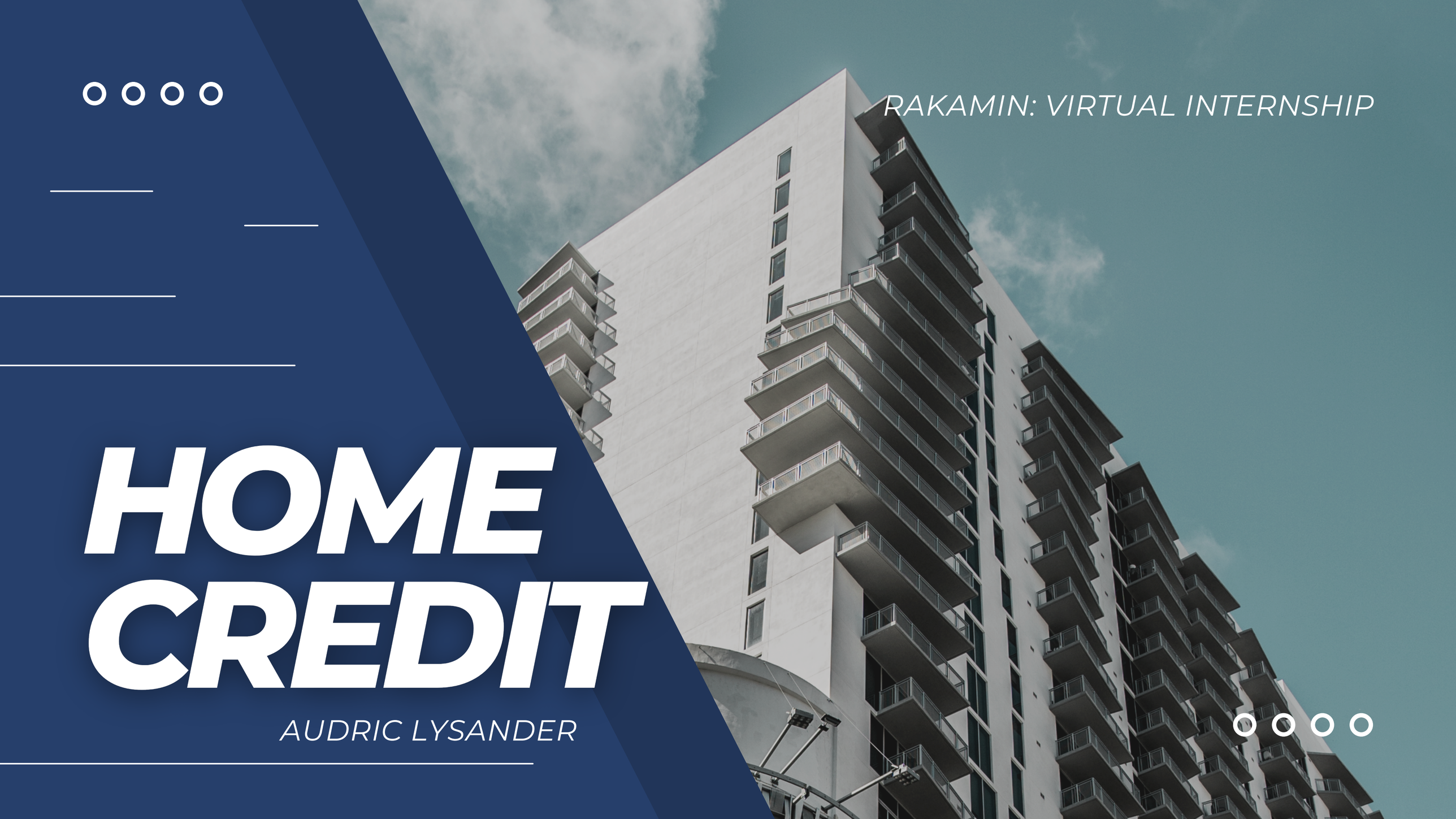

RAKAMIN: VIRTUAL INTERNSHIP

HOME CREDIT

AUDRIC LYSANDER

TABLE OF CONTENTS

THE OBJECTIVE

Home Credit is currently using various statistical methods and Machine Learning to make credit score predictions. Now, we ask you to unlock the full potential of our data. By doing so, we can ensure that customers who are able to make repayments are not turned down when applying for loans and that loans can be provided with a principal, maturity, and repayment calendar that will motivate customers to succeed. Evaluation will be carried out by checking how deep your understanding of the analysis you are working on is. For the record, you need to use at least 2 Machine Learning models where one of them is Logistic Regression.

DATA PREPARATION

121 columns with
307511 data

INITAL DATA QUALITY REPORT

More than 40 columns
with missing value.

ANALYSIS & FINDING

Customers have varying ages, but most are customers aged 30-40 years, so most of their work experience is between 0-10 years. Most of the customers also only have secondary education, so many of them have income from working as laborers who have an income between 90000 - 180000.

Most of the customers who applied for loans (64%) were married. Customers who are married and have a total family of 6-13 family members are more likely to have difficulty making payments because they have many other needs or lack good financial planning.

MODEL

Basic Logistic Regression

The results of the model score for the train and test data are around 91%, so it can be said to be a good fit model, but a recall value of 0% is obtained. This recall value can cause problems, because when the model predicts that the customer will not have a problem with payment, the actual data is that the customer has a problem with payment, so this needs to be handled with optimization.

Logistic Regression with Hyperparameter Tuning

For the logistic regression model which has been optimized with imlanaced data handling and hyperparameter tuning, it gets a score of 71% and 69% for the train and test data, so it can be said to be a goodfit model. Furthermore, an increase in recall was obtained to 58%. However, it is still in the low category, so it will be tried again with other models.

MODEL

Random Forest Classifier

The recall value for the two ensemble models is still 0, so data balancing must be done. In addition, the train and test scores have scores that are too far apart, namely 0 and 90%, so that it can be said to be an underfit model.

Random Forest Classifier with Hyperparameter Tuning

The results of the Random Forest Classifier were obtained as reported above, where the recall value increased to 36% and the train and test scores showed a goodfit model, but a recall value of 36% still could not be said to be a good model, so adjustments had to be made at the preprocessing / hyperparameter tuning / etc.

THANK YOU

github.com/AudricLysander/Home-Credit