Project 4

Final Presentation

Census Income Analysis

Analyzing Income Disparities Using Census Data

A Data Science Exploration of Gender, Education, and Work Impact on Income

Presented by

Conner Dekok, Quinn Daley, Mackenzie Deets, Sanjana Prabhakar

Context

Problem Statement:

- Income inequality is a critical global issue.
- Exploring socioeconomic factors like gender, education, and work helps inform policy decisions

Why the 1994 Census Dataset?

- Historical Value: Captures income patterns from a transformative decade
- Diversity: Rich dataset with over 32,000 records and 15 variables covering education, work, and demographics

Dataset:

- Number of Records: 32,561
- Target Variable: Income (>50K or <=50K)
- Key Features: Age, Workclass, Education, Gender, Marital Status, Hours per Week, etc

	Age	Workclass	Final Weight	Education	EducationNum	Marital Status	Occupation	Relationship	Race	Gender	Capital Gain	capital loss	Hours per Week	Native Country	Income
0	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States	<=50K
1	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States	<=50K
2	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States	<=50K
3	53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States	<=50K
4	28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba	<=50K
5	37	Private	284582	Masters	14	Married-civ-spouse	Exec-managerial	Wife	White	Female	0	0	40	United-States	<=50K
6	49	Private	160187	9th	5	Married-spouse-absent	Other-service	Not-in-family	Black	Female	0	0	16	Jamaica	<=50K

Data cleaning process

- Addressed missing values
- The Workclass column contained entries like “?”, which indicated missing data
- Filter out rows where 'Workclass' is 'never-worked' or 'without-pay'
- Converted entries in columns like Workclass to lowercase and stripped leading/trailing spaces
- Rare categories in workclass like “Never-worked and “wothout-pay” were combined into an “other” category

```
pd.unique(df_census_income["Workclass"])
```

```
array([' State-gov', ' Self-emp-not-inc', ' Private', ' Federal-gov',
       ' Local-gov', ' ?', ' Self-emp-inc', ' Without-pay',
       ' Never-worked'], dtype=object)
```

Workclass	count
Private	22696
Self-emp-not-inc	2541
Local-gov	2093
?	1836
State-gov	1298
Self-emp-inc	1116
Federal-gov	960
Without-pay	14
Never-worked	7

dtype: int64

Workclass	count
private	22696
self-emp-not-inc	2541
local-gov	2093
Unknown	1836
state-gov	1298
self-emp-inc	1116
federal-gov	960

dtype: int64

PySpark for Data Processing

Why PySpark?

- PySpark provided a scalable and efficient framework to handle our dataset
- Enabled seamless integration of distributed computing for data preprocessing

What We Did:

- Converted the dataset from Pandas to PySpark DataFrame to take advantage of distributed processing
- Separated the target variable (Income >50k) from the feature set for clearer analysis
- Preprocessed feature columns by encoding categorical variables into numerical formats
- Converted PySpark DataFrames back to Pandas for machine learning compatibility

Key Outcomes:

- Achieved efficient data handling with PySpark's distributed processing capabilities
- Simplified further transformations (e.g., one-hot encoding) using Pandas
- Ensured the dataset was ready for predictive modeling with clean and encoded features

Age	Workclass	Final Weight	Education Rank	Marital Status	Occupation	Relationship	Race	Gender	Hours per Week	Native Country
39	state-gov	77516	13	Never-married	Adm-clerical	Not-in-family	White	Male	40	United-States
50	self-emp-not-inc	83311	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	13	United-States
38	private	215646	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	40	United-States
53	private	234721	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	40	United-States
28	private	338409	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	40	Other
37	private	284582	14	Married-civ-spouse	Exec-managerial	Wife	White	Female	40	United-States
49	private	160187	5	Married-spouse-a...	Other-service	Not-in-family	Black	Female	16	Other
52	self-emp-not-inc	209642	9	Married-civ-spouse	Exec-managerial	Husband	White	Male	45	United-States
31	private	45781	14	Never-married	Prof-specialty	Not-in-family	White	Female	50	United-States
42	private	159449	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	40	United-States
37	private	280464	10	Married-civ-spouse	Exec-managerial	Husband	Black	Male	80	United-States
30	state-gov	141297	13	Married-civ-spouse	Prof-specialty	Husband	Asian-Pac-Islander	Male	40	Other
23	private	122272	13	Never-married	Adm-clerical	Own-child	White	Female	30	United-States
32	private	205019	12	Never-married	Sales	Not-in-family	Black	Male	50	United-States
40	private	121772	11	Married-civ-spouse	Craft-repair	Husband	Asian-Pac-Islander	Male	40	Other
34	private	245487	4	Married-civ-spouse	Transport-moving	Husband	Amer-Indian-Eskimo	Male	45	Other
25	self-emp-not-inc	176756	9	Never-married	Farming-fishing	Own-child	White	Male	35	United-States
32	private	186824	9	Never-married	Machine-op-inspct	Unmarried	White	Male	40	United-States
38	private	28887	7	Married-civ-spouse	Sales	Husband	White	Male	50	United-States
43	self-emp-not-inc	292175	14	Divorced	Exec-managerial	Unmarried	White	Female	45	United-States

