Nuray Ozden (nyo3)
Rahma Tasnim (rt429)

Women's Professional Standing: Income Over Time

Research Question

Our main area of research is into women's professional standings after having children. More specifically, we want to provide analytically explore these sub-genres:

- Can we predict if a person will go on maternity or paternity leave?
- How does a man's professional standing compare to a woman's after having children?
- What is the probability a woman will return to the same (or higher) professional standing after having children?

Background Information

In the 1970s, the proportion of women with higher education degrees to men with higher education degrees increased until flatlining in the 2010s. Along with this, the proportion of women employed to men employed has steadily increased from the 1970s but started flatlining and even started to slightly decrease since the the 2000s. Despite being more educated, women are having a difficult time entering the workforce. This issue is exacerbated when women decide to have children with this idea of the "motherhood penalty" where women in the workforce are discriminated against because of the fact that they are mothers.

Both being women in STEM, this topic is very important to us because it provides insight into how the workplace currently operates around women who choose to have children; or the lack thereof. This analysis will be extremely valuable for any young adult as it can help them determine the current reality of having children while working, and ways to adjust their lifestyle accordingly.

Dataset Description

We will be utilizing the data from IPUMS CPS, which supplies census and survey data on a number of demographics (based on gender, socioeconomic status, veteran status, race, etc.) and different sanctions of life, such as economic development, household structure, food security, and more. We have chosen a dataset that captures the information on both men and women with children, including features such as, but not limited to, income, occupation, and the number of children. The dataset will also have information on men as well, as a comparison measure for certain models. We will use this data to gain insight into women's professional lives before and after having children, then see how this differs from men's professional lives before and after having children.

Decision Tree to Predict if Employees go on Maternity/Paternity Leave

Overview

Nuray Ozden (nyo3)
Rahma Tasnim (rt429)

To better understand our main target variable for this paper, which is whether women go on maternity leave or not, we decided to unwrap this variable to understand what other features help predict whether people go on parental leave. We decided to look at both maternity and paternity leave in order to understand the concept as a whole before unpacking their differences, which we explore below. Doing so will allow us to understand what other variables are correlated with maternity/paternity leave, and in theory, people able to predict whether people will go on leave based on a few attributes. We have constructed a Classification Decision Tree to predict if an individual will go on maternity/paternity leave.

Raw Data

For this decision tree, the data extrapolated is from the IPUMS database. The features taken from the large database are **`Year`**, **`Age`**, **`Sex`**, **`IncWage`**, and **`Health`** with the target variable being **`WhyAbsnt`**, which represents the reason why a person left work. These are the features chosen for this model because they are attributes of people that could have an influence over whether they go on maternity/paternity leave or not. Some of the attributes have less of an influence and some have more, which we will explore further later on. The data is taken from the US from 2010 to 2022, to look at the most recent data since the average income in the US has increased over the past few decades. Additionally, all of the data is for individuals who are employed (so cleaning out the people who are unemployed) to only look at people who *had* work to leave before going on maternity/paternity leave. Before data cleaning, there were 2,433,005 data points in the dataframe.

Cleaning

The cleaning applied to the dataset dramatically decreased the amount of data used in the model. First, the **`IncWage`** column had NaN, 0, and NUI (not in universe) values, which resulted in the entire row being removed; this was chosen over replacing with the average or adding a Boolean indicator, for example, because the rows that had one of these values usually also had missing data in other columns. Next, only the **`Age`** between 25 and 65 were chosen for the dataset because these are the “working ages”. In the **`WhyAbsnt`** feature, there were some NUI values, so those rows were removed as well. Lastly, all NaN in all other columns were dropped.

