Link to the Github repo.

Introduction

In this project, we're looking at how different things like age, whether someone is male or female, their race, if they're married, and their education level can affect how much money sales representatives make every hour. We're using a big set of data called the Current Population Survey (CPS) that gives us information about how much people earn every week and other details about them. We're going to make four different models, or ways of guessing earnings, starting with a really simple one and then making them more complex. We want to see which model does the best job at guessing how much money people make. We'll check how good our models are by looking at their errors and how well they work on new data they haven't seen before.

Data preparation

Data Acquisition
The dataset is loaded from a remote source into our environment. This dataset, derived from the Current Population Survey (CPS), provides comprehensive information about individuals' earnings and other demographic details, crucial for our analysis.

Enhancing the dataset. Feature engineering
In this stage, our objective is to refine the dataset, ensuring it's well-prepared for the predictive modeling process. We begin by transforming weekly earnings into hourly wages, providing a consistent basis for comparing earnings across different individuals. Additionally, we convert various categorical variables such as gender, race, and educational qualifications into numerical format, facilitating their integration into our statistical models.

Recognizing the nuanced relationship between age and earnings, we introduce polynomial terms of age up to the fourth power. This approach allows our models to capture potential non-linear patterns, offering a more detailed understanding of how earnings may vary with age.

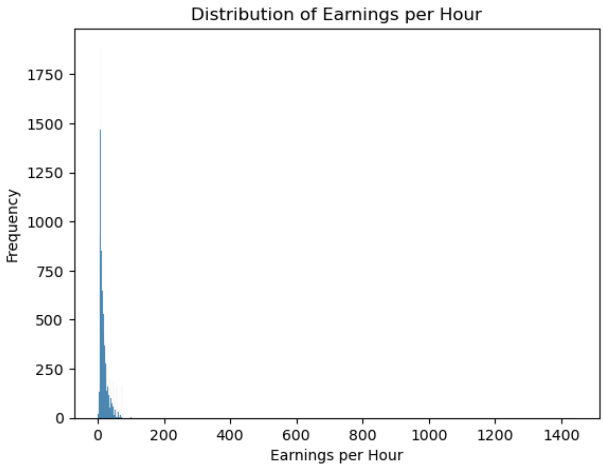
We also encode demographic variables like marital status and the presence of children, along with job-related factors such as employment sector, industry codes, and union membership. These variables often play a significant role in determining earnings and are, therefore, crucial components of our analysis.

Through these transformations, we aim to enrich our dataset, ensuring that it accurately mirrors the complexities inherent in the labor market. This meticulous preparation sets the stage for developing robust models capable of making insightful predictions about hourly earnings.

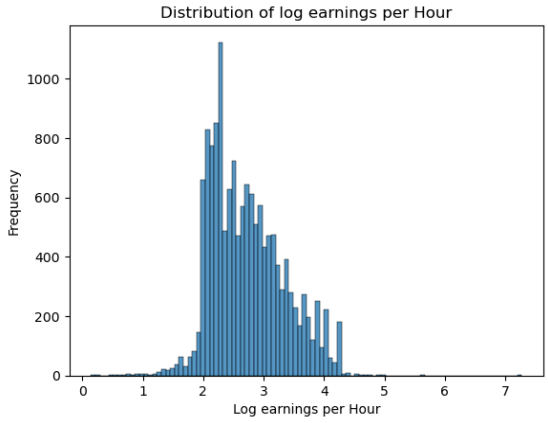
Our objective is to forecast the hourly earnings specifically for individuals in sales roles.
We'll focus our analysis on sales and associated occupations, as classified by the Census codes [4700, 4965].
The primary variable of interest in this analysis is the hourly earnings. To align our dataset with this goal, it's essential to refine our data, selecting only the relevant records that correspond to our area of study.

Initially, our dataset contained 149k records. Once we filtered to include only those pertaining to sales occupations, the dataset was reduced to 14577 observations. Thus, our focused dataset for analysis comprises 14577 entries. It's unclear if any anomalies are errors, especially in cases like his, where the individual is a sales manager in wholesale. Positions like these can command exceptionally high earnings, often due to large commissions from significant contracts.

Utilizing log earnings seems to be a reasonable approach, particularly because the wage distribution is not even and leans towards the lower end. This distribution exhibits a long tail extending to the right, indicating that a few individuals have significantly higher wages. By transforming earnings into their logarithmic scale, we can address this skewness and achieve a more normalized distribution for our analysis.