only showing top 20 rows

	Age	Final Weight	Education Rank	Hours per Week	Workclass_Unknown	Workclass_federal-gov	Workclass_local-gov	Workclass_private	Workclass_self-emp-inc	Workclass_self-emp-not-inc	...	Relationship_Wife	Race_Amer-Indian-Eskimo	Race_Asian-Pac-Islander	Race_Black
0	39	77516	13	40	0	0	0	0	0	0	...	0	0	0	0
1	50	83311	13	13	0	0	0	0	0	1	...	0	0	0	0
2	38	215646	9	40	0	0	0	1	0	0	...	0	0	0	0
3	53	234721	7	40	0	0	0	1	0	0	...	0	0	0	1
4	28	338409	13	40	0	0	0	1	0	0	...	1	0	0	1
5	37	284582	14	40	0	0	0	1	0	0	...	1	0	0	0
6	49	160187	5	16	0	0	0	1	0	0	...	0	0	0	1
7	52	209642	9	45	0	0	0	0	0	1	...	0	0	0	0
8	31	45781	14	50	0	0	0	1	0	0	...	0	0	0	0
9	42	159449	13	40	0	0	0	1	0	0	...	0	0	0	0
10	37	280464	10	80	0	0	0	1	0	0	...	0	0	0	1
11	30	141297	13	40	0	0	0	0	0	0	...	0	0	1	0
12	23	122272	13	30	0	0	0	1	0	0	...	0	0	0	0
13	32	205019	12	50	0	0	0	1	0	0	...	0	0	0	1
14	40	121772	11	40	0	0	0	1	0	0	...	0	0	1	0
15	34	245487	4	45	0	0	0	1	0	0	...	0	1	0	0
16	25	176756	9	35	0	0	0	0	0	1	...	0	0	0	0
17	32	186824	9	40	0	0	0	1	0	0	...	0	0	0	0
18	38	28887	7	50	0	0	0	1	0	0	...	0	0	0	0
19	43	292175	14	45	0	0	0	0	0	1	...	0	0	0	0

Cleaned Data

```
Income Greater than 50k
0    24699
1     7841
Name: count, dtype: int64
```

The total values that are 0 are under $ 50,000 while the value of 1 is over $ 50,000.

	Age	Workclass	Final Weight	Education Rank	Marital Status	Occupation	Relationship	Race	Gender	Hours per Week	Native Country	Income Greater than 50k
0	39	state-gov	77516	13	Never-married	Adm-clerical	Not-in-family	White	Male	40	United-States	0
1	50	self-emp-not-inc	83311	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	13	United-States	0
2	38	private	215646	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	40	United-States	0
3	53	private	234721	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	40	United-States	0
4	28	private	338409	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	40	Other	0

This is the completed data frame after cleaning out the unneeded information along with adding a value of 1 or 0 correlation to reaching the $ 50,000 threshold.

Logistic Regression

Model Description:

- Logistic Regression is a simple yet effective classifier for binary outcomes
- Used here to predict whether income is greater than or less than 50K

Attempt 1: Baseline Model

- Approach: Random state 1
- Accuracy: 84%
- Observation: The accuracy of all models regardless of random state was 84%.

