Technical Assessment

Neil Menghani

Introduction

dataiku

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.

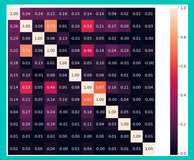
Steps:

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

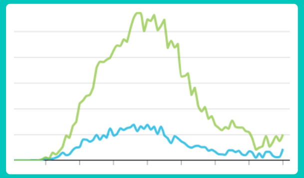
Methodology

Explore Data

Determine correlations between variables



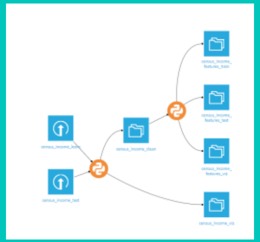
Graphically analyze trends in key variables

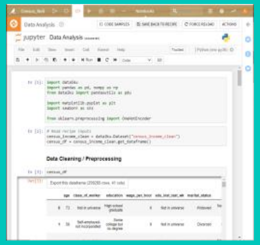


Prepare Data

Build a data pipeline for:

- Data Cleaning & Preprocessing
- Feature Engineering





Model

Build Machine Learning models for binary classification problem:

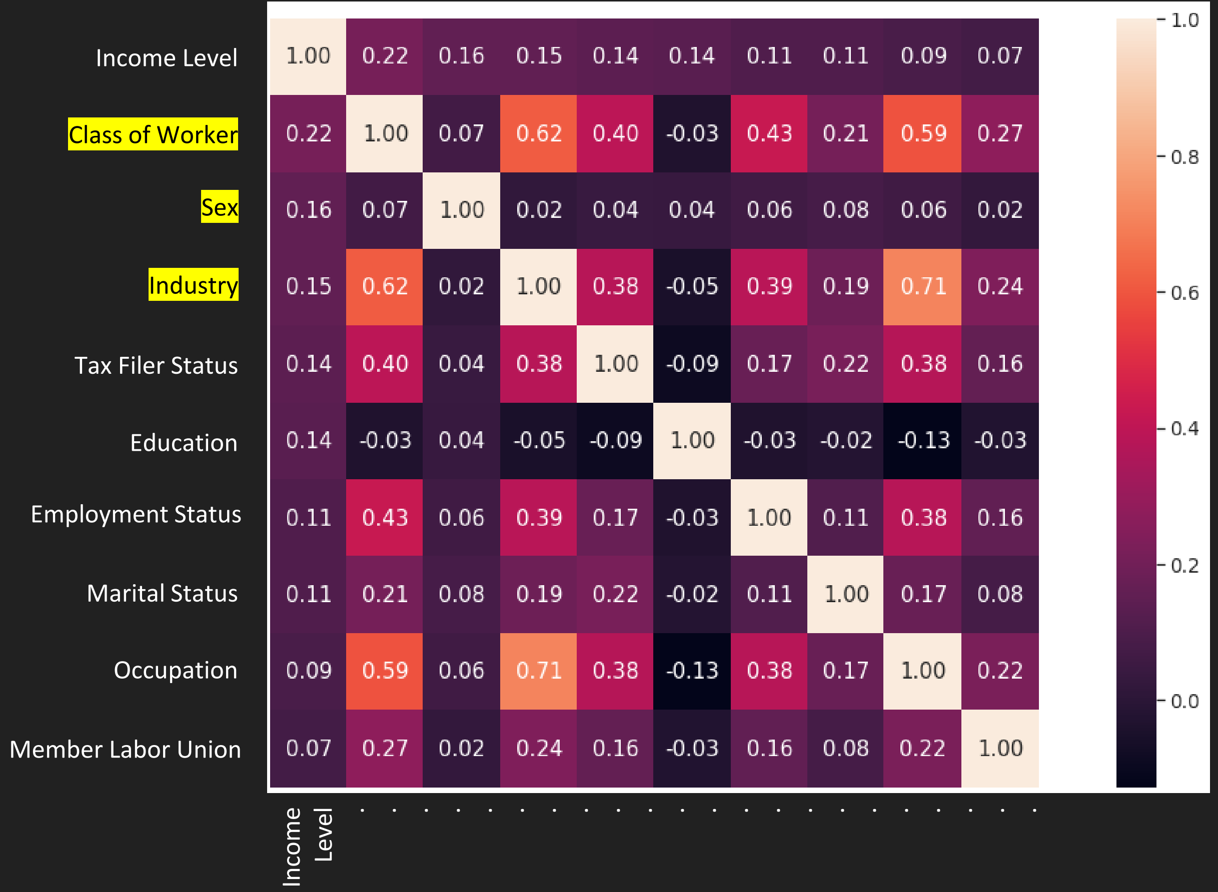
- Logistic Regression
- Tree-Based Models
- Neural Network