Feature Transformations

The feature transformations allow the data to be read by the model to have a successful output. The first feature transformation was to convert a categorical variable into numerical values for the **`Sex`** feature, transforming ‘Male’ to 0 and ‘Female’ to 1. Next, we converted the ordinal variable **`Health`** into numerical values as well; originally, it gave a health score of 'excellent', 'very good', 'good', 'fair', or 'poor'. This was translated to a scale from 1 to 5, where ‘excellent’ is 5 and ‘poor’ is 1. Lastly, the target variable **`WhyAbsnt`** was transformed in order to only look at the people who went on maternity/paternity leave. The possible values for this

categorical variable were 'maternity/paternity leave', 'vacation/personal days', 'weather affected job', 'other family/personal obligation', 'school/training', 'child care problems', 'labor dispute', 'own illness/injury/medical problems', 'civic/military duty', and 'other'. We thus converted the 'maternity/paternity leave' option to the value 1 and the other options to value 0 in order to have a classification label for our Decision Tree. While we considered also assigning the 'child care problems' value to a 1, this does not equal maternity/paternity leave in all cases; we only want to analyze parents leaving while their child is an infant. This is an overview of the cleaned and transformed input features fed to the model: there are 24,538 data points.

	Year	**Age**	**Sex**	**Health**	**IncWage**
std	3.67922145291938	11.250190924437446	0.4943922647094189 4	1.009618413744217	73275.99663722566
min	2010.0	25.0	0.0	1.0	7.0
mean	2015.7868611948813	44.23779444127476	0.5747412177031543	3.7868611948814084	53748.89489770967
max	2022.0	65.0	1.0	5.0	1749999.0
count	24538.0	24538.0	24538.0	24538.0	24538.0
75%	2019.0	54.0	1.0	5.0	65000.0
50%	2016.0	44.0	1.0	4.0	40000.0
25%	2013.0	35.0	0.0	3.0	21000.0

Table 1: Summary of Transformed Input Features

Models

To use this data to predict whether an individual goes on maternity/paternity leave, we decided that a Classification Decision Tree using an ensemble method was the best tactic, where the label is "Went on Maternity/Paternity leave" or "Did not go on Maternity/Paternity leave", extracted from the **`WhyAbsnt`** column. A Decision Tree was chosen as it is an excellent model for non-linear data, it is relatively fast to compute, it can tell us which features are the most important, and it can help identify relationships between variables. A Bagging algorithm was chosen to create an "aggregate tree" because it creates a strong, more accurate model, reduces probability of the tree overfitting, and it generalizes well over new data. The chosen number of trees to average over was 10. For the rest of the section, the resulting tree from the 10 weak-learners will be referred to as the "aggregate tree." The train/test split for all of the models was 80% and 20%.

First, an aggregate tree was formed over all of the features (**`Year`**, **`Age`**, **`Sex`**, **`Health`**, and **`IncWage`**) with no limit on the maximum depth of the tree. The average accuracy using the train/test split specified above was **0.88**. Then, the classifier was evaluated using cross-validation, which was done 10-fold. The accuracy score was: **0.87 (+/- 0.05)**. Here is a glimpse of the first three layers of the aggregate tree, as well as the feature importances:

Nuray Ozden (nyo3)
Rahma Tasnim (rt429)

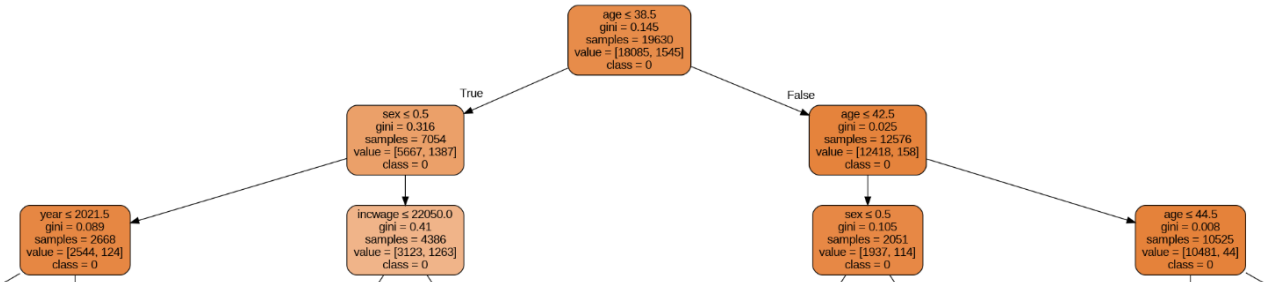

Figure 1: First Three Layers of Decision Tree

Additionally, to analyze the important features, we used the `feature_importances_` attribute of the DecisionTree in the Scikit package that indicates which features carried the most weight. As shown by the table, **`Year`** is the least indicative of whether a person goes on maternity/paternity leave, whereas **`Age`** is the most indicative.