Attempt 2: Feature Selection

- Approach: Random state 2
- Accuracy: 84%

Attempt 1: Baseline Model

- Approach: Random state 7
- Accuracy: 84%

	precision	recall	f1-score	support
0	0.87	0.93	0.90	6130
1	0.72	0.56	0.63	2005
accuracy			0.84	8135
macro avg	0.79	0.74	0.76	8135
weighted avg	0.83	0.84	0.83	8135

K-Nearest Neighbors (KNeighborsClassifier)

Model Description:

- K-Nearest Neighbors (KNN) is a simple and effective algorithm that predicts the class of a data point based on the majority class of its nearest neighbors.
- Used to classify whether income exceeds $50K based on proximity in feature space.

Attempt 1: Baseline Model

- Approach: Neighbors number 6
- Accuracy: 82%
- Observation: The baseline model resulted in an 82% accuracy.

Attempt 2: Adjusted Number of Neighbors

- Approach: Neighbors number 8
- Accuracy: 82%
- Observation: The adjusted model resulted in an 82% accuracy.

Attempt 3: Weighted Neighbors

- Approach: Neighbors number 3
- Accuracy: 81%
- Observation: The weighted model resulted in an 81% accuracy.

	precision	recall	f1-score	support
0	0.93	0.84	0.89	6762
1	0.47	0.69	0.56	1373
accuracy			0.82	8135
macro avg	0.70	0.77	0.72	8135
weighted avg	0.85	0.82	0.83	8135

	precision	recall	f1-score	support
0	0.93	0.85	0.89	6658
1	0.51	0.69	0.59	1477
accuracy			0.82	8135
macro avg	0.72	0.77	0.74	8135
weighted avg	0.85	0.82	0.83	8135

	precision	recall	f1-score	support
0	0.89	0.86	0.87	6345
1	0.55	0.62	0.58	1790
accuracy			0.81	8135
macro avg	0.72	0.74	0.73	8135
weighted avg	0.81	0.81	0.81	8135

Random Forest Model

Model Description:

- The Random Forest is a non-parametric algorithm that splits the data based on feature values to make predictions
- Used here to classify whether income exceeds $50K

Attempt 1: Baseline Model
Approach: Estimator 1500, Random state 68
Accuracy: 83%
Observation: The accuracy started at baseline at 82%.

Confusion Matrix

	Predicted 0	Predicted 1
Actual 0	5607	523
Actual 1	860	1145

Accuracy Score : 0.8299938537185003
Classification Report

	precision	recall	f1-score	support
0	0.87	0.91	0.89	6130
1	0.69	0.57	0.62	2005
accuracy			0.83	8135
macro avg	0.78	0.74	0.76	8135
weighted avg	0.82	0.83	0.82	8135

Attempt 2: Depth Limitation
Approach: Estimator 200, Random state 68
Accuracy: 83%
Observation: Increasing the depth led to an 83% accuracy.

Confusion Matrix

	Predicted 0	Predicted 1
Actual 0	5599	531
Actual 1	856	1149

Accuracy Score : 0.8295021511985249
Classification Report

	precision	recall	f1-score	support
0	0.87	0.91	0.89	6130
1	0.68	0.57	0.62	2005
accuracy			0.83	8135
macro avg	0.78	0.74	0.76	8135
weighted avg	0.82	0.83	0.82	8135

Attempt 3: Optimized Splitting and Pruning
Approach: RandomForestClassifier(n_estimators=2000, max_depth=15,
min_samples_split=10,
min_samples_leaf=5,
random_state=68
Accuracy: 84%
Observation: Optimizing the data led to 84% accuracy.

