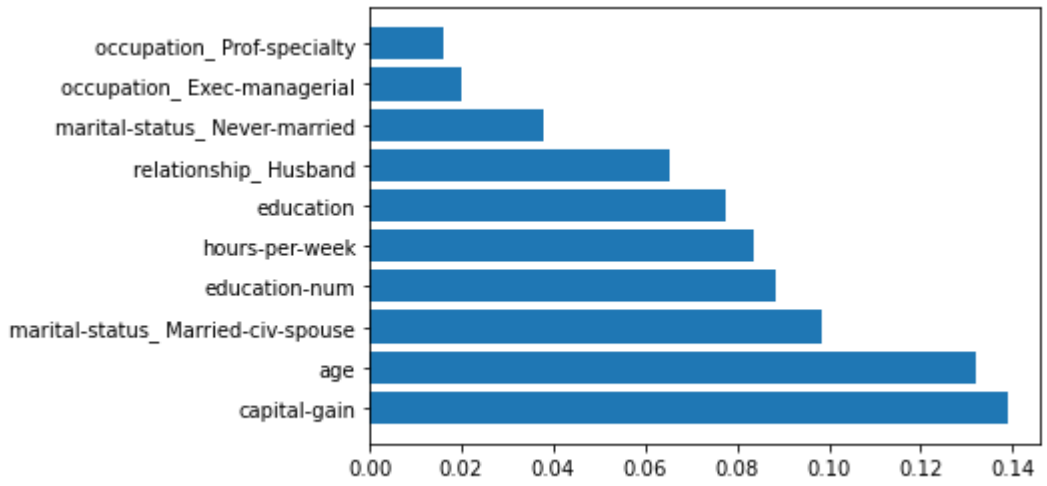
SALARY PREDICTOR PROJECT

	age	workclass	fnlwgt	education	education-num	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	hours-per-week	native-country	salary
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
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
32556	27	Private	257302	Assoc-acdm	12	Married-civ-spouse	Tech-support	Wife	White	Female	0	0	38	United-States	<=50K
32557	40	Private	154374	HS-grad	9	Married-civ-spouse	Machine-op-inspct	Husband	White	Male	0	0	40	United-States	>50K
32558	58	Private	151910	HS-grad	9	Widowed	Adm-clerical	Unmarried	White	Female	0	0	40	United-States	<=50K
32559	22	Private	201490	HS-grad	9	Never-married	Adm-clerical	Own-child	White	Male	0	0	20	United-States	<=50K
32560	52	Self-emp-inc	287927	HS-grad	9	Married-civ-spouse	Exec-managerial	Wife	White	Female	15024	0	40	United-States	>50K

