INTRODUCTION TO PROBLEM

In this project, I will predict whether someone is approved for a loan based on financial and personal data. I am interested in seeing what factors heavily affect the prediction of getting accepted for a loan. These factors can include the length of credit history a person has or previous loan defaults. I found this interesting because a majority of people today have to take out a loan, whether it is for college or to start a new business.

- **What factors are most important in predicting loan approval?**

INTRODUCTION TO DATASET

https://www.kaggle.com/datasets/lorenzozoppelletto/financial-risk-for-loan-approval?resource=download

The dataset that is being used is from Kaggle and titled “Financial Risk For Loan Approval” by Lorenzo Zoppelltto. It has 20,000 records of financial and personal data to help determine if someone is approved for a loan. It also contains a Risk Score Regression but this will be cleaned from the dataset during the preprocessing step. This dataset contains 36 features/columns. This dataset is very recent as it was last modified on September 3rd of this year.

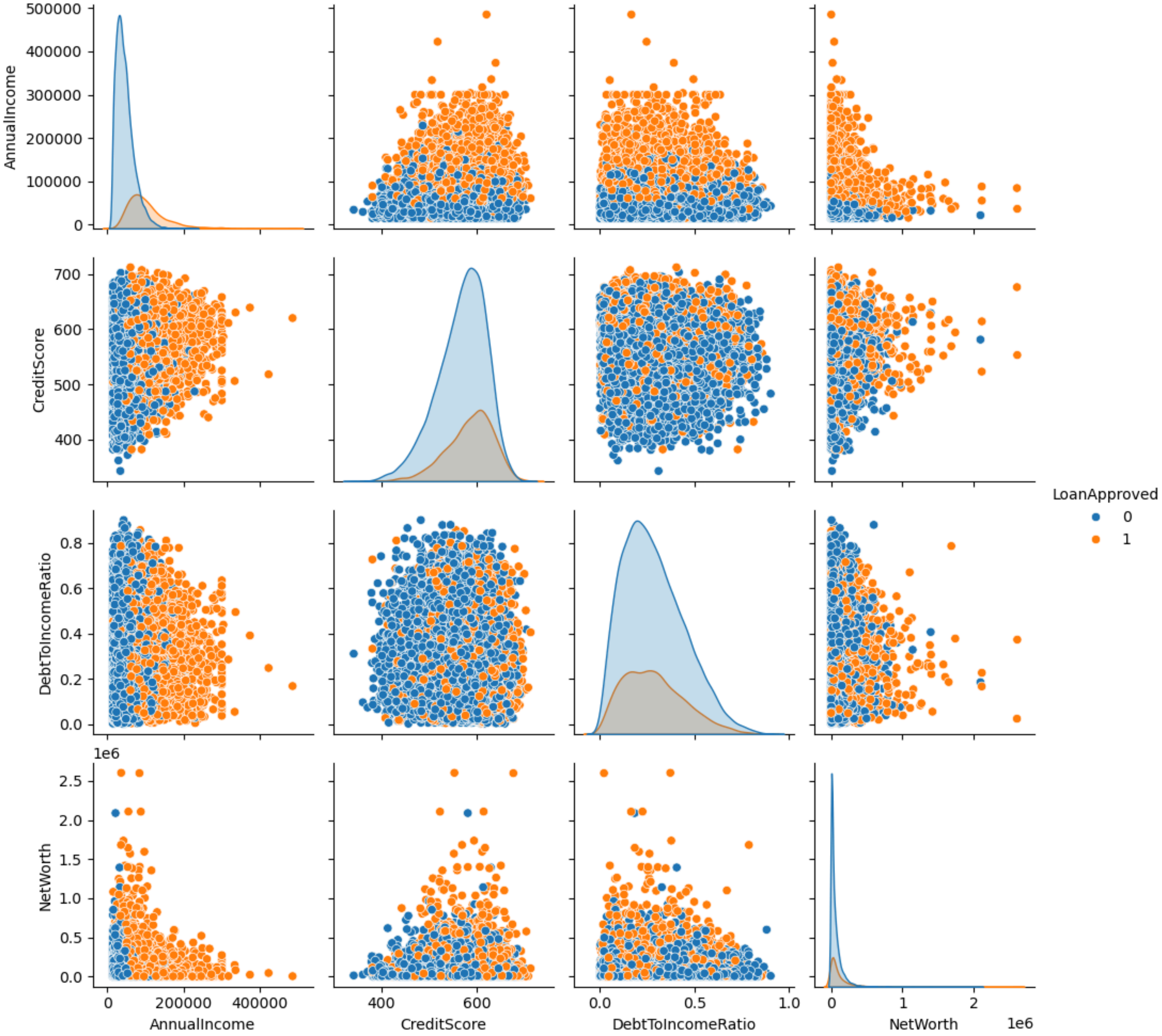
PREPROCESSING

This dataset was originally very clean. The first step I took was to drop the AppicationDate column. This column was dropped because it provided no relevant information and the majority of the data points were null. The next step I took was to drop the RiskScore column. This column was dropped because you could determine whether someone was approved for a loan solely based on this column. If the risk score was above 46.5 they were not approved. This would skew my data if left in and would not allow me to determine what other factors are contributing to the loan approval.

Next, I utilized the *pandas.get_dummies* function from the Pandas library in Python. This function allowed me to convert the columns with categorical data to numerical data. Doing this would allow the machine-learning algorithm to work more efficiently and accurately. The following columns were changed: LoanPurpose, HomeOwnershipStatus, MaritalStatus, EducationLevel, and EmploymentStatus.

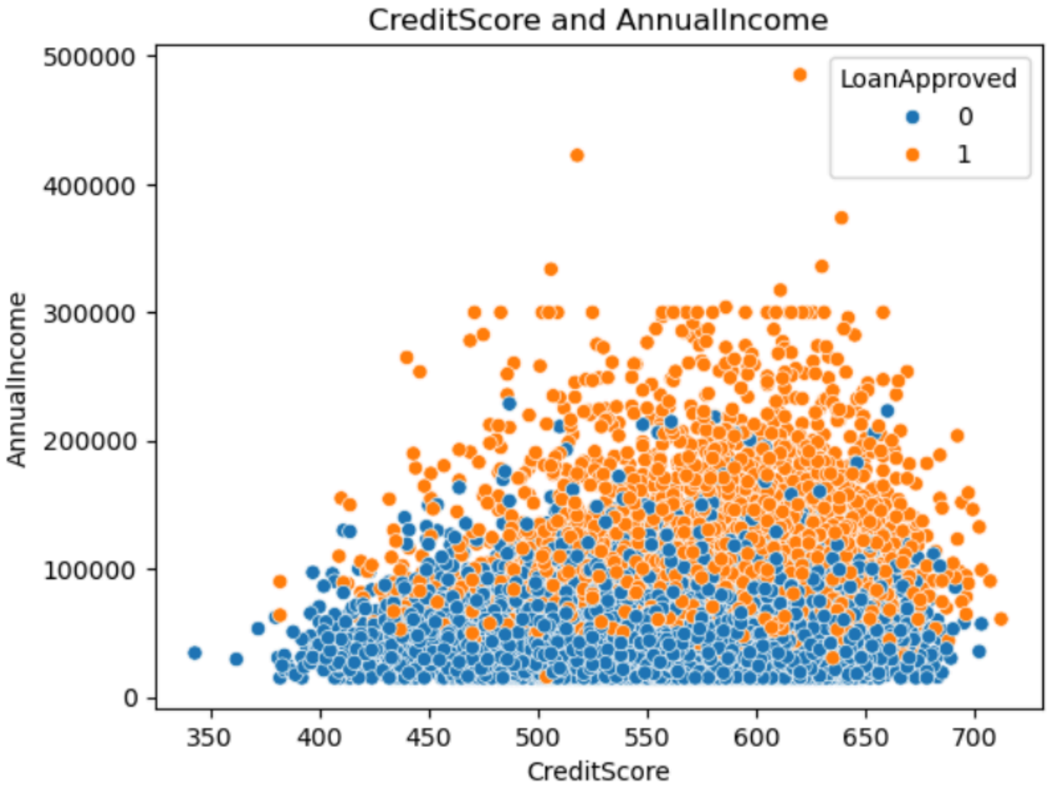
DATA VISUALIZATION

Fig 1.



