



Supervised Machine Learning Capstone Project

US Census Income Data (1994 – 1995)



Content

- Background
- Problem Statement
- Data Set
- Analytical Questions
- Exploratory Data Analysis
- Data Manipulation
- Data Visualization
- Machine Learning Models
- Model Evaluation
- Final Thoughts



Background

The prominent inequality of income continues to be a pressing problem especially in the United States despite federal laws protecting against pay discrimination by race, ethnicity, and gender.

The principle of universal moral equality ensures sustainable development and improve the economic stability of a nation.



Problem Statement

"Given various features, the aim is to build a predictive model to determine the income level for people in US. The income levels are binned at below 50K and above 50K."

Data Set

UCI Census Income Dataset has been used for the purpose.

- This data set contains weighted census data extracted from the 1994 and 1995 current population surveys conducted by the U.S. Census Bureau.
- Classification has been done to predict whether a person's annual income in US falls in the income category of either greater than 50K Dollars or less equal to 50K Dollars category based on a certain set of demographic and employment related attributes.

Analytical Questions

1. What features, within the provided dataset, are most determinant of income level?
2. Which prediction models or algorithms perform best in terms of speed, accuracy, and explainability combined?
3. How good is our trained model in predicting the income level given a set of demographic and employment features?

Exploratory Data Analysis

Basic statistics for this data set

- Number of instances in data = 199523
- Duplicate or conflicting instances : 46627
- Number of attributes = 40 (continuous : 7 nominal : 33)
- Target variable = "Income Level" binned at the \$50K level (binary classification problem).

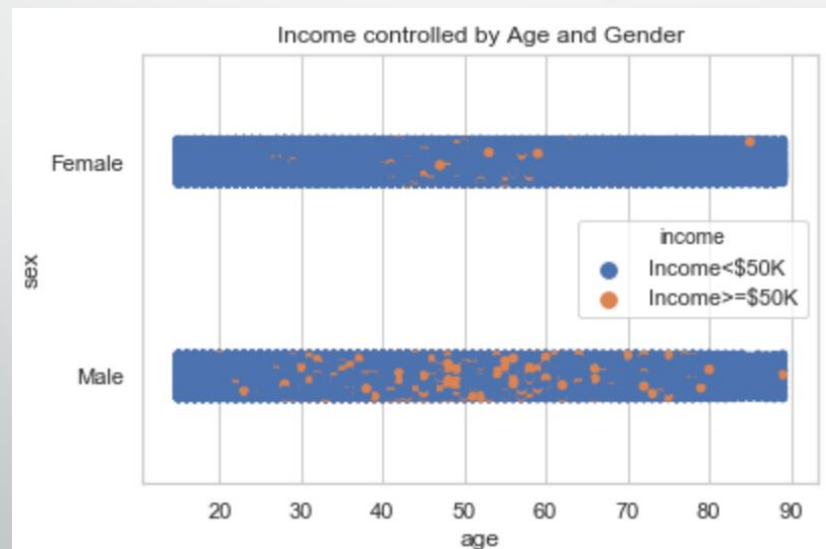
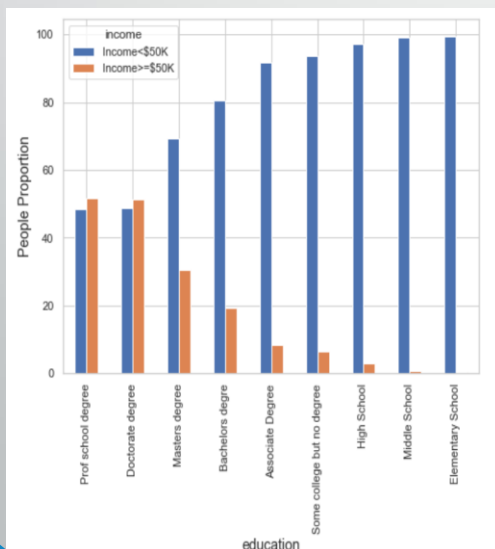
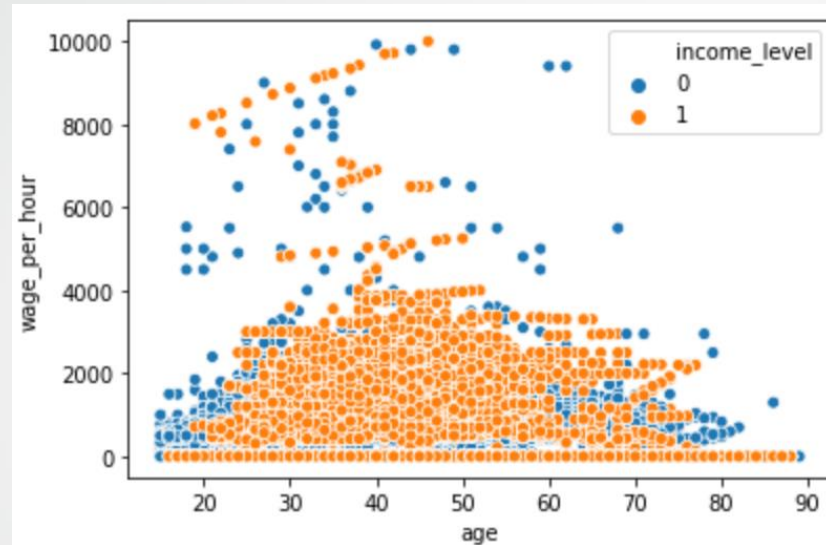
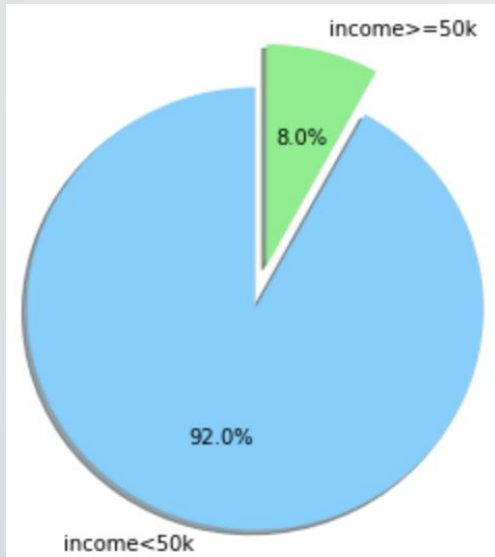
Train-Test-Split

