Rakamin Academy X HOME CREDIT

Home Credit Scorecard Model

Home Credit Indonesia Data Scientist Project Based Internship Program

Presented by
Rais Yufli Xavierullah

Rais Yufli Xavierullah

About You

Ex-Entrepreneurship Assistant transitioning into a Data Scientist role after completing Hacktiv8 Data Science Bootcamp. Over two years of experience in government I was in charge of helping entrepreneurs move up their class level by helping to develop legality and also training. Core skills include providing actionable insights from modeling and statistical analysis.

Experience

- Entrepreneurship Assistant - Suku Dinas PPKUKM East Jakarta (2021 – 2023)
- Research and Development - Production System and Automation Laboratory Assistant (2018 – 2020)
- Engineer Staff Intern - Zi Argus (2019 – 2019)

Rakamin Academy

Case Study

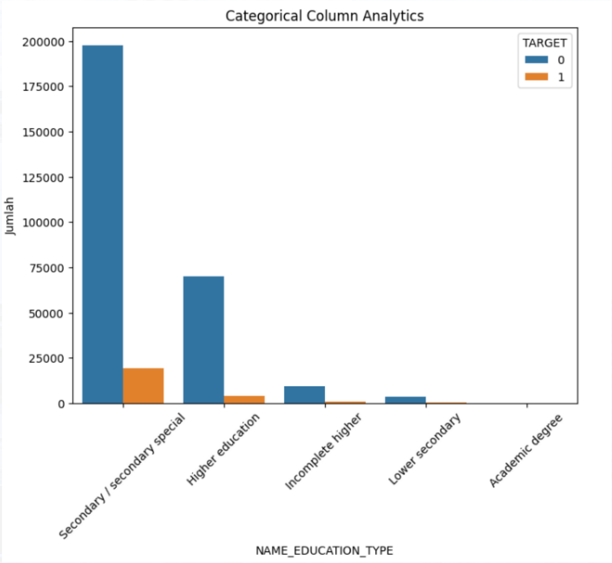
Home Credit has data about people who make loans to banks. This data contains 122 features regarding all existing customers. As a data scientist, I was assigned to create a model used by companies so that it could be used to predict which customers would have difficulty paying their debts or which would not.

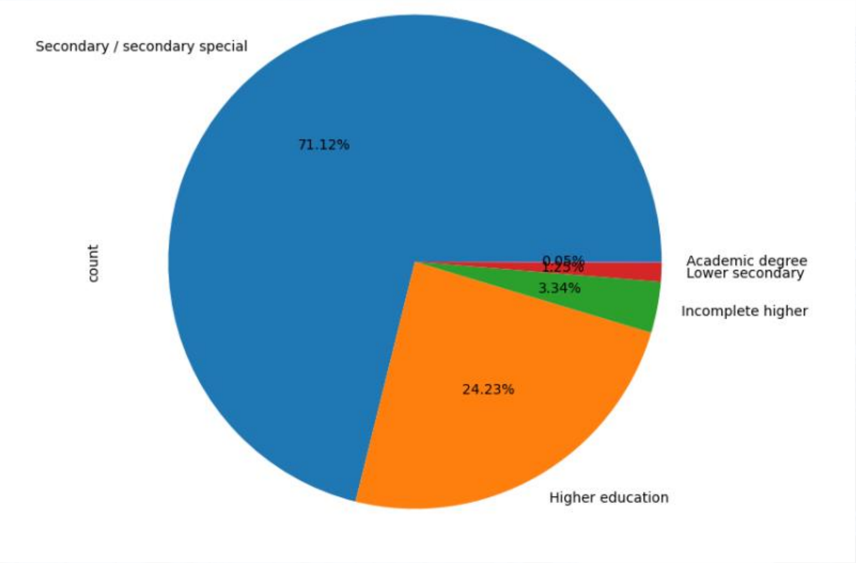
Dataset

The dataset has 122 features used for modeling. I try to reduce the features that I think are important so that the model does not become overfitting and makes computing efficient. I reduced several columns because these columns had a lot of missing value data, there was a lot of cardinality in the data, and some columns had a very small correlation with the target, so I used the data for modeling as many as 49 features.

