Dataset Overview

The dataset was obtained from the Census Bureau Database of America in 1994 by Ronny Kohavi and Barry Becker. The dataset was created for census purposes, but later became a benchmark for advanced statistic analysis and machine learning, such as Naive Bayes and Nearest Neighbor, to predict the income group of a person. However, the dataset that is used for this analysis is only ⅔ of the original dataset.

This version contains 32561 rows, or 30162 rows if missing values are removed. The dataset has 17 features/columns with examples of the features are age, education, and marital status. The features mainly can be divided into 3 types, demographic, employment, and economic.

Motives

One of the columns is hours per week which represents the amount of hours that a person works in a week. This analysis will build a model to find what features may influence the hours per week and try to make a prediction based on these features.

Model Overview

The model in this analysis uses backward eliminations which the model will start with a bunch of features and gradually remove the features with low significance. However, to make sure that our first model won't get too big and make the process complex, the initial features will be selected intuitively.
Initial features with their reason/expectations:

- Age: Unknown correlation since older people may work more until they retire.
- Workclass: Positive correlation for a working person and zero correlation for others
- Education Num: Positive Correlation since higher education may lead to higher workload
- Sex: Males will have a higher correlation than Female
- Marital Status: Unknown correlation, but it's interesting to analyze
- Gross income group: Positive correlation for a group with income higher than 50K

- Occupation: Some types of jobs may require a higher amount of working hours.

Model Results

We can assess which column contributes/influences hour per week by seeing the feature left on our independent variables. From the appendix, some of the features include age, sex (Male and Female), education_num, and some types of jobs like Sales and Farming. However, some features were dropped, such as ‘Widowed’, indicating insignificance.

Since some of the columns are categorical variables, we need to set a default value for each of them. Our default value is:

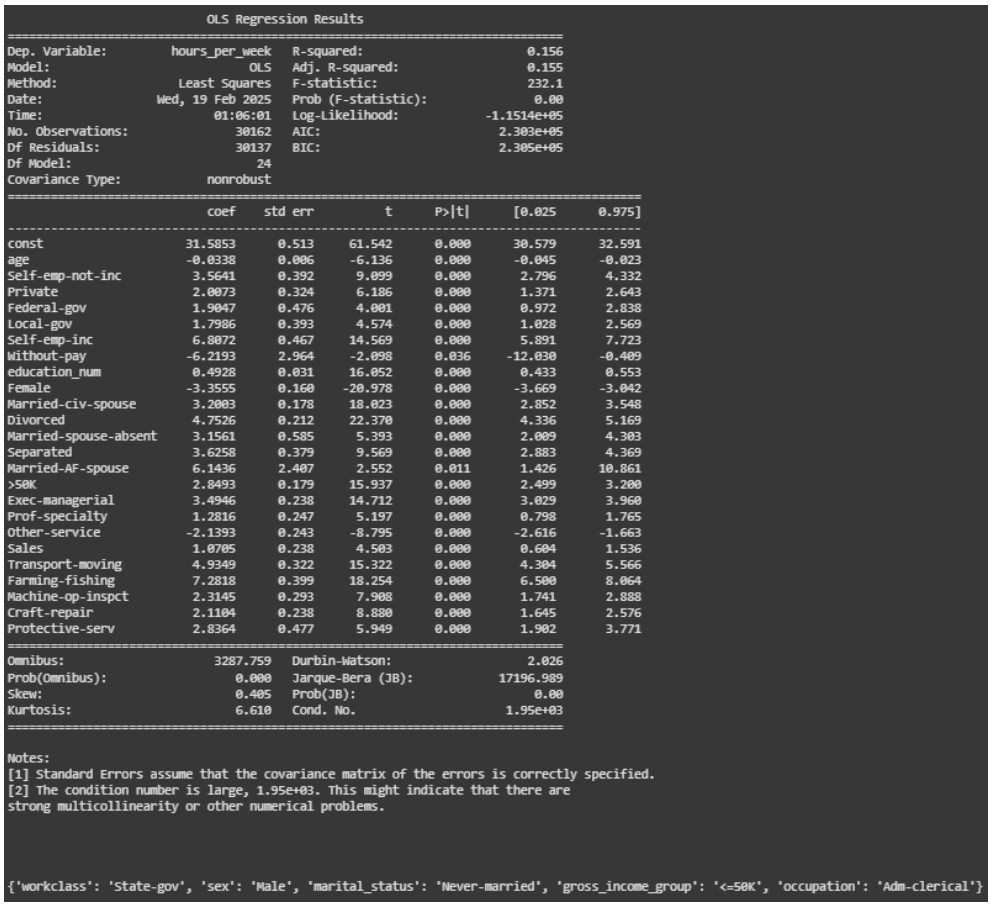
- Workclass: State-Gov
- Sex: Male
- Marital Status: Never Married
- Gross-Income-Group: Less than 50K
- Occupation: Adm-clerical

