Maddie Marchionna

Yana Pathak

Prajusha Reddy

Urvi Sharma

Debt Repayment Prediction

[Link to code](https://colab.research.google.com/drive/1JgmJ7PC1yE3MfKeCymD1PyWWkeQ9feSw?usp=sharing)

**Introduction**

Seeing if people are more likely to repay their debt, or if they are good credit card holders, is based on historical data and criteria provided by them, such as credit score or payment history, to predict the probability of future mishaps. But, these rules are pretty rigid and do not take other important factors into account. Past models also do not account for economic changes and fluctuations, and therefore lose some predictive power and credibility.

Logistic model is a way that banks use for credit scoring. Logistic is helpful for binary classification, and also to calculate coefficients of every feature. With the development of machine learning algorithms, there are better predictive models that can be used for credit card scoring, such as Decision Trees and others.

We are trying to solve the problem of figuring out what criteria are most helpful in predicting whether or not a person will repay their debt and whether or not they are in good standing as a credit-card holder by using the decision tree and k-nearest neighbors algorithms.

We used two tables for our data that display credit record and application record. The application record contains the applicants’ personal information that we used for predicting, and the credit record has data for the users’ behaviors with their credit card. The data on the credit record table consists of how many months of data there are, and the status on whether or not payments are overdue, as well as how overdue they are (status). The data included on the application record table are gender, car ownership, property ownership, amount of children, annual income, income category, education level, marital status, way of living, birthday, days of employment, mobile phone, work phone, occupation, and family size. For this project, we focused on car ownership, property ownership, annual income, and birthday.

**Methodology**

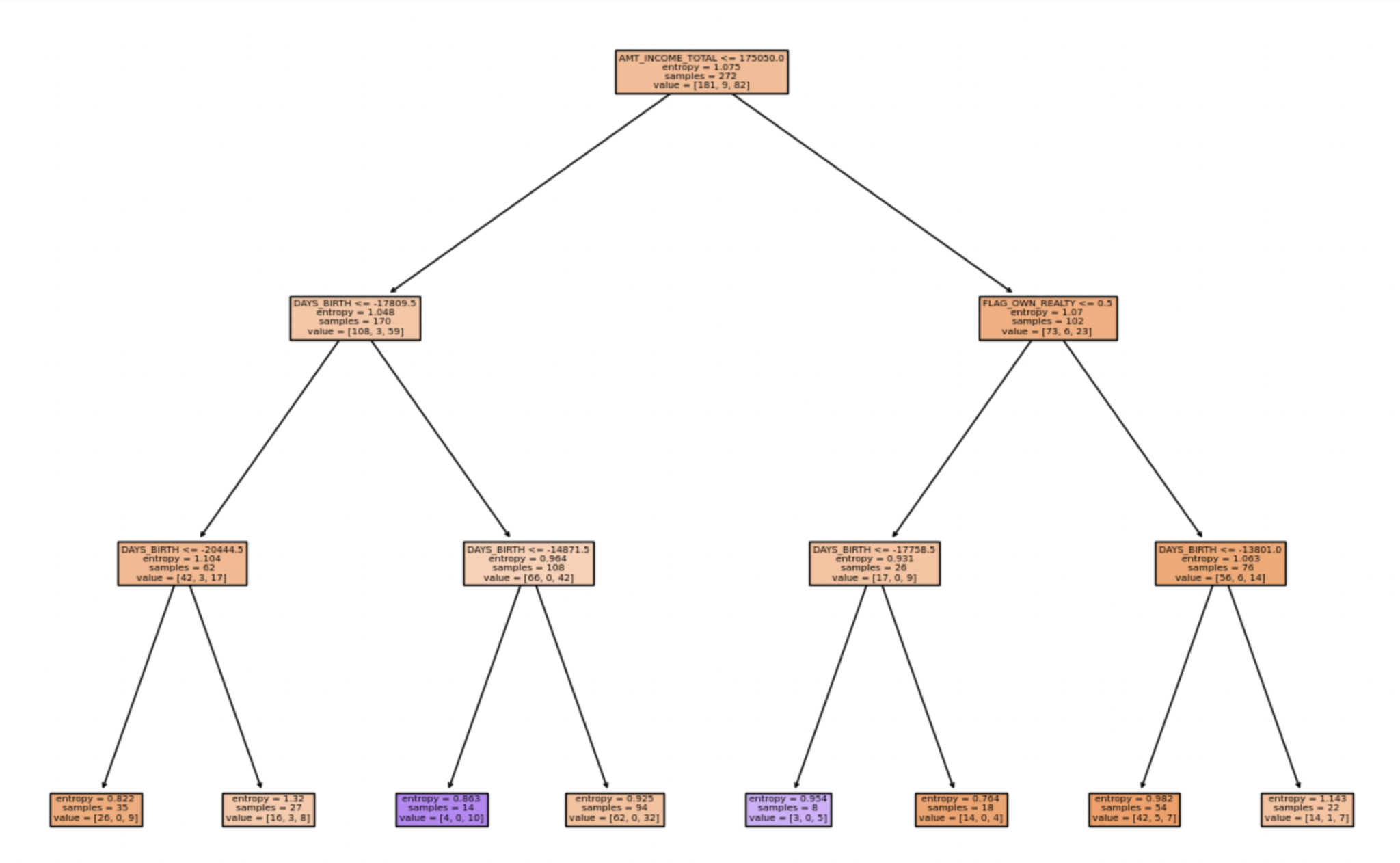
After collecting the datasets of past credit card applications and their corresponding approval or rejection outcomes, we preprocessed the data regarding missing values, assigning qualitative data numerical values, and scaling numerical values to a common range. Then, we split the dataset into training and testing sets, and set aside a proportion of the data for evaluating the performance of the training model.