Feature Name	Weight
`Year`	0.006
`Age`	0.516
`Sex`	0.323
`IncWage`	0.105
`Health`	0.049

Table 2: Weights of Features

A high accuracy and cross-validation score can oftentimes be indicative of overfitting, so some adjustments were made to the model and it was created again. In this instance, we limited the maximum depth of the tree to be four layers (which we discovered to be the optimal depth); the Bagging algorithm was still utilized for these new weak-learners. The resulting accuracy score was **0.92** and cross-validation score was **0.90**. An increase in both the accuracy and cross-validation score indicates that the previous tree, which had no limit on the depth, was overfitting to the data more than the aggregated tree with four layers.

However, just to ensure that the decision tree was not overfitting, we decided to compute a covariance matrix to expose any undercover correlations between features. The covariance matrix (not shown) revealed that the covariance between **`IncWage`** and **`Year`** was 19162, meaning there is a strong linear relationship between the two features and we do not need both in our Decision Tree. Thus, we re-performed the Decision Tree without the **`Year`** feature, but still with a maximum depth of 4 and using the Bagging algorithm to form our aggregate tree. Our resulting accuracy was **0.91** and our cross-validation score was **0.92**. These increased accuracy and cross-validation scores indicate that the previous model may have been overfitting, and this one is more generalized. The new importances of the features are shown in the table below, which, as you can see, are very similar to the previous model's importances.

Overall, the Classification Decision Tree with the Bagging algorithm was a successful model on the data that accurately classifies whether a person goes on maternity/paternity leave or not 91% of the time. The features chosen were good features to choose for this model, and we can see that the **`Age`** feature has the most influence on the label, whereas **`Health`** has the least

influence. Intuitively, this makes a lot of sense. This exploration of the Decision Tree taught us about the bias/variance tradeoff, and how to improve a model so it does not overfit to the training data, and what it means to choose valuable features. To improve this model for the future, we would explore other features in the IPUMS database that may be good decision features to predict the label.

Feature Name	Weight
Age	0.519
Sex	0.324
IncWage	0.104
Health	0.052

Table 3: Weights of Features

Linear Regression Model to Compare Professional Standing

Raw Data

The features taken for this dataset include: **Year, Age, Sex, Marst, Nchild, EmpStat, WhyAbsnt, WkStat, Educ**, and **IncWage** where **ncwage** is the target variable. These features were chosen to see how they would work together to impact the income of an individual in the workforce. Each of these features have categories that would indicate more information than actually needed for this regression, so some feature transformations needed to be done.

Cleaning + Feature Transformation

Because all of the features were categorical features, we transformed them into one-hot features with the exception of education. The feature transformations look like this **Sex**: 1 for female, 0 for male, **Marst**: 1 for married, 0 otherwise, **Nchild**: 1 for having at least 1 child, 0 otherwise, **EmptStat**: 1 for it they are currently at work, 0 otherwise, **Wkstat**: 1 for full-time, 0 otherwise. Instead of a one-hot feature, **Educ** became an ordinal variable where 4 is for graduate degrees or higher, 3 is for undergraduate degrees, 2 is for some college, 1 is for highschool, and 0 is for less than highschool. We also removed any NIU and NA values along with any major outliers.

Model

To help decide which features to use for the model, we created pairplots of the raw and transformed data. The raw data did not show any visible relationships. After we transform our data, we create another pairplot. We find that this also did not show any visible relationships between the features.