EDA

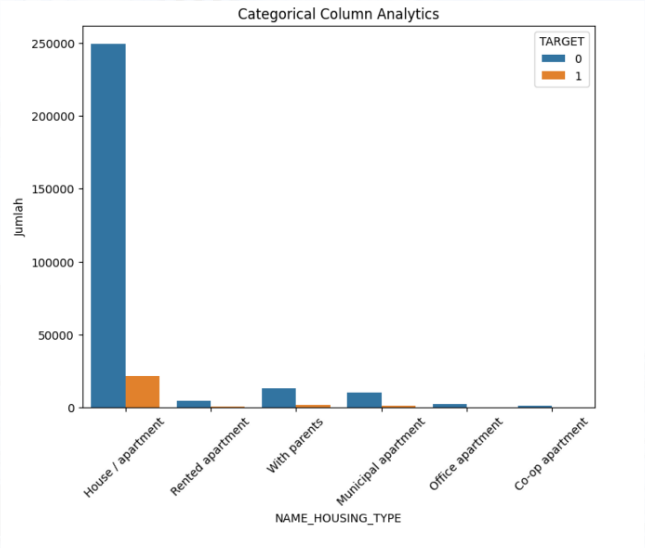
Rakamin Academy

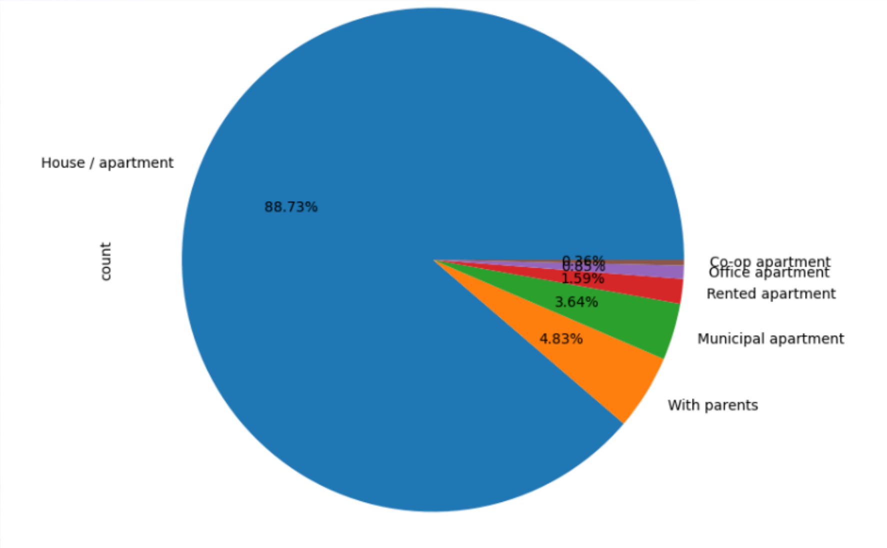




From the dataset, it was found that the people who borrowed the most money had difficulty paying or did not have low levels of education, namely secondary specials with a percentage of 72% of all borrowers. And those who borrow the least money are people who have an academic degree with a percentage of 0.05%

EDA





From the dataset, it was found that the people who borrowed the most money and had difficulty paying or not were people who lived in their own apartments or houses with a percentage of 88.73%. And those who borrow the least money are people who live in non-own ownership, namely cooperative apartments with a percentage of 0.36%

Feature selection

Rakamin Academy

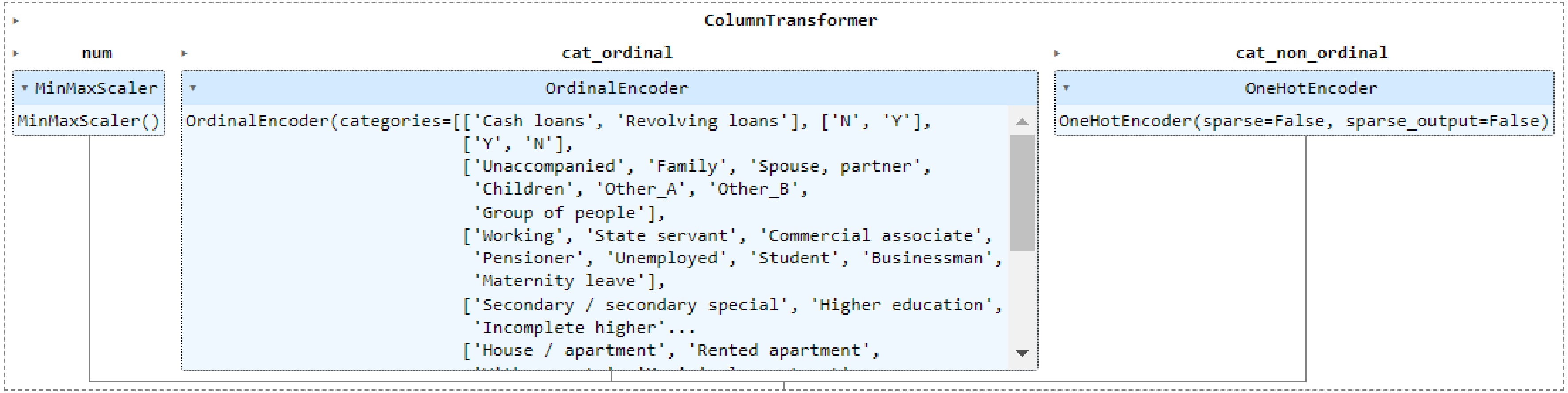
Kendalltau for Tables Numeric

```
The column that has correlation with loan status are ['CNT_CHILDREN',
'AMT_INCOME_TOTAL', 'AMT_CREDIT', 'AMT_GOODS_PRICE', 'REGION_POPULATION_RELATIVE',
'DAYS_BIRTH', 'DAYS_EMPLOYED', 'DAYS_REGISTRATION', 'DAYS_ID_PUBLISH', 'FLAG_EMP_PHONE',
'FLAG_WORK_PHONE', 'FLAG_PHONE', 'CNT_FAM_MEMBERS', 'REGION_RATING_CLIENT',
'REGION_RATING_CLIENT_W_CITY', 'HOUR_APPR_PROCESS_START', 'REG_REGION_NOT_LIVE_REGION',
'REG_REGION_NOT_WORK_REGION', 'REG_CITY_NOT_LIVE_CITY', 'REG_CITY_NOT_WORK_CITY',
'LIVE_CITY_NOT_WORK_CITY', 'OBS_30_CNT_SOCIAL_CIRCLE', 'DEF_30_CNT_SOCIAL_CIRCLE',
'OBS_60_CNT_SOCIAL_CIRCLE', 'DEF_60_CNT_SOCIAL_CIRCLE', 'DAYS_LAST_PHONE_CHANGE',
'FLAG_DOCUMENT_2', 'FLAG_DOCUMENT_3', 'FLAG_DOCUMENT_6', 'FLAG_DOCUMENT_8',
'FLAG_DOCUMENT_9', 'FLAG_DOCUMENT_11', 'FLAG_DOCUMENT_13', 'FLAG_DOCUMENT_14',
'FLAG_DOCUMENT_15', 'FLAG_DOCUMENT_16', 'FLAG_DOCUMENT_18', 'FLAG_DOCUMENT_21']
```