The plairplots above show how the following 4 features relate to each other: AnnualIncome, CreditScore, DebtToIncomeRatio, and NetWorth. As you can see the data is very scattered and cannot be used to draw any conclusions.

Fig 2.



The scatterplot above shows the relationship between Credit Score and Annual Income. It also provides information about loan approval status. As you can see, the majority of approved loans are concentrated where credit scores and annual income are increasing. This could provide a correlation between these two variables. You can also see that when Annual Income is very low there are rarely any instances where loan approval gets accepted.

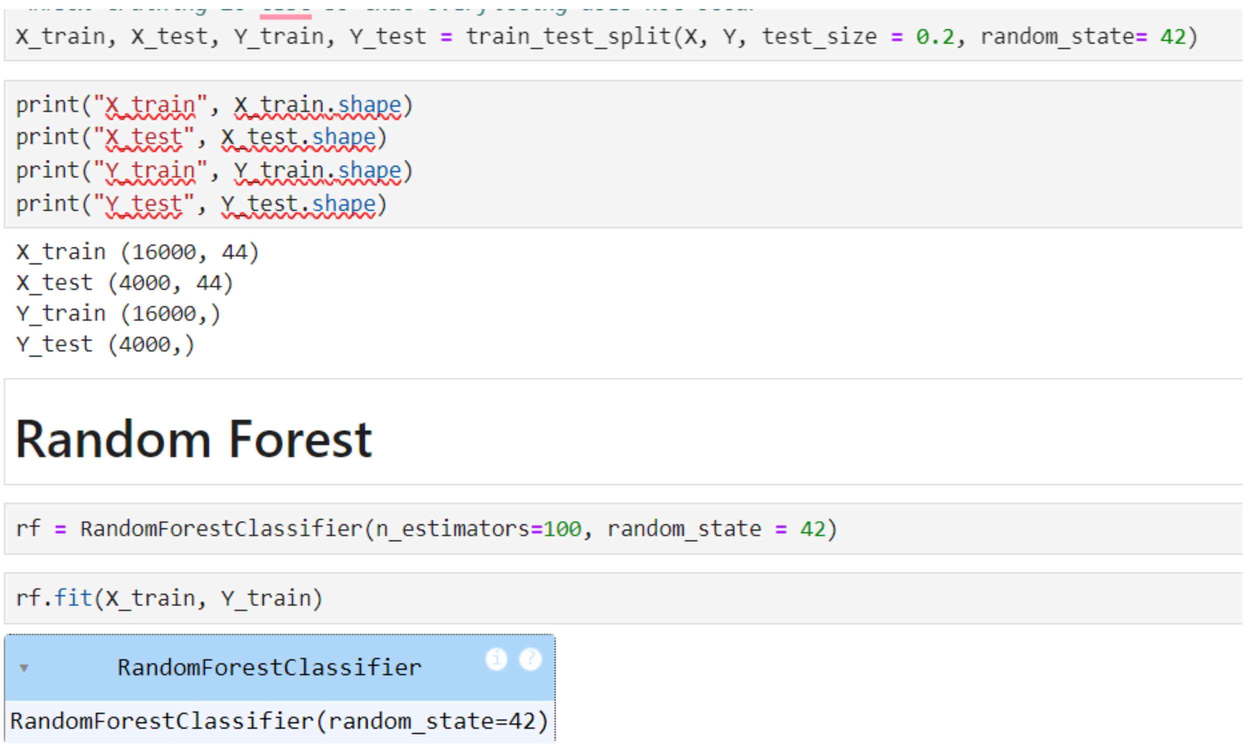
I found it interesting how data points with a high credit score but low annual income were not approved for a loan. Based on this I believe that the amount of Annual Income someone has will be very important when it comes to predicting outcomes of loan approvals.