Nuray Ozden (nyo3)
Rahma Tasnim (rt429)

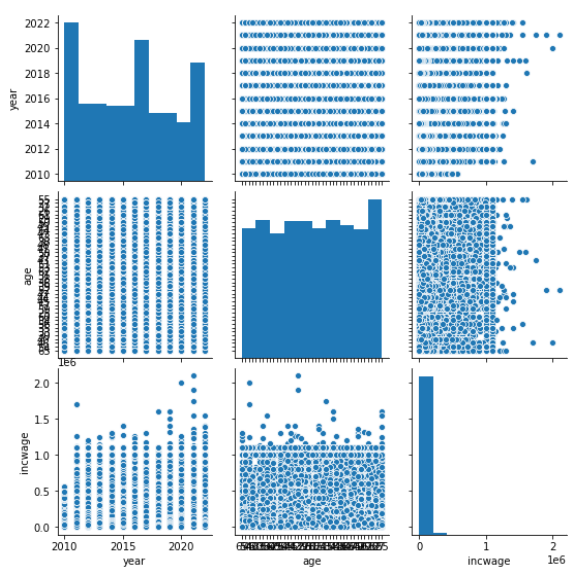

Figure 2: Seaborn Pairplot of Raw Data

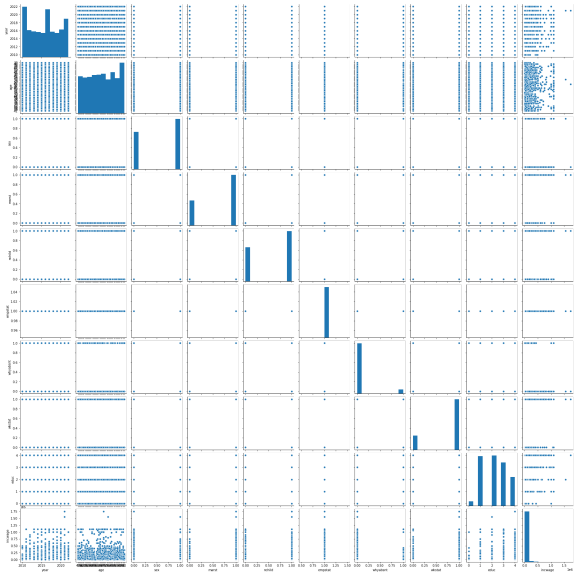

Figure 3: Seaborn Pairplot of Transformed Data

In this model, we take features that seem to impact an individual's standing at work. These features include: **`Sex`**, **`Marst`**: marital status, **`Nchild`**: the number of children the individual has, **`EmptStat`**: if they are actually at work, **`WhyAbsnt`**: if they are on maternity or paternity leave, **`WkStat`**: if they are working full time, and **`Educ`**: their education level. We find that all these features are significant with the exception of **`EmpStat`** and **`WhyAbsnt`**, when looking at their p values. After looking at the evaluative measures of the model we find that the R^2 value is relatively low and the mean squared error is relatively high for both the validation and test sets.

R^2 on validation set:	0.11054904176524027
MSE on validation set:	5480209430.126057
R^2 on test set:	0.13455379298933767
MSE on test set:	3879393279.51452

Table 4: Evaluative Measures of Linear Regression Model

This indicates that the model is not doing a good job predicting the target variable, **`IncWage`**. Because of this, we decided to add a Ridge Regression to improve the model. However, we find that after adding the ridge, the evaluative stay relatively the same:

Ridge R^2 on validation set:	0.11054882426700263
Ridge MSE on validation set:	5480210770.206567
Ridge R^2 on test set:	0.13455409803207996

Ridge MSE on test set:	3879391912.1496224

Table 5: Evaluative Measures of Ridge Regression Model

We decided to remove the non-significant features **`EmpStat`** and **`WhyAbsnt`** from the linear regression model to see if that would improve the evaluative measures. However, this actually worsens the model, reducing the R^2 value even more and increasing the mean squared error. As a one last ditch effort to improve the model, we look at the very bare bone features that should impact the income of an individual in the workforce. We look at only the **`Sex`**, **`Nchild`**, and **`Educ`** feature to see how they work together to predict **`IncWage`**. This combination of features improved the model, however, it is still a bad predictor of an individuals' income.

