Technical Assessment

Neil Menghani

Introduction

Problem Statement

Can we identify characteristics associated with a person making more or less than $50,000 per year?

Proposed Solution

Develop a data analysis and modeling pipeline using data collected by the U.S. Census Bureau.

Steps:

1. Explore dataset for clear trends
2. Prepare dataset for modeling
3. Develop a predictive model to determine income level above or below $50,000

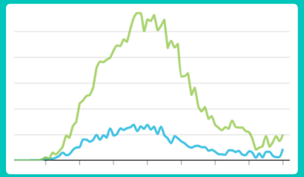
Methodology

dataiku

Explore Data

Determine correlations between variables

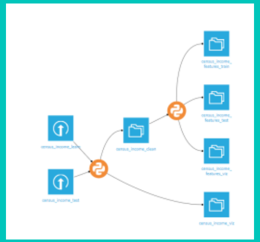
Graphically analyze trends in key variables

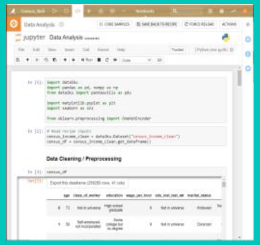


Prepare Data

Build a data pipeline for:

- Data Cleaning & Preprocessing
- Feature Engineering





Model

Build Machine Learning models for binary classification problem:

- Logistic Regression
- Tree-Based Models
- Neural Network

Evaluate models, extract feature importance, and draw conclusions

Technologies Used: Python (pandas, scikit-learn, keras/tf); Dataiku (Charts, Recipes, Jupyter Notebooks)

Exploratory Analysis

Correlation Matrix

Figure 1a: Correlation matrix of numerical features.

	Income Level											
Income Level	1.00	0.26	0.24	0.22	0.18	0.15	0.14	0.14	0.04	0.02	0.01	0.01
Weeks Worked in Year	0.26	1.00	0.08	0.75	0.01	0.10	0.53	0.21	0.27	0.20	0.01	0.03
Capital Gains	0.24	0.08	1.00	0.06	0.13	-0.01	0.05	0.05	0.02	-0.00	0.01	0.00
# Employed by Employer	0.22	0.75	0.06	1.00	0.01	0.08	0.46	0.14	0.24	0.19	0.02	0.04
Dividends from Stocks	0.18	0.01	0.13	0.01	1.00	0.04	0.05	0.10	-0.00	-0.01	0.00	-0.00
Capital Losses	0.15	0.10	-0.01	0.08	0.04	1.00	0.08	0.06	0.02	0.01	0.00	0.01
Veterans Benefits	0.14	0.53	0.05	0.46	0.05	0.08	1.00	0.67	0.18	0.11	0.00	0.04
Age	0.14	0.21	0.05	0.14	0.10	0.06	0.67	1.00	-0.00	0.04	0.00	-0.00
Own Business	0.04	0.27	0.02	0.24	-0.00	0.02	0.18	-0.00	1.00	0.05	0.01	0.01
Wage / Hour	0.02	0.20	-0.00	0.19	-0.01	0.01	0.11	0.04	0.05	1.00	0.00	0.01
Year	0.01	0.01	0.01	0.02	0.00	0.00	0.00	0.00	0.01	0.00	1.00	0.01
Instance Weight	0.01	0.03	0.00	0.04	-0.00	0.01	0.04	-0.00	0.01	0.01	0.01	1.00

Figure 1b: Correlation matrix of categorical features.

	Income Level									
Income Level	1.00	0.22	0.16	0.15	0.14	0.14	0.11	0.11	0.09	0.07
Class of Worker	0.22	1.00	0.07	0.62	0.40	-0.03	0.43	0.21	0.59	0.27
Sex	0.16	0.07	1.00	0.02	0.04	0.04	0.06	0.08	0.06	0.02
Industry	0.15	0.62	0.02	1.00	0.38	-0.05	0.39	0.19	0.71	0.24
Tax Filer Status	0.14	0.40	0.04	0.38	1.00	-0.09	0.17	0.22	0.38	0.16
Education	0.14	-0.03	0.04	-0.05	-0.09	1.00	-0.03	-0.02	-0.13	-0.03
Employment Status	0.11	0.43	0.06	0.39	0.17	-0.03	1.00	0.11	0.38	0.16
Marital Status	0.11	0.21	0.08	0.19	0.22	-0.02	0.11	1.00	0.17	0.08
Occupation	0.09	0.59	0.06	0.71	0.38	-0.13	0.38	0.17	1.00	0.22
Member Labor Union	0.07	0.27	0.02	0.24	0.16	-0.03	0.16	0.08	0.22	1.00