MODELING

The model that I used was the Random Forest Classifier. I picked this model because a big part of the inquiries I had was looking at what factors impact the prediction of getting accepted for a loan.

Random forest is a supervised machine learning algorithm that uses decision trees to predict data. Since this algorithm uses decision trees it can also be used to determine which features are important when making a prediction. Each tree gives a vote/classification and then the forest will choose the classification with the most votes. It is less prone to overfitting and can handle missing values while still being accurate. The random forest also works well with datasets with a lot of data. This model works best for my dataset considering that it contains 20,000 entries and my inquiries.

The Random forest classifier can be slow and difficult to interpret. compared to other algorithms. The trees were randomly trained using Bootstrap Aggregation.



```
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size = 0.2, random_state= 42)
```

```
print("X_train", X_train.shape)
print("X_test", X_test.shape)
print("Y_train", Y_train.shape)
print("Y_test", Y_test.shape)
```

```
X_train (16000, 44)
X_test (4000, 44)
Y_train (16000,)
Y_test (4000,)
```

Random Forest

```
rf = RandomForestClassifier(n_estimators=100, random_state = 42)
```

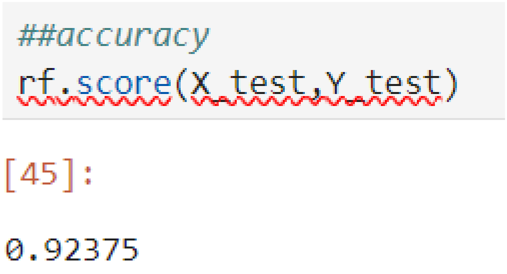
```
rf.fit(X_train, Y_train)
```

```
▾ RandomForestClassifier
RandomForestClassifier(random_state=42)
```

My model had a test size of 20% and a training size of 80%. These numbers were picked to ensure that overfitting does not occur. The forest had 100 decision trees and used a random state of 42. I chose 100 decision trees because too many decision trees can make the model hard to interpret and too few can make the model less accurate. 42 is regarded as a commonly used number for random state.

EVALUATING

I used a variety of metrics to evaluate my models' performance. The first was accuracy, this shows the proportion of correct predictions made over the total number of predictions.



```
##accuracy
rf.score(X_test,Y_test)
```

```
[45]:

0.92375
```

I was surprised that my model was able to predict loan approval by about 92%; I thought this number would be around 85%.

Next, I used a confusion matrix and classification report to see a more detailed view of my models' performance.

Fig 3.

