpler binary logistic regression.

Decision Tree Classifier

The decision trees performed well and had the same performance metrics as the logit model with a shorter run time. However, decision tree models are easy to overfit in the learning process which is not useful for generalizing the data, and can have an unstable variance. For these reasons and more we chose the simpler binary logit over the decision tree model.

Model Choice-Logistic Regression with CV



Considering the metrics above, the model that we felt was best able to predict geography was the logistic regression. Logit models also provide coefficients in terms of probability which allow us to interpret the results with ease, and there is a low risk of overfitting the model. Logistic regressions also retain the benefit of easy updating with new data, and the sigmoid function allows logits to work with linear and non-linear problems.

Section 6: Conclusion

Overall, we considered our variable choice, desired interpretation, scores and run times to choose the best models for labor data. This taught us about the strengths of individual models, and how to compare models to find the best fit for a specific research question.

We found that logistic regression performs well on binary variables and is flexible to different labor variables. It was the best fitting model for predicting employment, gender and geographic region. It's main strengths for predicting labor variables are that it allows for a non-linear decision boundary and gives a probabilistic interpretation that is useful for reporting its results. However, decision tree models can be preferable on balanced data where there are many non-linear patterns, as it was the best model choice for predicting wage. We think that these findings would be useful to economists and data scientists who are interested in selecting a strong model to address questions about labor demographics.