Exploratory Analysis

data iku

Plots of Key Features

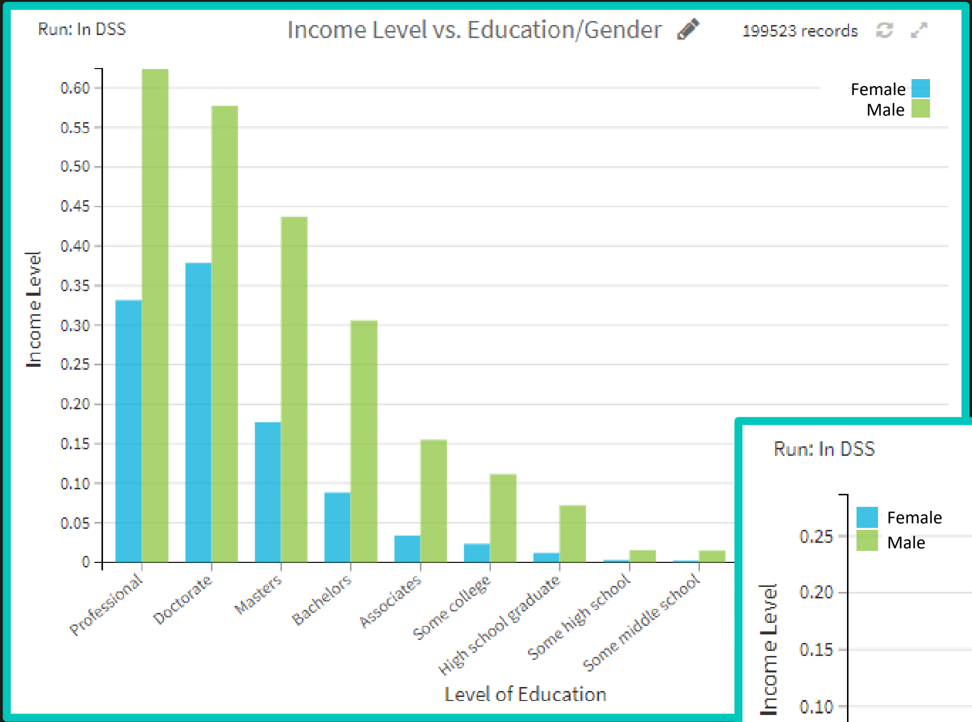
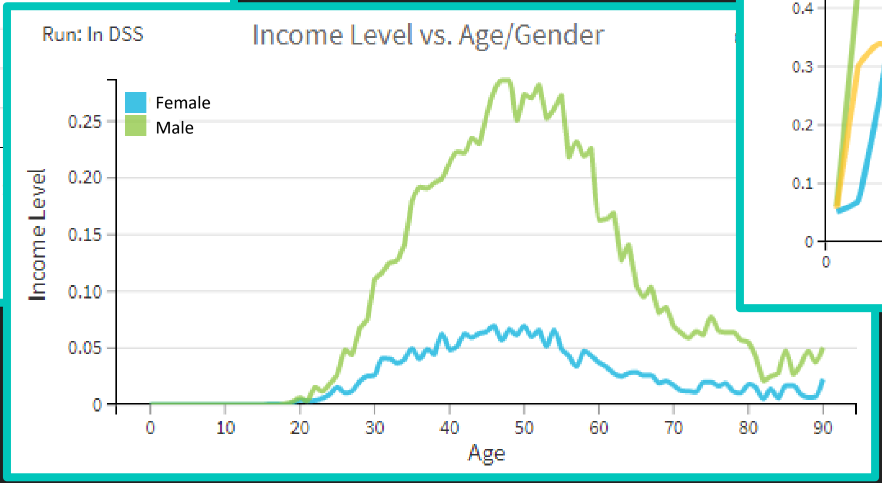


Figure 2a: Plot above shows average income level at various education levels (for each gender). Higher levels of education appear to translate to higher income.

Figure 2b: Plot below shows income level as a function of age (for each gender). Income appears to increase until mid-40s and then fall off.