R^2 on validation set:	0.1451579643882177
MSE on validation set:	3582515769.0021124
R^2 on test set:	0.1451993131278917
MSE on test set:	3485653837.0642095

Table 6: Evaluative Measures of Updated Linear Regression Model

After looking at the OLS Regression results, we find that the coefficient for **`Sex`** shows that men have higher salaries than women even after accounting for the effect of number of children and education. The negative sign of the **`Sex`** coefficient shows that women earn less than men. Interestingly enough, we see that the coefficient for **`Educ`** shows that women have higher education attainment than men do, even with children. Despite being more educated than men, women are less monetarily compensated.

Logistic Regression Model for Professional Standing Probability

Data

For the logistic regression model, we used the same data set as used for the linear regression model as it had already been cleaned and transformed into one-hot features but we added a new feature **`Occly`** which has codes for different occupations into the dataset as well. To make the analysis easier, we looked at occupations that would be considered as data analyst or operations research analysts as it is relevant to our career interests. This would also make it easier for a logistic regression model to be run on the data considering that we want to keep characteristics the same to address our question. We split the data into a training, validation, and test set.

Model

To find the probability that a woman will return to the same (or higher) professional standing after having children, we run a logistic regression model using the same data set as the linear regression model with features **`Year, Age, Sex, Nchild, EmpStat,`**

Nuray Ozden (nyo3)
Rahma Tasnim (rt429)

`WkStat,WhyAbsnt` to predict the target variable **`IncWage`** as our standard of professional standing. We needed to take a subset of the training data to run the logistic regression on because the model kept timing out every time we ran the fitting. We ended up getting an accuracy score of **0.034**. This suggests that it is not a good predictor of a woman's probability of returning to the same (or higher) professional standing after having children. In fact, an accuracy this low indicates that the model is predicting at random. To mitigate this, we decided to use the new dataset including the **`Occly`** feature and then also decided to create a binary variable called **`Returned`** indicating whether the woman returned to the same or higher professional standing after having children using this data to improve the model. We created this variable by comparing the income of a woman after having children to the mean income of women in similar education, marital status, and number of children. This way, we can accurately assume that women who earn as much or more than the mean income of their peers with similar characteristics have returned to the same or higher professional standing.

We then train the logistic regression model using this **`Returned`** feature along with all the other features in the new dataset. We evaluate the model using accuracy, precision, recall, and F1. The high accuracy score of **0.74** suggests that the model is effective in predicting whether a woman returned to the same or higher professional standing after having children. The low recall score of **0.46** suggests that the model may have difficulty identifying all cases of women who returned to the same or higher professional standing. The precision score of **0.66** shows that among the cases that the model identifies as returning to the same or higher professional standing, 66.1% are actually correctly predicted. The F1 score of **0.54** is a weighted average of the precision and recall scores, showing that the model is predicting the target variable **`IncWage`** moderately.

Conclusion

We used many of the techniques and findings from ORIE 4741 to create this project and extract meaning from our dataset about the professional standings of women who went on maternity leave. We were able to learn a lot about how to create not just any model with any features, but how to be selective and calculative in creating a meaningful model that gave us the insight we were searching for. For example, we learned about which characteristics were the most indicative of a person going on maternity/paternity leave using the Decision Tree, the relationship between Sex and salary from the Linear Regression, and how to combine features to predict a women's professional standing using Logistic Regression.

For future exploration, we would like to expand the dataset to include information from other countries, to see if we can run the same models above but compare across the world. For the Decision Tree Classifier, an extension would be to use a Random Forest Algorithm to explore the features more in depth, and see if any overfitting is a result of any of them. Additionally, we originally wanted to run the linear regression model across different occupational industries; if we were to continue this project, we would create that model and see if there are any significant differences across different industries.