Analysis Predicting Voting Decisions Using K-Nearest Neighbours

Author: Rohan Chawla

This analysis aimed to determine whether individuals had already made a voting decision or remained undecided based on their income, education level, and marital status. The dataset included demographic and economic attributes, with the dependent variable, Vote, categorized as either decided or Undecided. Boxplot was created to visually explore how income and education related to voting decisions. These plots indicated differences between the two voter groups: individuals who had already decided generally had higher average incomes(*Graph1.2)*, whereas undecided voters tended to have slightly higher education levels(*Graph1.3*). Such insights helped confirm the importance of these variables and guided the selection of predictors for the predictive model.

There was an imbalance in the data; more voters had already decided. So, SMOTE was applied. This allowed the model to learn equally well from both groups, resulting in more accurate predictions. The chosen algorithm, K-Nearest Neighbours (KNN), classifies new voters based on similarities to previously observed individuals. After training the model and evaluating its performance(*Table 1.1)* through cross-validation, the KNN model achieved an overall accuracy of **81.3%**, correctly identifying most voters' decision statuses. The Area Under the ROC Curve (AUC), a measure of the model's ability to distinguish between the two groups, was **0.858**, indicating strong performance. The ROC curve (Graph1.1)confirmed that the model was significantly better than random guessing. A confusion matrix *(Table1.2)* clarified the results, showing that the model correctly identified **1,111 decided voters** and **503 undecided voters. However,** some misclassifications occurred, with **300 decided voters misclassified as undecided** and **87 undecided voters misclassified as agreed**. One of the key strengths was the model's high recall value (**93%**) for undecided voters, indicating a particularly effective ability to identify individuals who had not yet made a decision.

In conclusion, this analysis demonstrates that demographic factors such as income, education, and marital status influence voter decisions. Boxplots were instrumental in visually confirming these relationships, guiding variable selection, and strengthening the predictive model. The KNN model's effectiveness in identifying undecided voters can be particularly valuable for political campaigns, policymakers, and organizations seeking to engage voters who have not yet decided.

Table1.1 Table showing performance metrics

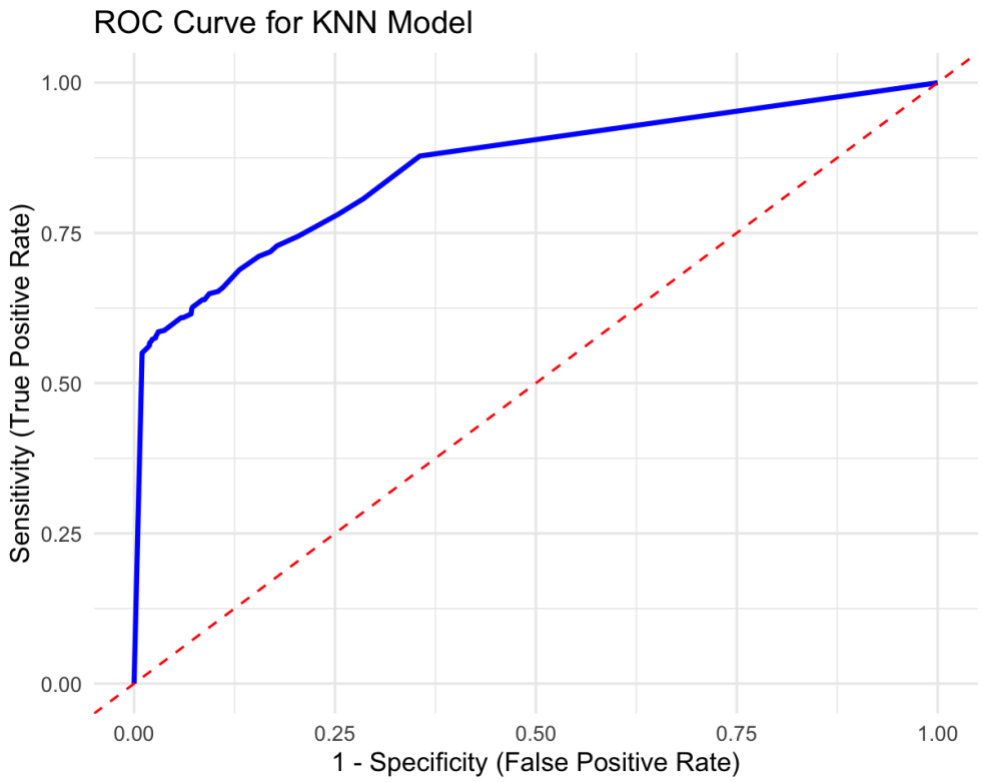
Metric	Estimate
Accuracy	0.813
Precision	0.794
Recall	0.93
ROC AUC	0.858
Sensitivity	0.93
Specificity	0.639

