University of North Carolina at Charlotte

Predicting Income Using Census Data

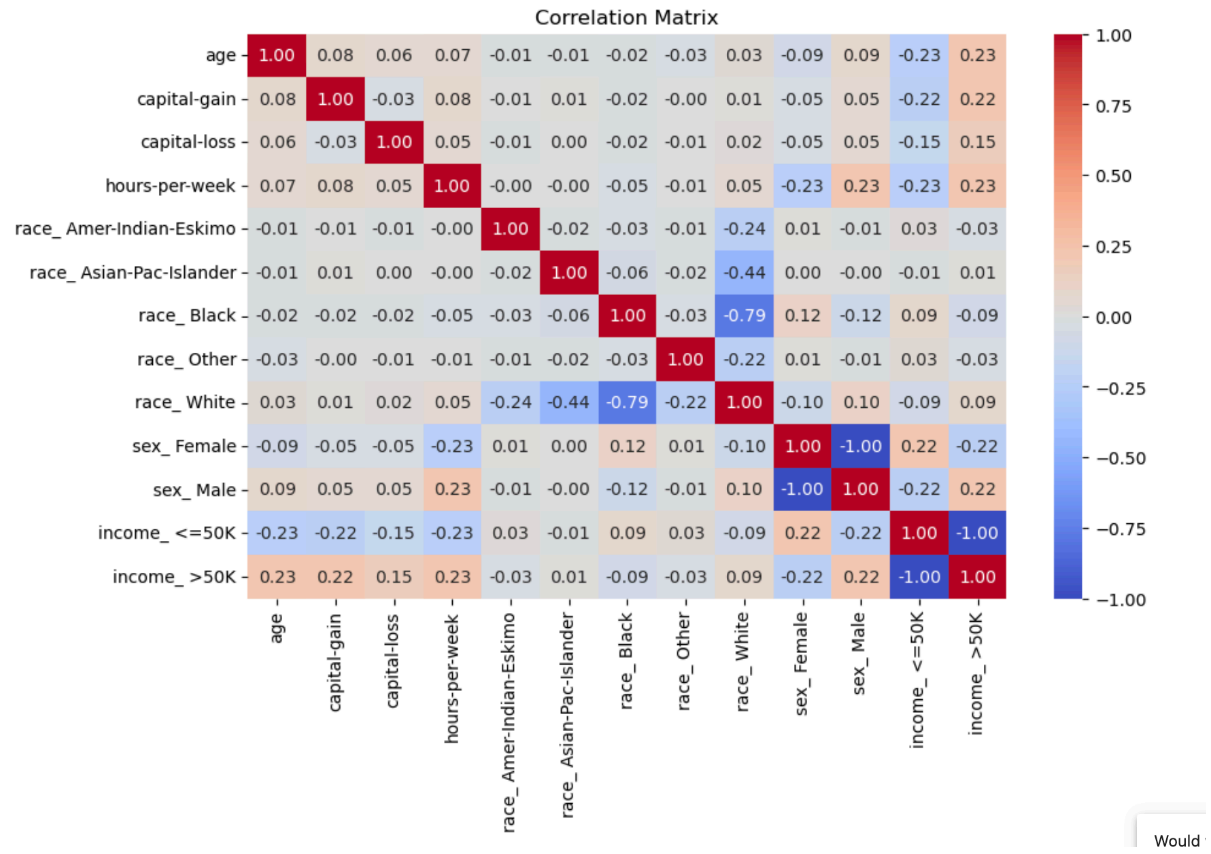
Michael Spitz
ITSC 3156 - Intro. to Machine Learning
Aileen Benedict
12/12/2024

Introduction:
I decided I wanted to analyze the Adult dataset. My research focuses on the adult dataset and analyzing it with complex models, I wanted to use some of the simpler models we learned to see how much more accurate a model can be. The adult dataset contains 14 features with 47000+ rows of data.