The dataset

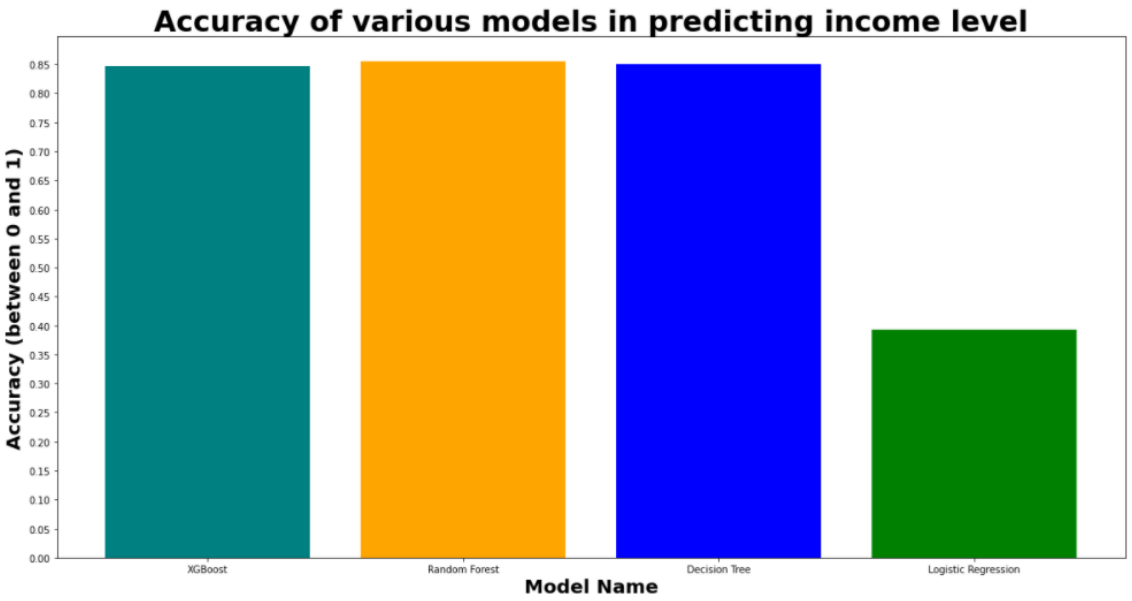
The dataset that I used contains various attributes about adults living across the world such as their age, education level, occupation, race, and marital status. The problem statement was what factors are most important in predicting the salary level of an individual.

I first got a summary of the dataset using the head and tail features. Once I did that, I checked whether there are any missing values for any of the columns in the dataset. I found out that the dataset contained no missing values, so there was no need for data cleaning. I then used the get_dummies feature on columns that contained non-numeric and non-orderable values. I then used the mapping feature in order to map education level in ascending order (pre-school was mapped to 0, and professional school was mapped to 14). I also mapped the salary feature. If a person's salary was above 50k, it was mapped to a value of 1. If a person's salary was below 50k, it was mapped to a value of 0. I then used MinMaxScaler to ensure that all the values are between 0 and 1 . I then removed duplicate rows as well as rows in which values for any particular column were more than 3 standard deviations away from the mean of that particular column. I then used feature importances in order to find out the 5 best predictors of a person's income level.



I found out that the 5 best predictors of a person's income level are their capital gain, their age, their marital status, the number of hours per week they work and their level of education respectively.

I then split the dataset into training data and testing data. I then instantiated the models to use in order to predict a person's income level. The models I used were decision trees, random forest, XGBoost, and logistic regression. I trained these models using the training dataset. In the case of the XG Boost model, I played around with the parameters and found out that the model performs optimally when the column sample by tree value is 0.5, the learning rate is 0.15, the maximum depth is 5, alpha is 10, and the number of estimators is 50. In the case of the Decision Tree classifier, I found out that the model performs the best when the criterion is gini, the splitter value is best, and when the maximum depth is 10. I also noticed that there is a significant decrease in performance when the splitter value is changed from best to random and when the maximum depth is increased from 10 to a three digit number. In the case of the random forest classifier, I found out that the model performs the best when the bootstrap value is true, the criterion value is entropy, the minimum impurity decrease value is 0, the minimum number of leaf samples is 10, the minimum sample split is 10, the minimum fraction leaf weight is 0, the number of estimators is 10, njobs is 1, and verbose is 0. Additionally, I found out that the model's performance significantly decreases when the minimum fraction leaf weight value and the minimum impurity decrease value changes. After training, each model predicted whether a person's salary was above 50k a year or not for every individual in the testing dataset. Then the accuracy of each model was determined by comparing the predicted salary level for every individual in the testing dataset with the actual salary level for every individual in the testing dataset. I also found various features of interest such as the Recall Score, the F1 Score, the Precision Score, and the Confusion Matrix for the XGBoost model.



The Random Forest classifier had the highest accuracy out of all the models at 85.51%, followed by the Decision Tree Classifier which had an accuracy of 85.1% , followed by the XGBoost model which had an accuracy of 84.63%, followed by the logistic regression model with an accuracy of 39.31%.