Chi Square for Tables Categorical

```
The column that has correlation with loan status are  ['NAME_CONTRACT_TYPE',
'CODE_GENDER', 'FLAG_OWN_CAR', 'FLAG_OWN_REALTY', 'NAME_TYPE_SUITE', 'NAME_INCOME_TYPE',
'NAME_EDUCATION_TYPE', 'NAME_FAMILY_STATUS', 'NAME_HOUSING_TYPE', 'OCCUPATION_TYPE',
'WEEKDAY_APPR_PROCESS_START']
```

Preprocessing



In the preprocessing stage I will normalize all the data. For numerical data I use the minmaxscaler method as a normalization method. For data that has a categorical type but is multilevel, I use the ordinalencoder method as a normalization method. For data that has a categorical type but is not stratified, I use the onehotencoder method as a normalization method

Model

Model KNN

```
------Clasification Report KNN Train-------
              precision    recall  f1-score   support

           0       0.92      1.00      0.96    224373
           1       0.62      0.07      0.12     19775

    accuracy                           0.92    244148
   macro avg       0.77      0.53      0.54    244148
weighted avg       0.90      0.92      0.89    244148


------Clasification Report KNN Test-------
              precision    recall  f1-score   support

           0       0.92      0.99      0.95     56093
           1       0.14      0.02      0.03      4944

    accuracy                           0.91     61037
   macro avg       0.53      0.50      0.49     61037
weighted avg       0.86      0.91      0.88     61037
```

Model Logistic Regression

```
------Clasification Report Logistic Regression Train---
              precision    recall  f1-score   support

           0       0.92      1.00      0.96    224373
           1       0.00      0.00      0.00     19775

    accuracy                           0.92    244148
   macro avg       0.46      0.50      0.48    244148
weighted avg       0.84      0.92      0.88    244148


------Clasification Report Logistic Regression Test---
              precision    recall  f1-score   support

           0       0.92      1.00      0.96     56093
           1       0.00      0.00      0.00      4944

    accuracy                           0.92     61037
   macro avg       0.46      0.50      0.48     61037
weighted avg       0.84      0.92      0.88     61037
```

Model

Model Random Forest

------Clasification Report Random Forest Train-------

	precision	recall	f1-score	support
0	1.00	1.00	1.00	224373
1	1.00	1.00	1.00	19775
accuracy			1.00	244148
macro avg	1.00	1.00	1.00	244148
weighted avg	1.00	1.00	1.00	244148

------Clasification Report Random Forest Test-------

	precision	recall	f1-score	support
0	0.92	1.00	0.96	56093
1	0.00	0.00	0.00	4944
accuracy			0.92	61037
macro avg	0.46	0.50	0.48	61037
weighted avg	0.84	0.92	0.88	61037

I use the KNN model because this model can predict target 1 even though the percentage is very small compared to other models that cannot predict at all

Conclusion

- The majority of people who borrow money are people with low levels of education, namely secondary special at 72% and also people who already have their own house or apartment at 88.73%. It can be assumed that people who have their own house or apartment are taking out home ownership loans, which means that the people who borrow the most are people who have their own house or apartment.

- In the model that I use, namely the KNN model, although its accuracy is 91%, it is smaller than other models, but this model can predict target 1 which has a percentage of 3% while the other models cannot predict target 1 at all.

- The reason why this model cannot predict target 1 is because the data for target 1 is only 8% of the total data, which means the data is not very imbalanced. So that the model can be better, you can use several methods by resampling by adding data or subtracting data from the data, giving weights, or adding more real data.

Github

- https://github.com/Raisyuflix/Loan_Prediction

Thank You

Rakamin Academy X HOME CREDIT