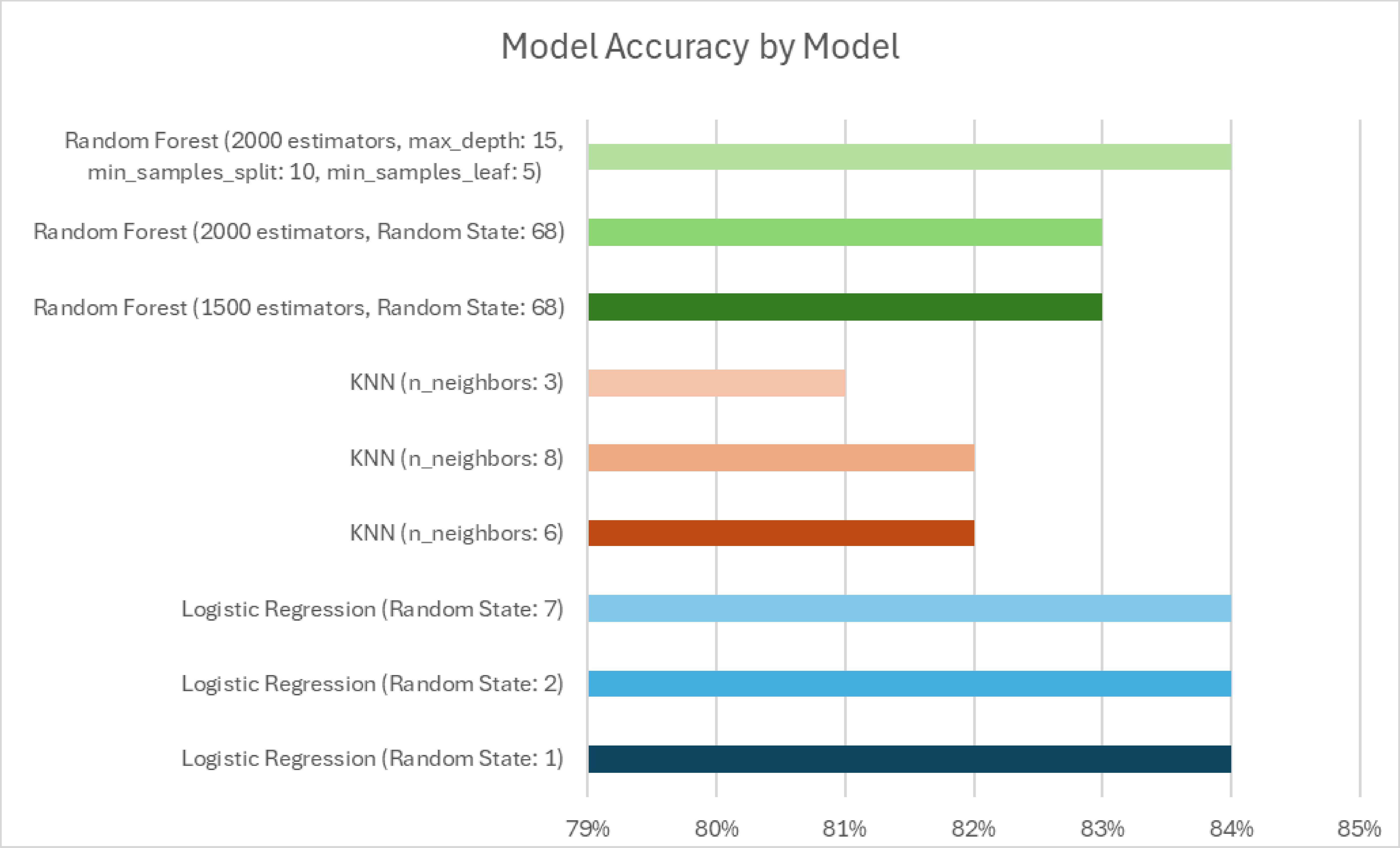
Confusion Matrix

	Predicted 0	Predicted 1
Actual 0	5769	361
Actual 1	933	1072

Accuracy Score : 0.8409342347879533
Classification Report

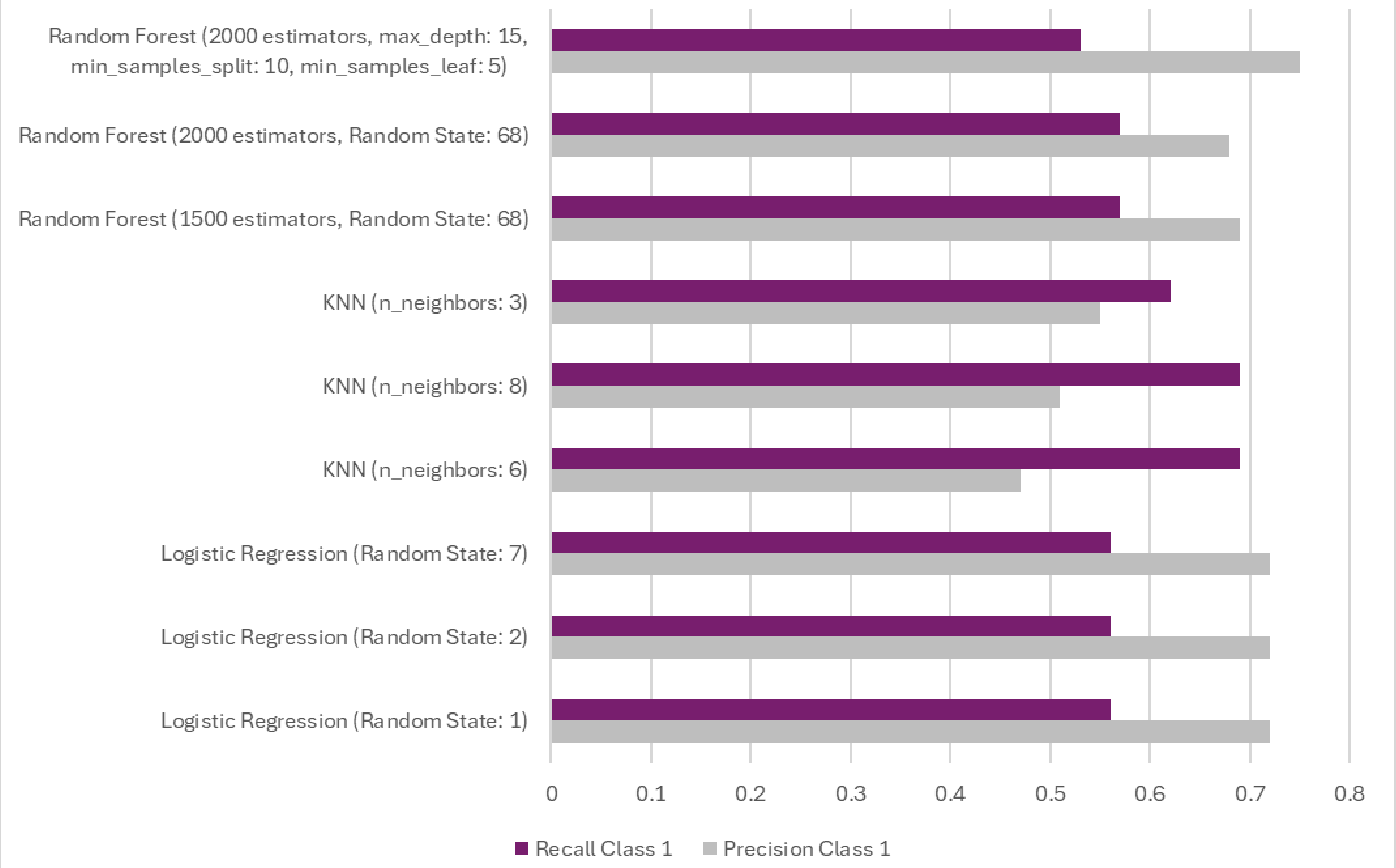
	precision	recall	f1-score	support
0	0.86	0.94	0.90	6130
1	0.75	0.53	0.62	2005
accuracy			0.84	8135
macro avg	0.80	0.74	0.76	8135
weighted avg	0.83	0.84	0.83	8135

Key Findings



Recall and Precision of Class 0 (Individuals making <$50k/year)

Recall and Precision of Class 1 (Individuals making =>$50k/year)



Best Performing Model

Accuracy

84%

Precision

Class 0: 86%
Class 1: 75%

Recall

Class 0: 94%
Class 1: 53%

Best-Performing Model (possibly...):

- The optimized Random Forest model with hyperparameters such as 2000 estimators, a maximum depth of 15, and adjusted minimum samples split and leaf values demonstrated superior performance. This model achieved the best precision for predicting class 1 while maintaining a good balance of precision and recall for class 0 as well as the highest overall accuracy.

Conclusion

Recap of Findings:

- **Income is influenced by multiple factors, with education level and gender being the most significant**
- **Work hours also play a role but have a lesser impact compared to other variables**

Importance of Addressing Income Disparity:

- **Highlights the need for gender equity and access to education to reduce income inequality**

Next Steps:

- **Suggest further analysis to explore causation rather than just correlation.**
- **Address bias or sampling issues in the dataset**

Thank you

For your attention