Table1.2 Table showing the confusion matrix

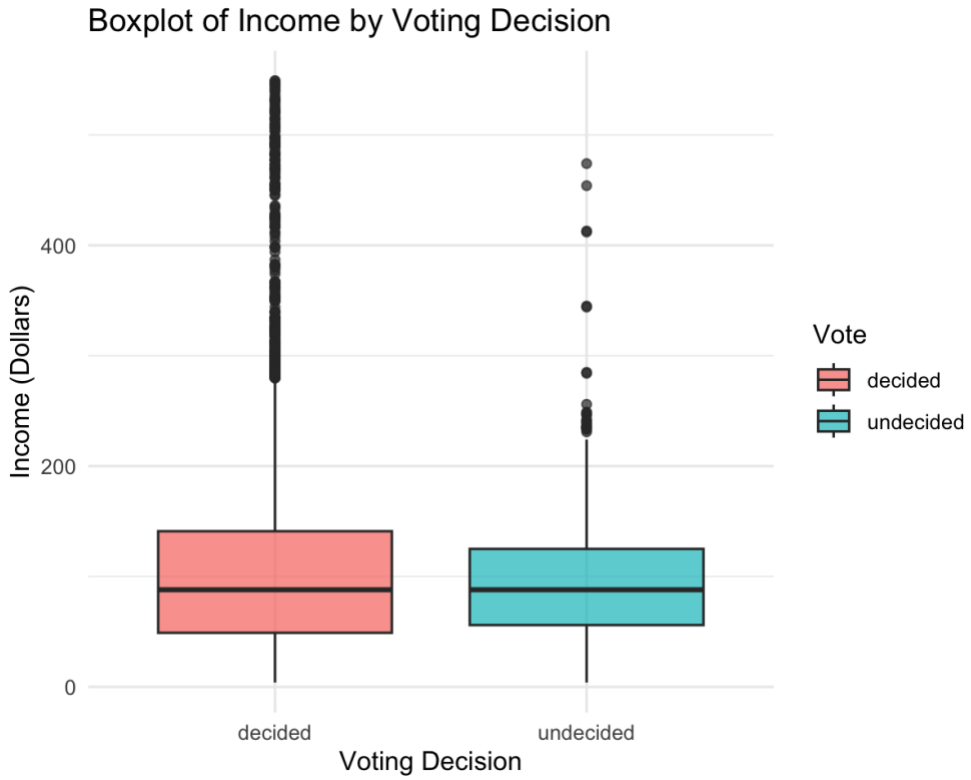
Prediction	Decided	Undecided
Decided	1111	300
Undecided	87	503

Reference:
https://chatgpt.com/#

Graph1.1 ROC curve for KNN Model

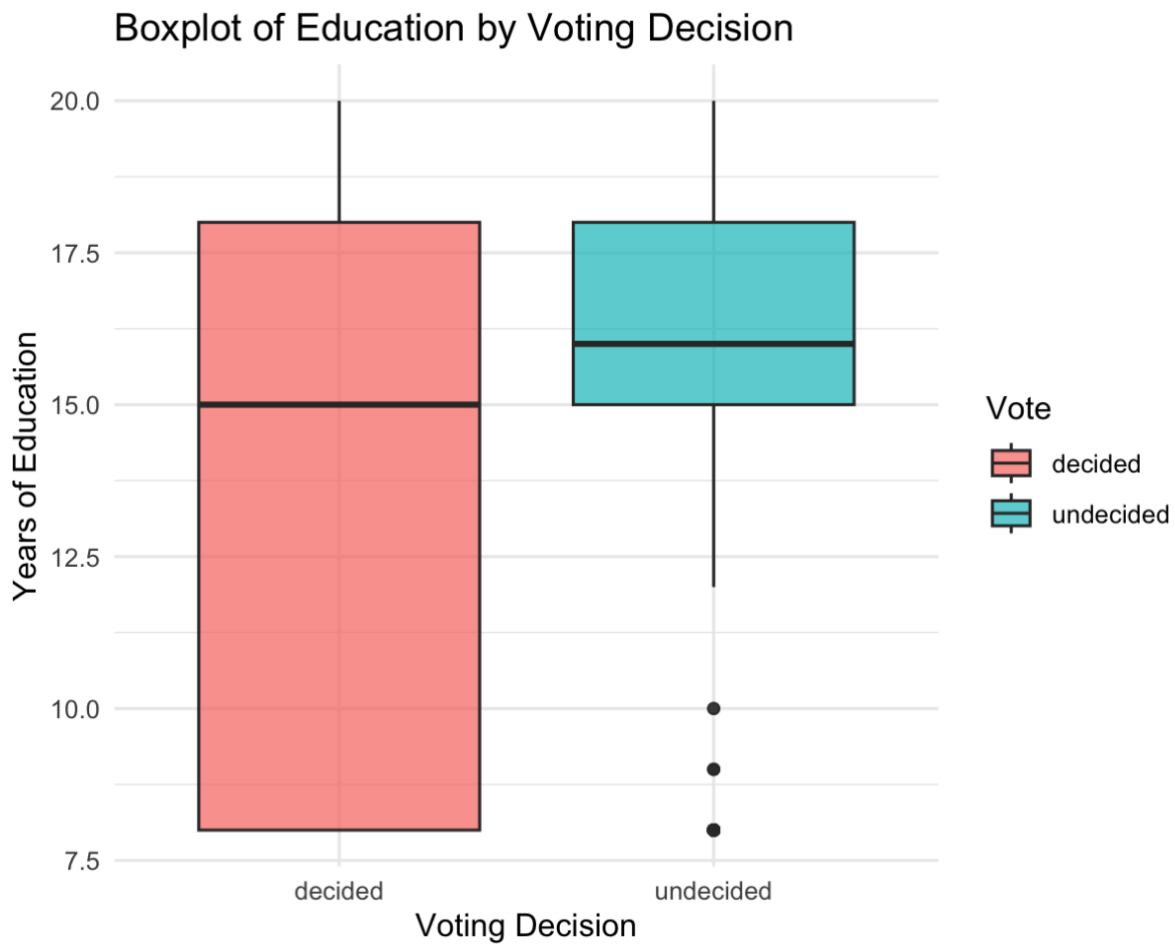


Graph1.2 Boxplot showing the relationship between Income(Dollars) and Voting decision



Reference:
https://chatgpt.com/#

Graph1.3 Boxplot showing the relationship between Education and Voting decision



Reference:
https://chatgpt.com/#