Evaluate models, extract feature importance, and draw conclusions

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

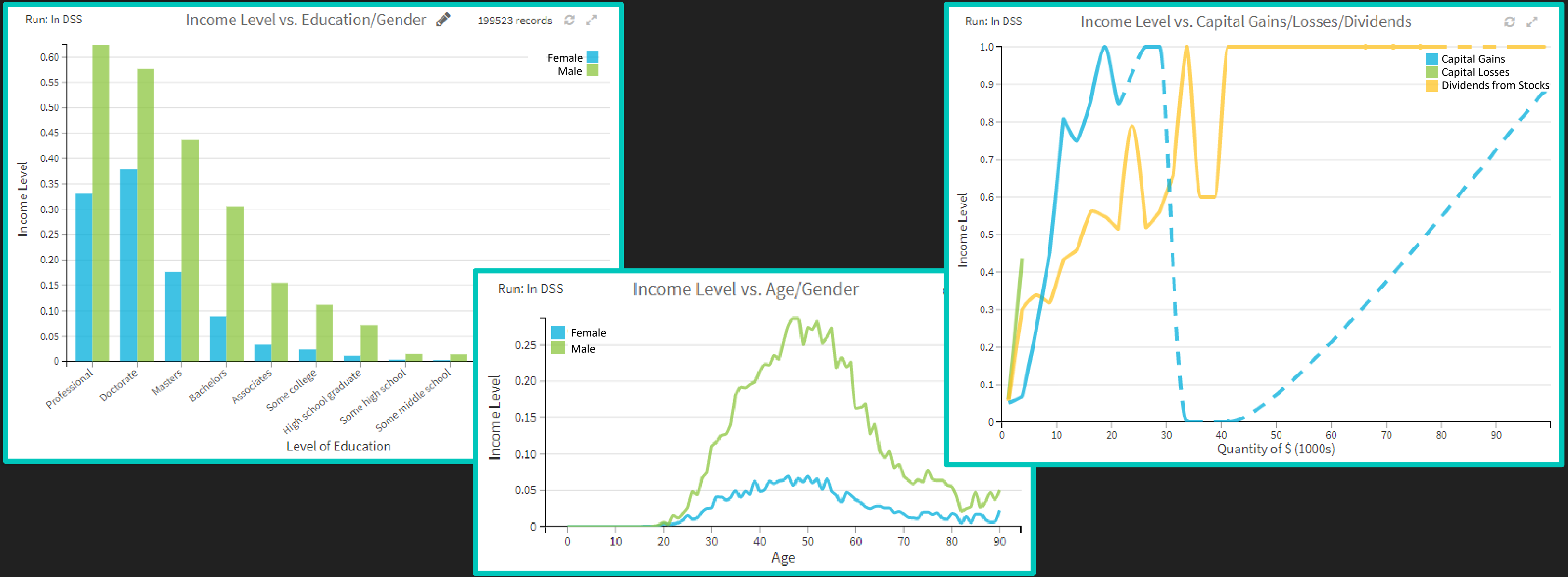
Exploratory Analysis

Correlation Matrix



Exploratory Analysis

Plots of Key Features



Data Preparation

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. 9th-12th 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

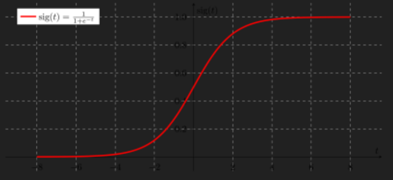
Machine Learning

Modeling Steps

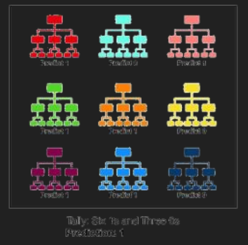
- Split the training set into train and validation sets (75/25 split)
  - Avoid learning on the test set
- Trained and tuned parameters for 3 types of machine learning models
  - Binary classification problem with target variable representing income above or below $50,000
- Generated test predictions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

Models Used

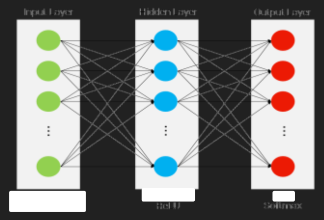
Logistic Regression



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



Neural Network (1 Hidden Layer)

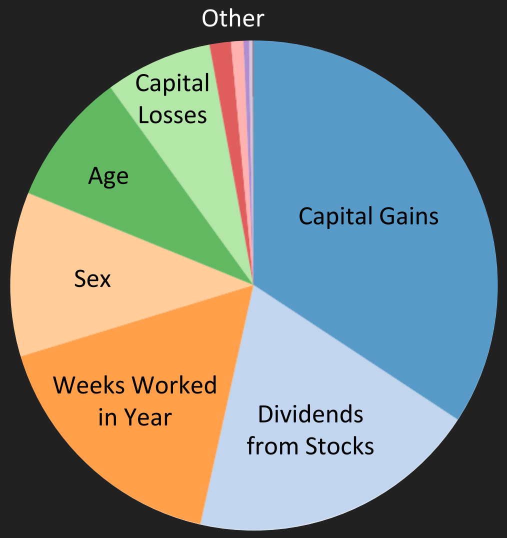


Model Evaluation

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

Feature Importance



Conclusions

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:
  - <u>Accuracy:</u> Gradient Boosted Ensemble Model
  - <u>AUC/F1-Score:</u> 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

Thank You!