Data cleaning:
The adult dataset is a dataset extracted from the 1990 US Census dataset, this dataset aims to predict whether income is more than $50000 or less. This dataset has a lot of custom fields and values that can be understood more at: https://archive.ics.uci.edu/dataset/2/adult

I decided to remove the feature 'fnlwgt'. This is a formula calculated by the team that made this dataset. I have no notation for it or anything, rendering it as a useless AND meaningless feature. I also dropped 'education-num', this was a numerical representation of the years spent in school. Once one hot encoded 'education' will give the same data. Finally 'native-country' and 'relationship' were dropped as they intuitively seem irrelevant to the task at hand. I decided to leave race and sex to find any general trends regarding both.

After removing these features we can take a look at the correlation matrix:

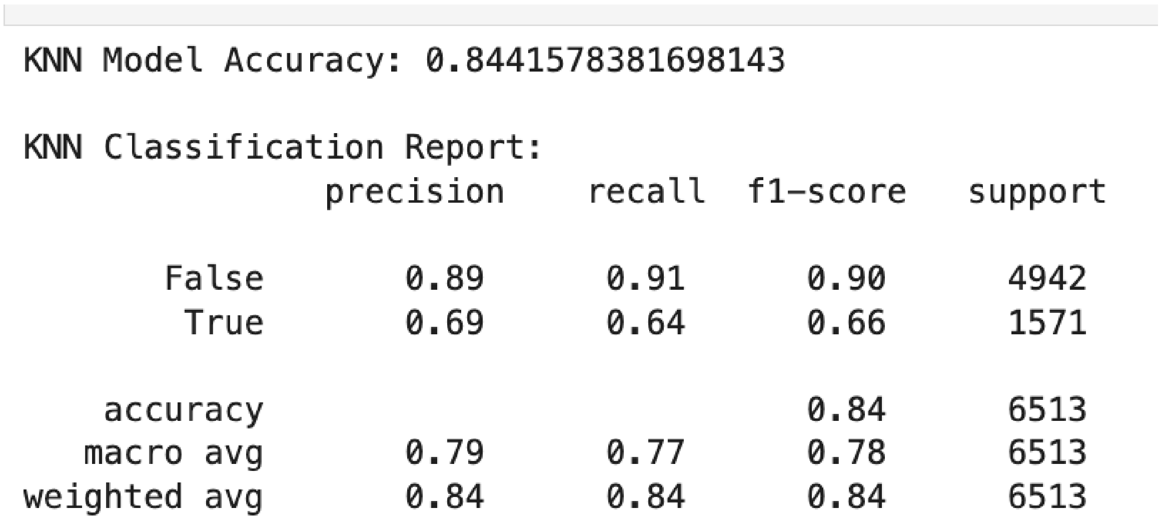


	age	capital-gain	capital-loss	hours-per-week	race_ Amer-Indian-Eskimo	race_ Asian-Pac-Islander	race_ Black	race_ Other	race_ White	sex_ Female	sex_ Male	income_ <=50K	income_ >50K
age	1.00	0.08	0.06	0.07	-0.01	-0.01	-0.02	-0.03	0.03	-0.09	0.09	-0.23	0.23
capital-gain	0.08	1.00	-0.03	0.08	-0.01	0.01	-0.02	-0.00	0.01	-0.05	0.05	-0.22	0.22
capital-loss	0.06	-0.03	1.00	0.05	-0.01	0.00	-0.02	-0.01	0.02	-0.05	0.05	-0.15	0.15
hours-per-week	0.07	0.08	0.05	1.00	-0.00	-0.00	-0.05	-0.01	0.05	-0.23	0.23	-0.23	0.23
race_ Amer-Indian-Eskimo	-0.01	-0.01	-0.01	-0.00	1.00	-0.02	-0.03	-0.01	-0.24	0.01	-0.01	0.03	-0.03
race_ Asian-Pac-Islander	-0.01	0.01	0.00	-0.00	-0.02	1.00	-0.06	-0.02	-0.44	0.00	-0.00	-0.01	0.01
race_ Black	-0.02	-0.02	-0.02	-0.05	-0.03	-0.06	1.00	-0.03	-0.79	0.12	-0.12	0.09	-0.09
race_ Other	-0.03	-0.00	-0.01	-0.01	-0.01	-0.02	-0.03	1.00	-0.22	0.01	-0.01	0.03	-0.03
race_ White	0.03	0.01	0.02	0.05	-0.24	-0.44	-0.79	-0.22	1.00	-0.10	0.10	-0.09	0.09
sex_ Female	-0.09	-0.05	-0.05	-0.23	0.01	0.00	0.12	0.01	-0.10	1.00	-1.00	0.22	-0.22
sex_ Male	0.09	0.05	0.05	0.23	-0.01	-0.00	-0.12	-0.01	0.10	-1.00	1.00	-0.22	0.22
income_ <=50K	-0.23	-0.22	-0.15	-0.23	0.03	-0.01	0.09	0.03	-0.09	0.22	-0.22	1.00	-1.00
income_ >50K	0.23	0.22	0.15	0.23	-0.03	0.01	-0.09	-0.03	0.09	-0.22	0.22	-1.00	1.00