The distribution of the natural logarithm of wages (lnw) appears significantly improved and more symmetrical. Therefore, we will proceed with lnw as our target variable for the forthcoming analysis.



Regression analysis

In our pursuit to create an accurate predictive model for hourly earnings, we've developed four distinct linear regression models, each with varying complexity and depth of analysis:

Model 1 - Basic Age Model: We start with the simplest model, focusing solely on the age variable and its higher powers (squared, cubed, and to the fourth power). This model aims to capture the non-linear impact of age on hourly earnings.

Model 2 - Adding Demographic Variables: Building on Model 1, we introduce demographic variables such as gender and race to the equation. This model seeks to understand how these factors, in addition to age, influence hourly earnings.

Model 3 - Including Marital Status, Education, and Children Presence: In this model, we include variables related to marital status, the presence of children, and education levels (Master's, Professional degree, Ph.D.). This model examines how personal and educational background correlates with earnings.

Model 4 - Incorporating Job-Related Variables and Interaction Terms: Our most complex model incorporates job-related factors such as the sector of employment and union membership. Additionally, we introduce interaction terms between age and these job-related variables to investigate if and how the impact of age on earnings varies across different employment conditions.

The selection of predictors in these models is driven by a comprehensive understanding of the factors that potentially influence hourly earnings, each model building upon the last to incorporate a richer set of characteristics.

Model 1 (Basic Age Model): This model serves as our baseline, focusing solely on the individual's age and its polynomial terms up to the fourth degree. The choice of age and its powers aims to capture not just the linear but also the more complex, non-linear effects of age on earnings, recognizing that the relationship between age and earnings is not necessarily straightforward.

Model 2 (Demographic Model): Building upon Model 1, this model integrates key demographic variables such as gender and race. These factors are known to impact earnings due to systemic trends in the labor market, like gender pay gaps and racial income disparities. Including these variables allows the model to account for these influences and understand how they interact with age in determining earnings.

Model 3 (Socio-Economic Model): This model further expands by incorporating variables related to marital status, the presence of children, and educational attainment. These socio-economic factors can significantly affect an individual's earning capacity, with marital status potentially reflecting dual-income dynamics, children introducing financial responsibilities, and education level often correlating with job opportunities and earning potential.

Model 4 (Occupational Model): The most comprehensive model includes job-related factors such as employment sector, union membership, and interaction terms between age and these job-related variables. These factors are crucial as they directly relate to an individual's occupation and work environment, which can significantly influence earnings. The interaction

terms specifically allow the model to explore if and how the effect of age on earnings varies across different job conditions and sectors.

The R-squared values are most favorable for the 3rd and 4th models, indicating a strong explanatory power. However, considering the simplicity of the 3rd model, it's presumed to be more suitable for use as a predictive model. Nonetheless, we plan to further validate and assess the performance of our models using cross-validation techniques.

The R-squared values are most favorable for the 3rd and 4th models, indicating a strong explanatory power. However, considering the simplicity of the 3rd model, it's presumed to be more suitable for use as a predictive model. Nonetheless, we plan to further validate and assess the performance of our models using cross-validation techniques (**Appendix**).

Cross Validation

The results from the 5-fold cross-validation reveal a consistent improvement in model performance as we transition from Model 1 to Model 4. The R-squared values indicate an increase in explanatory power, peaking at 0.29 for both Models 3 and 4, suggesting that these models are better at capturing the variance in hourly earnings. However, Model 3, with fewer variables (8 vs. 15) and coefficients (20 vs. 30), achieves a similar level of predictive accuracy as Model 4, as evidenced by comparable RMSE values (0.52 for both models) and a lower BIC(22361 vs 22438), indicating a better balance between model complexity and performance. The cross-validation results, with nearly identical average RMSE values for Models 3 and 4 across all folds, further support the notion that Model 3 is a more efficient choice for predicting hourly earnings without sacrificing predictive power. This analysis underscores the importance of considering model simplicity alongside predictive accuracy, especially when incremental increases in complexity do not translate into significant performance gains.

Prediction

We are eager to see how these models perform with real-world data. It's important to understand their practical applicability and effectiveness in live scenarios.