- The data was split into train/test in $\frac{2}{3}$ and $\frac{1}{3}$ proportions.

Data Manipulation

- The values represented by 'Not in universe', 'Not in universe or children' and 'Not in universe under 1 year old' are replaced by unknown.
- Education, Marital Status, Class Of Worker and Tax Filer Status is classified in multiple similar categories.
- Since Migration type Features are missing 50% values and are deleted.
- Country type Features have been filled with the mode values.
- Observations with education = "Children", capital gains = 99999 and age = 90 are deleted.

Data Visualizations



- 92% have income less than 50K and 8% have income greater than 50K
- White and Asian-Pacific-Islanders earn salary more than 50K
- Greater distribution of income > 50k among males between age 30 – 65, and moderate distribution of income > 50k among females between age 45 – 60 gender pay gap.
- Doctorate, Professional School, and Masters degree holders are making salary more than 50K.

Machine Learning Models

- Preprocess and Transform the data set.
- Use SMOTE to balance the data set.
- Split the data set between Training data set and Test data set at 67% and 33% proportions.
- Explore and Evaluate various ML Algorithms.
- Interpret and Report results.
- Refine and Improve the results by tuning hyper parameters.

ML Algorithms

- Logistic Regression
- K- Nearest Neighbors
- Decision Trees
- Random Forest
- Extra Trees
- Gradient Boost
- ADA Boost
- Bagging
- SVM
- Naïve Bayes

Model Evaluation

Making predictions on this data should atleast give us ~94% accuracy. However, while working on imbalanced problems, accuracy is considered to be a poor evaluation metrics because:

1. Accuracy is calculated by ratio of correct classifications / incorrect classifications.
2. This metric would largely tell us how accurate our predictions are on the majority class (since it comprises 94% of values). But, we need to know if we are predicting minority class correctly. We're doomed here.