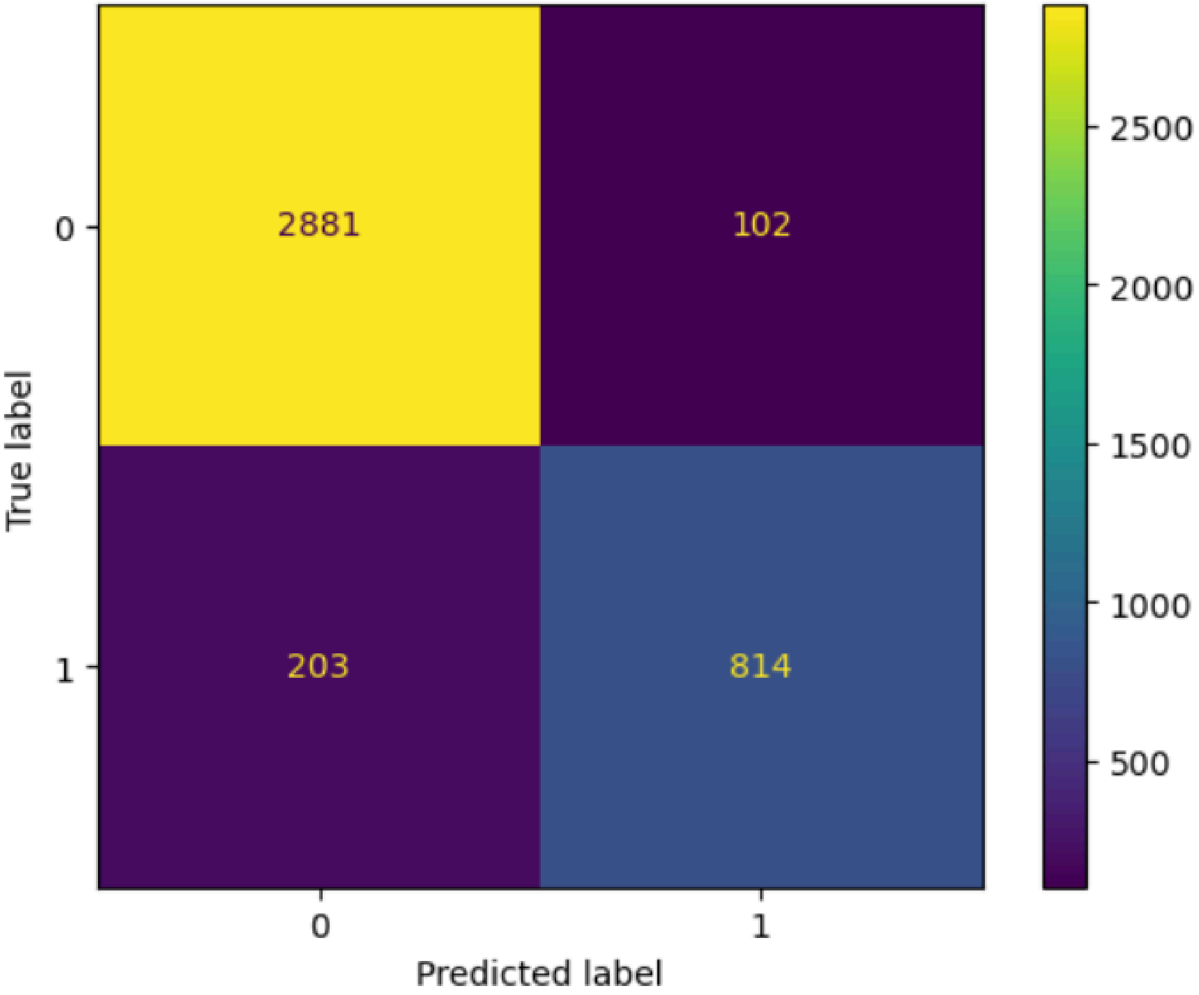
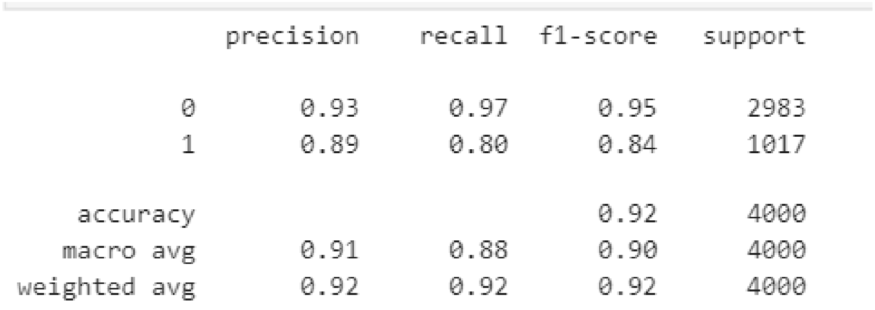