After preprocessing and splitting the data, we applied machine learning algorithms, specifically decision trees and k-nearest neighbors. For the decision tree algorithm, we trained the data using a specified set of parameters. These parameters were a maximum depth of 3, a minimum samples per leaf of 8, and a random state of 110. These parameters were chosen based on the performance of the decision tree in the testing data. Decision trees were used to create a model that predicted whether a person would repay their debt based on a set of rules and criteria. K-NN was helpful for this task as it found the k-nearest neighbors to a given data point in the training data, and classified this data point based on the most common class of its nearest neighbors.

After training the predictive model, it was applied to new credit card applications in order to predict the likelihood of approval or rejection. The performance could also be evaluated on a validation set, and the model could be refined by adjusting the parameters as needed.

**Results**

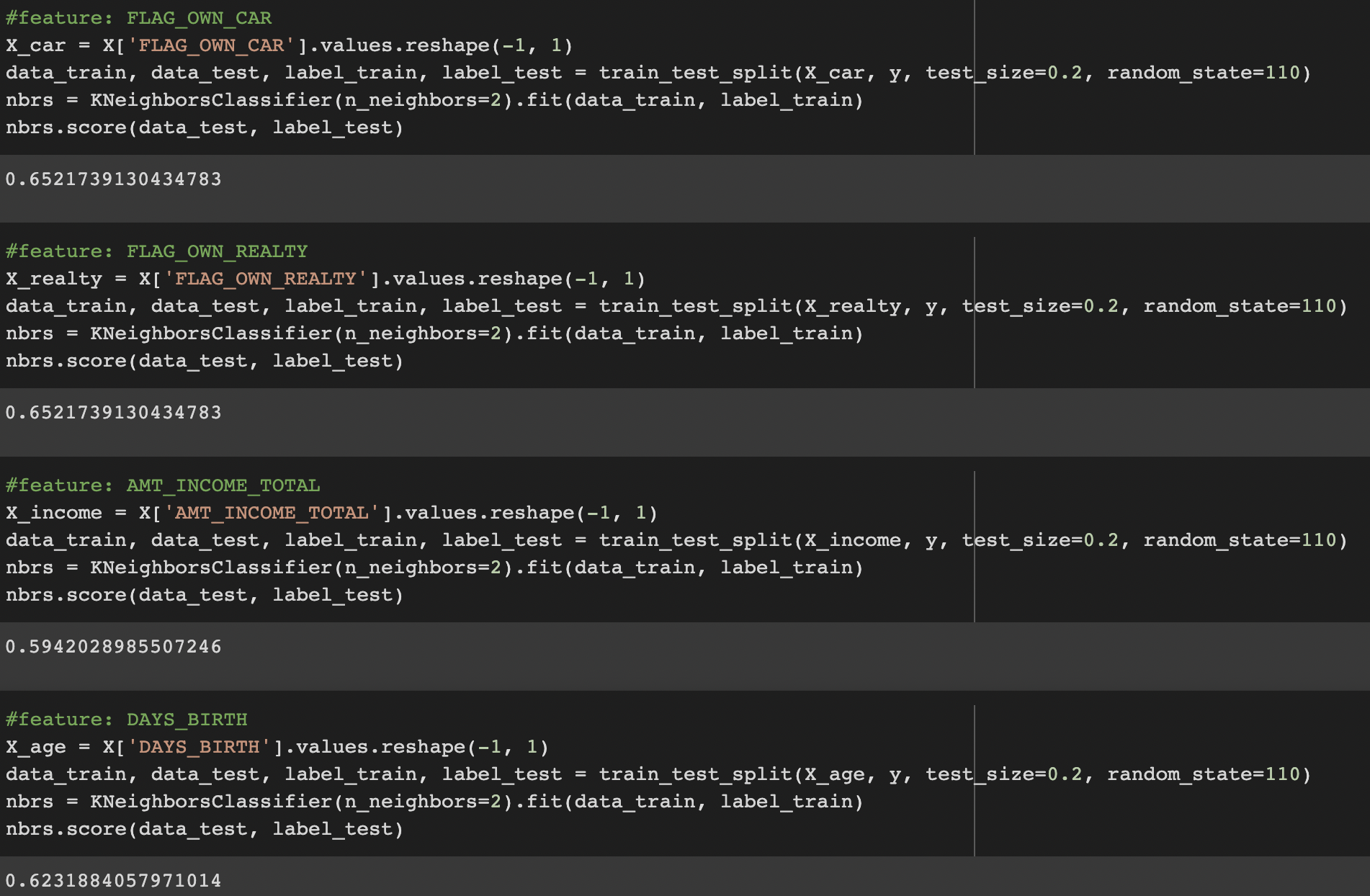
Decision Tree

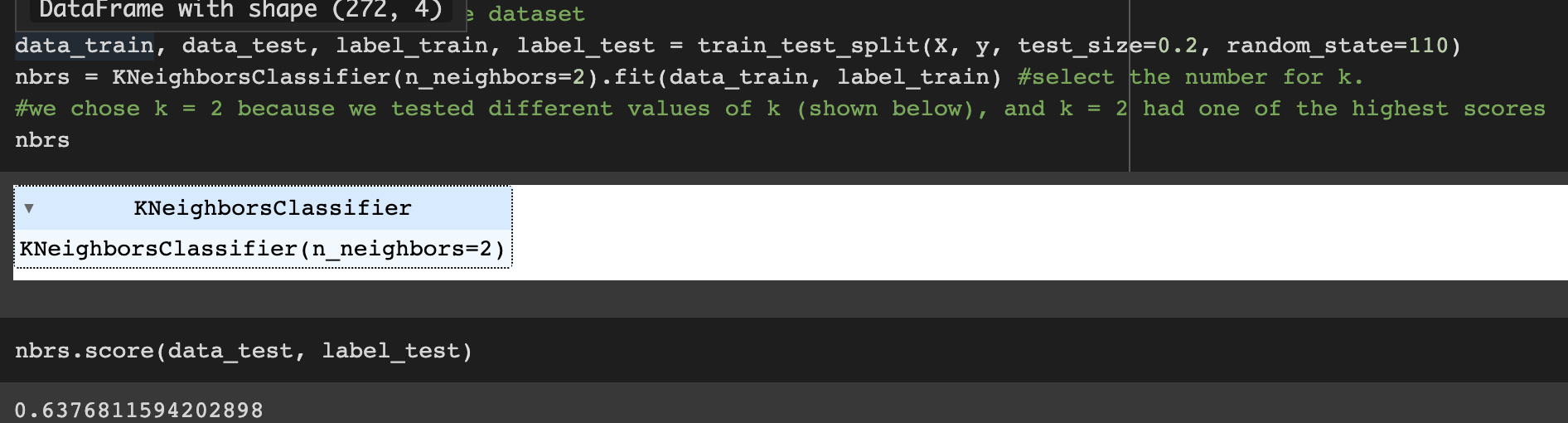


Shown above is our decision tree. The orange shaded nodes represent the nodes in which a client is more likely to repay their debt, whereas the purple nodes represent a client that is not likely to repay their debt. The darker nodes show that they are not as homogenous as the lighter ones, meaning that they lean more towards one classification. The decision tree shows that