The model's prediction for a 30-year-old, white, married sales representative male with no children and a Ph.D. degree suggests an hourly earnings of 3.5 on the logarithmic scale. When this figure is translated back to the actual earnings scale, it corresponds to an estimated hourly earning of approximately 37.86. If we look at the direct hourly earnings estimate from the model in levels, it's slightly lower, at around 37.06. These estimates reflect the model's understanding of the relationship between an individual's characteristics—particularly the high level of education in this case—and their earning potential.

The 80% prediction interval provides a range for the expected hourly earnings, accounting for the inherent uncertainty in the model's estimates. On the log scale, this interval spans from 2.83 to 4.17, which when converted to actual earnings, translates to a range of about 19.37 to 73.99. Meanwhile, the direct level prediction offers a narrower interval, ranging from 14.78 to 59.34.

These intervals capture the variability and potential uncertainty in the hourly earnings for an individual with these specific characteristics, offering a realistic understanding of the possible range of earnings and emphasizing the model's confidence in its predictions.
The generalizability and external validity of our models are promising due to the comprehensive nature of the Current Population Survey (CPS) dataset, which captures a wide array of demographic and occupational variables. However, this validity may be moderated by specific factors such as temporal changes in economic conditions, evolving labor market dynamics, and regional disparities that were not present in the dataset. While the inclusion of a diverse range of predictors enhances the robustness of the models, there's a potential risk of overfitting with more complex models, which may limit their applicability to new or future data. Consequently, while our models offer valuable and broadly representative insights within the dataset's context, cautious application and ongoing validation are necessary when extending these findings to wider or evolving scenarios.

Appendix

	Dependent variable: lnw			
	(1)	(2)	(3)	(4)
Intercept	0.919***	1.027***	0.636**	0.644***
	(0.300)	(0.296)	(0.248)	(0.248)
afram		-0.058**	-0.050**	-0.050**
		(0.024)	(0.023)	(0.023)
age	0.052	0.061*	0.108***	0.106***
	(0.037)	(0.036)	(0.036)	(0.036)
age:fedgov				0.003
				(0.004)
age:locgov				0.006**
				(0.003)
age:nonprof				-0.002
				(0.003)
age:stagov				-0.004
				(0.005)
age:union				-0.000
				(0.002)
agecu	-0.000**	-0.000*	-0.000	-0.000
	(0.000)	(0.000)	(0.000)	(0.000)
agequ	0.000***	0.000**	0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)
agesq	0.002	0.001	-0.001	-0.001
	(0.002)	(0.002)	(0.002)	(0.002)
asian		0.082***	0.053*	0.052*
		(0.030)	(0.030)	(0.030)
child0			0.107**	0.109**

		(0.050)	(0.050)
child1		0.197***	0.199***
		(0.054)	(0.054)
child2		0.071	0.073
		(0.056)	(0.056)
child3		0.146***	0.148***
		(0.050)	(0.050)
child4pl		0.115**	0.116**
		(0.049)	(0.049)
divorced		-0.000	0.000**
		(0.000)	(0.000)
ed_MA		0.358***	0.361***
		(0.028)	(0.028)
ed_Phd		0.541***	0.537***
		(0.093)	(0.093)
ed_Profess		0.328***	0.329***
		(0.095)	(0.095)
fedgov			0.053
			(0.158)
female	-0.275***	-0.261***	-0.261***
	(0.009)	(0.009)	(0.009)
hisp	-0.133***	-0.123***	-0.123***
	(0.012)	(0.012)	(0.012)
locgov			-0.254**
			(0.123)
married		0.118***	0.118***
		(0.015)	(0.015)
nevermar		-0.006	-0.006

			(0.017)	(0.017)
nonprof				0.011
				(0.131)
stagov				0.190
				(0.203)
union				0.040
				(0.055)
white		0.140***	0.128***	0.130***
		(0.021)	(0.021)	(0.021)
widowed			-0.042	-0.040
			(0.041)	(0.041)
Observations	14577	14577	14577	14577
R^2	0.199	0.266	0.291	0.292
Adjusted R^2	0.199	0.266	0.290	0.290
Residual Std. Error	0.550 (df=14572)	0.527 (df=14567)	0.518 (df=14557)	0.518 (df=14547)
F Statistic	1402.961*** (df=4; 14572)	682.753*** (df=9; 14567)	13738.837*** (df=19; 14557)	9110.705*** (df=29; 14547)
Note:				*p<0.1; **p<0.05; ***p<0.01