In such situations, we should use elements of confusion matrix.

		Actual	
		+	-
Predicted	Y	True positives	False positives
	N	False negatives	True negatives

Following are the metrics we'll use to evaluate our predictive accuracy:

- Sensitivity = True Positive Rate ($TP / (TP + FN)$) – It says, 'out of all the positive (majority class) values, how many have been predicted correctly'.
- Specificity = True Negative Rate ($TN / (TN + FP)$) – It says, 'out of all the negative (minority class) values, how many have been predicted correctly'.
- Precision = $(TP / (TP + FP))$
- Recall = Sensitivity
- F score = $2 * (Precision * Recall) / (Precision + Recall)$ – It is the harmonic mean of precision and recall. It is used to compare several models side-by-side. Higher the better.

Model Evaluation

MODEL/METRIC	Accuracy		Sensitivity		Specificity		FP Rate		Precision		F-Score		Time	
	Imbalanced	Balanced	Imbalanced	Balanced	Imbalanced	Balanced	Imbalanced	Balanced	Imbalanced	Balanced	Imbalanced	Balanced	Imbalanced	Balanced
Random Forest	93.49%	96.67%	32.80%	95.26%	98.70%	98.07%	1.30%	1.93%	68.46%	98.01%	44.35%	96.62%	0m:05s	1m 40s
Bagging	93.58%	96.34%	40.13%	95.22%	98.16%	97.47%	1.84%	2.53%	65.25%	97.41%	49.70%	96.30%	1m:02s	1m 31s
Extra Trees	92.97%	96.22%	31.99%	94.64%	98.21%	97.80%	1.79%	2.20%	60.56%	97.73%	41.87%	96.16%	0m:08s	2m 42s
Decision Tree	91.45%	95.17%	47.44%	95.44%	95.22%	94.91%	4.78%	5.09%	46.03%	94.94%	46.73%	95.19%	00:11s	14s
KNN	92.51%	93.14%	28.54%	99.60%	98.01%	86.67%	1.99%	13.33%	55.15%	88.20%	37.62%	93.55%	16m:03s	13m 6s
Gradient Boosting	94.12%	92.83%	39.43%	91.67%	98.81%	94.00%	1.19%	6.00%	74.03%	93.85%	51.45%	92.75%	1m:44s	3m 14s
Adaboost	93.72%	92.81%	34.07%	92.30%	98.84%	93.32%	1.16%	6.68%	71.57%	93.25%	46.16%	92.77%	0m:25s	57s
Logistic Regression	93.72%	81.78%	35.75%	86.89%	98.70%	76.67%	1.30%	23.33%	70.20%	78.83%	47.37%	82.67%	0m:06s	13s
Naïve Bayes	34.82%	78.98%	95.18%	98.29%	29.63%	59.66%	70.37%	40.34%	10.41%	70.90%	18.76%	82.38%	00:01s	3s
SVM**	92.10%	78.12%	0.10%	73.32%	100.00%	82.91%	0.00%	17.09%	100.00%	81.08%	0.20%	77.00%	19m:35s	5h 2m 32s

** SVM model took 5 hours to complete. The metrics seems way out of proportion.

Final Thoughts

Conclusions


After comparing the model performance using the confusion matrix and accuracy of the model, we can certainly say that the Random Forest model is better fit to predict the income of an individual based on the census data.

Practical Uses

- Establish structures, policies, objectives in every organization to ensure gender balance
- This analysis can be useful to the banks or other financial institutes to know the American salary distribution and target potential borrowers and segment them appropriately

Future Considerations

May consider doing additional analysis that focuses on dimension reduction such as Principal Component Analysis (PCA) or Linear Discriminant Analysis (LDA). Since this data set represents classification problem, LDA can be used as a preprocessing step in Machine Learning and pattern classification applications.



Q&A