To interpret this model further, we can see the coefficient. We can see that age has a negative coefficient, which means that every increase in age will lead to a decrease of 0.034 hours per week. We also have a negative coefficient on Females (-3.3355) which means, that holding other features value the same, females will have 3.335 hours less than males. Besides, we also have a positive coefficient on education_num (0.5), which means that more educated people will have higher working hours.

For prediction purposes, a person with a default value on every categorical class will have 31.583 - 0.034 * age + 0.5 * education_num hours. If the same case but it’s a Female, it will be the original value - 3.355. Further, if instead of Less than 50K, the income group is At least 50K, then the value will be 2.9 higher.

Lastly, our model has an R^2 of 0.156 which means that our model can explain 15.6% of the data variance. Further, we can see that the RMSE is 11 which means that, on average, our model prediction is 11 hours far off from the true value. These results indicate poor performance, leading a further research.

Appendix



OLS Regression Results

Dep. Variable:	hours_per_week	R-squared:	0.156
Model:	OLS	Adj. R-squared:	0.155
Method:	Least Squares	F-statistic:	232.1
Date:	Wed, 19 Feb 2025	Prob (F-statistic):	0.00
Time:	01:06:01	Log-Likelihood:	-1.1514e+05
No. Observations:	30162	AIC:	2.303e+05
Df Residuals:	30137	BIC:	2.305e+05
Df Model:	24		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	31.5853	0.513	61.542	0.000	30.579	32.591
age	-0.0338	0.006	-6.136	0.000	-0.045	-0.023
Self-emp-not-inc	3.5641	0.392	9.099	0.000	2.796	4.332
Private	2.0073	0.324	6.186	0.000	1.371	2.643
Federal-gov	1.9047	0.476	4.001	0.000	0.972	2.838
Local-gov	1.7986	0.393	4.574	0.000	1.028	2.569
Self-emp-inc	6.8072	0.467	14.569	0.000	5.891	7.723
Without-pay	-6.2193	2.964	-2.098	0.036	-12.030	-0.409
education_num	0.4928	0.031	16.052	0.000	0.433	0.553
Female	-3.3555	0.160	-20.978	0.000	-3.669	-3.042
Married-civ-spouse	3.2003	0.178	18.023	0.000	2.852	3.548
Divorced	4.7526	0.212	22.370	0.000	4.336	5.169
Married-spouse-absent	3.1561	0.585	5.393	0.000	2.009	4.303
Separated	3.6258	0.379	9.569	0.000	2.883	4.369
Married-AF-spouse	6.1436	2.407	2.552	0.011	1.426	10.861
>50K	2.8493	0.179	15.937	0.000	2.499	3.200
Exec-managerial	3.4946	0.238	14.712	0.000	3.029	3.960
Prof-specialty	1.2816	0.247	5.197	0.000	0.798	1.765
Other-service	-2.1393	0.243	-8.795	0.000	-2.616	-1.663
Sales	1.0705	0.238	4.503	0.000	0.604	1.536
Transport-moving	4.9349	0.322	15.322	0.000	4.304	5.566
Farming-fishing	7.2818	0.399	18.254	0.000	6.500	8.064
Machine-op-inspct	2.3145	0.293	7.908	0.000	1.741	2.888
Craft-repair	2.1104	0.238	8.880	0.000	1.645	2.576
Protective-serv	2.8364	0.477	5.949	0.000	1.902	3.771

Omnibus:	3287.759	Durbin-Watson:	2.026
Prob(Omnibus):	0.000	Jarque-Bera (JB):	17196.989
Skew:	0.405	Prob(JB):	0.00
Kurtosis:	6.610	Cond. No.	1.95e+03

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.95e+03. This might indicate that there are
strong multicollinearity or other numerical problems.

{'workclass': 'State-gov', 'sex': 'Male', 'marital_status': 'Never-married', 'gross_income_group': '<=50K', 'occupation': 'Adm-clerical'}

Image 1: Summary of the Final Model with Default Value as a dict