Fig.4

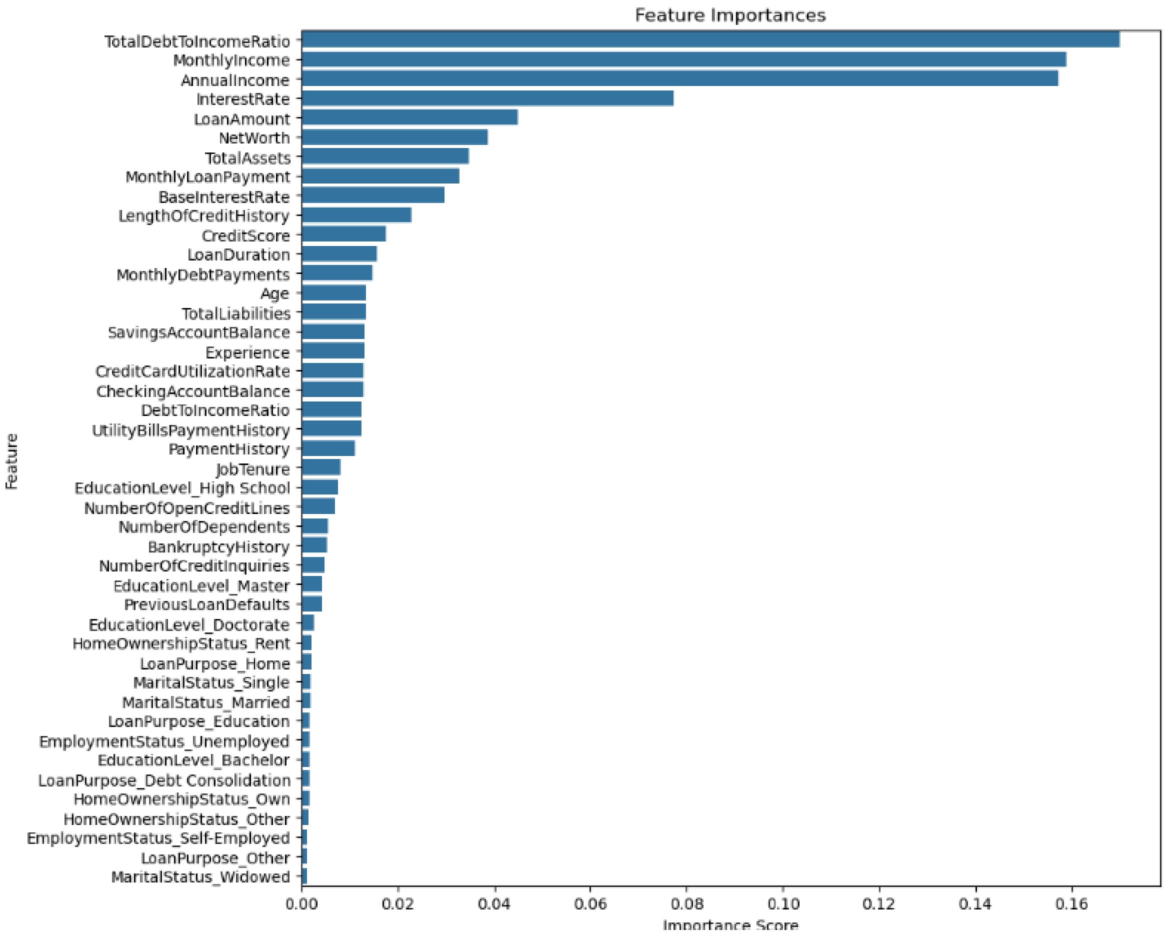


```
              precision    recall  f1-score   support

           0       0.93      0.97      0.95      2983
           1       0.89      0.80      0.84      1017

    accuracy                           0.92      4000
   macro avg       0.91      0.88      0.90      4000
weighted avg       0.92      0.92      0.92      4000
```

Lastly, I wanted to see what features were most important in predicting loan approval.



I found this bar graph very interesting. It shows that TotalDebtToIncomeRatio is the most important feature when predicting loan approval. Another thing I found interesting was that Age was more important than Education Level. Also, it looks like the features containing financial data were weighted more heavily than features containing personal data.

STORYTELLING/IMPACTS

After interpreting and creating my visualizations I have learned a lot about the features that contribute to loan approval. The most important features when it comes to predicting loan approval using the Random Tree Classifier are TotalDebtToIncomeRatio and Monthly/Annual Income. I found it interesting that the model weighted monthly income higher than annual income even though they are related. This data can be used to help others understand the different factors that go into the loan approval process.

References

https://pandas.pydata.org/docs/reference/index.html
https://scikit-learn.org/1.5/api/sklearn.metrics.html
https://seaborn.pydata.org/index.html