AMT\_INCOME\_TOTAL, which represents annual income, is the most important predictor for debt repayment; it is the root node from which the data begins to split. Then it is shown that the left child node further splits on DAYS\_BIRTH, which is the person’s age; it is also an important predictor for debt repayment. As you can see, for the values less than the given value in the node (meaning people who are older than that certain age, given that the values are negative since they represent how many days ago a person was born), it is more likely that the person will repay their debt. The right child node splits on FLAG\_OWN\_REALTY, which is whether or not the person owns property. It shows that people who own property are more likely to repay their debt. Overall, the decision tree shows that annual income is the most important predictor of debt repayment, and a person’s age and whether or not they have property are also important predictors of debt repayment. Our decision tree has a score of 0.6376, which means that our model correctly classifies 63.8% of our data.

K- Nearest Neighbors

We used the k-nearest neighbors algorithm on each feature to see which one had the highest score.



We chose a k value of 2, since it is what gave us the highest score after experimenting with other values. The number of nearest neighbors was the same for each feature since it would be less biased that way. As shown above, when we ran the k-nn model on the different features, the features that returned the highest scores were FLAG\_OWN\_CAR and FLAG\_OWN\_REALTY, which are whether or not a person owns a car and whether or not a person owns property. The score for both of them was the same: 0.6522, meaning a 65.2% accuracy in predicting the correct classification. Surprisingly, annual income (AMT\_INCOME\_TOTAL) had the lowest score of 0.5942, or 59.4% accuracy, which is the opposite of the decision tree. DAYS\_BIRTH, or age, returned a score of 0.6232, or 62.3% accuracy. Along with using the k-nn model on each feature, we ran it on the whole dataset.

Running it on the whole dataset returned a score of 0.6377, or 63.8% accuracy, which is the same as the decision tree’s score.

**Conclusions**

Through both the decision tree and k-nearest neighbors, we can conclude that whether or not a person owns property and their age are both important predictors of debt repayment. The decision tree shows that annual income was the most important predictor, and k-nn showed that whether or not someone owns a car or property are the most important predictors. Perhaps there are other underlying factors which did not lead us to get higher accuracy scores; maybe it was since our data was large.

Through this project, we were able to experiment with the decision tree and k-nearest neighbors algorithms to help us predict debt repayment outcomes, and see which features are the most useful predictors. Along with that, we were able to learn new techniques to manipulate and clean our very large dataset before running the two models on the datasets.