As you can see males and females both have the same amount of correlation to having income greater than 50000. Black and white races also have the same correlation for having greater than 50k and having less than 50k. This is interesting and presses against my hypothesis of whites having a higher correlation.

Methods:
I decided to use knn as well as linear regression, I wanted to see how similarly they would perform. I only used 5 neighbors on the knn as I didn't want to be hindered by the curse of dimensionality. I tried other depths but none of them were as good as 5, 1 overfit too much while going any higher than 5 caused issues.

Results:



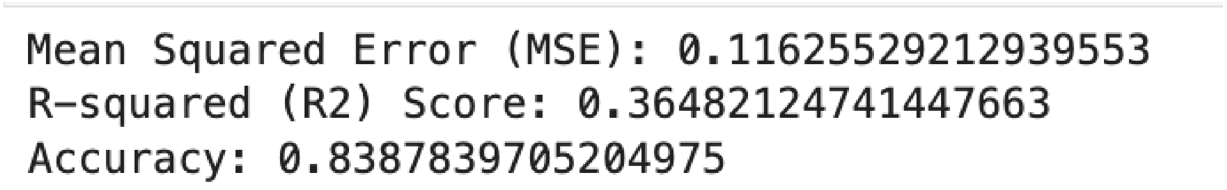
```
KNN Model Accuracy: 0.8441578381698143

KNN Classification Report:
              precision    recall  f1-score   support

       False       0.89      0.91      0.90      4942
        True       0.69      0.64      0.66      1571

    accuracy                           0.84      6513
   macro avg       0.79      0.77      0.78      6513
weighted avg       0.84      0.84      0.84      6513
```

The knn algorithm recorded an accuracy of 84% at a depth of 5. Meanwhile the linear regression algorithm managed an accuracy of 83%.



```
Mean Squared Error (MSE): 0.11625529212939553
R-squared (R2) Score: 0.36482124741447663
Accuracy: 0.8387839705204975
```

The models performed quite similarly.

Conclusions:
The models performed quite similarly which was a big surprise to me. I assumed with the large amount of dimensions a lack of standardization the knn model would completely outclass the linear regression model. This was not the case. The models that were used in my research yielded an accuracy of about 87% which is interesting considering they use 5 more features and a more complex model. By removing features and using a simple model we can achieve a similar result. I learned that complex isn’t always better which goes hand in hand with the curse of dimensionality. I also learned that it isn’t that hard to create a model so long as you keep the steps of the machine learning pipeline in mind. This was a lot simpler than creating a knn or linear regression model from scratch. Overall I learned making a machine learning model isn’t

that hard once you've gone and made one on your own. It gives you a deeper understanding which allows you to understand the parameters and meaning of things better.

Acknowledgements:
I used chatGPT to help with a couple of errors regarding dropping columns(I ended up just needing to restart the kernel). I also used the notes and chatgpt to try to standardize the data, I ended up not standardizing the data and just one hot encoding instead. This lead to good results.

Citations: (research study I was building off of for my personal research which led to using the same dataset)
Akash, A. K., Lokhande, V. . S., Ravi, S. N., & Singh, V. (2021). Learning Invariant Representations using Inverse Contrastive Loss. Proceedings of the AAAI Conference on Artificial Intelligence, 35(8), 6582-6591. https://doi.org/10.1609/aaai.v35i8.16815