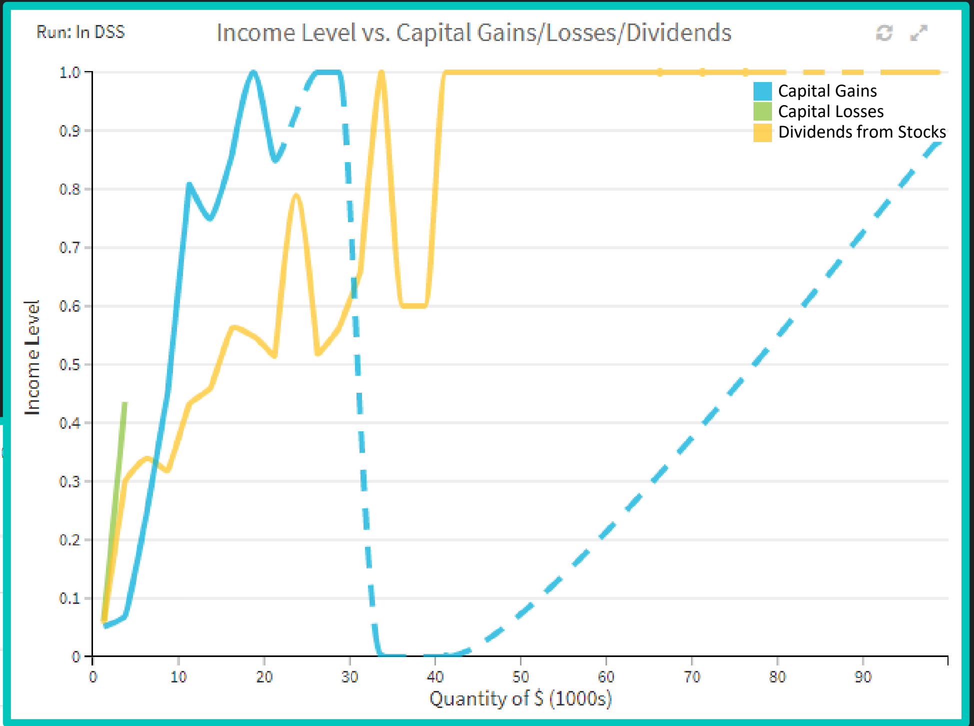


Figure 2c: Plot above shows income level as a function of dollars in capital gains, capital losses, and dividends from stocks. Income appears to increase as a function of these features.

Data Preparation

dataiku

Data Cleaning & Preprocessing

- Removed Null Values
- Converted Columns into Binary Features
  - Sex, Veterans, Year, and Hispanic
- Removed Redundant Columns
  - State, Household Family Status, Move Within Region
- Normalized Features
  - No Improvement to Model

Feature Engineering

- Reduced categories for categorical features (avoid wide data)
  - Re-categorized education into various levels (e.g. 9^{th}-12^{th} into Some High School)
  - Re-categorized country of birth into regions (e.g. Panama and Honduras into Central America)
- Generated one-hot encoding for categorical features

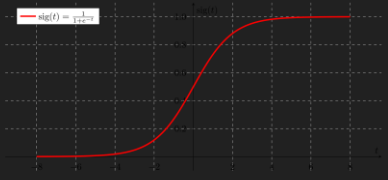
Machine Learning

Modeling Steps

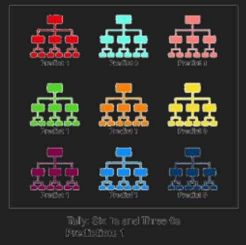
- Split the training set into train and validation sets (75/25 split)
  - Avoids learning on the test set
- Trained and tuned parameters for 3 types of machine learning models
  - Binary classification problem with target variable representing income above or below $50,000
- Generated test predictions for evaluation of model

Models Used

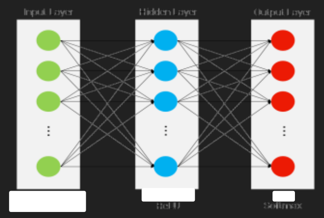
Logistic Regression



Tree-Based Ensemble Methods (Random Forest, Gradient Boosting)



Neural Network (1 Hidden Layer)



Model Evaluation

Model Results

	Train Accuracy	Validation Accuracy	Test Accuracy	AUC-ROC	F1-Score
Logistic Regression	0.9518	0.9515	0.9521	0.6821	0.7335
Random Forest	0.9656	0.9523	0.9526	0.6516	0.7108
Gradient Boosting	0.9640	0.9538	0.9539	0.6846	0.7398
Neural Network	0.9518	0.9523	0.9516	0.7066	0.7485

Figure 3a: Table shows training, validation, and testing accuracy; area under curve (ROC); and f1-score. Best-performing models for each metric are highlighted in yellow.

Feature Importance

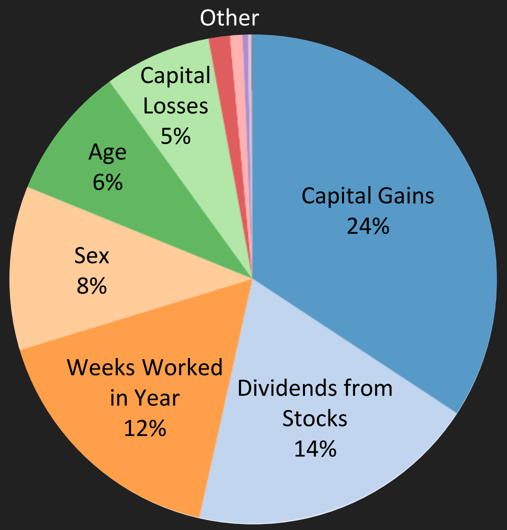


Figure 3b: Pie chart shows importance of each feature. Low-importance features grouped into Other category.

Conclusions

dataiku

Findings

- Key Variables According to Model:
  - Capital Gains, Dividends from Stocks, and Capital Losses
  - Weeks Worked in a Year
  - Sex and Age
- Best-Performing Models:
  - Accuracy: Gradient Boosted Ensemble Model
  - AUC/F1-Score: Neural Network with 1 Hidden Layer

Future Work

- Handle categorical features differently for Tree-Based Methods
- Use under-sampling methods to balance target variable
- Introduce other demographic datasets such as physical and mental traits
- Modify Neural Network
  - Introduce more hidden layers
  - Train for more epochs