Home Credit Default Risk

1. Description

Many people struggle to get loans due to insufficient or non-existent credit histories. And, unfortunately, this population is often taken advantage of by untrustworthy lenders.

Home Credit Group

Home Credit strives to broaden financial inclusion for the unbanked population by providing a positive and safe borrowing experience. In order to make sure this underserved population has a positive loan experience, Home Credit makes use of a variety of alternative data--including telco and transactional information--to predict their clients' repayment abilities.

While Home Credit is currently using various statistical and machine learning methods to make these predictions, they're challenging Kagglers to help them unlock the full potential of their data. Doing so will ensure that clients capable of repayment are not rejected and that loans are given with a principal, maturity, and repayment calendar that will empower their clients to be successful.

2. Udacity Final Project

This directory contain all code that was used for the Udacity Data Scientist Nanodegree Program

3. Define the Problem

For this project, the problem statement is given to us , develop an algorithm to predict the default of home credit .

Project Summary: Many people struggle to get loans due to insufficient or non-existent credit histories. And, unfortunately, this population is often taken advantage of by untrustworthy lenders.

In this project, we ask you to complete the analysis of which customers of home credit were likely default. In particular, we ask you to apply the tools of machine learning to predict which customers defaulted.

Project Metrics: Default customer can be predicted using less variable at credit risk perspective. So selected model specification must be explainable and applicable.

4. The 4 C's of Data Cleaning: Correcting, Completing, Creating, and Converting

In this stage, data should have been cleaned

Correcting abnormal values and outliers Completing missing information Creating new features for analysis Converting fields to the correct format for calculations and presentation. Correcting: Reviewing the data, there should have been analyzed to be any abnormal or non-acceptable data inputs. In addition, age and income may have outlier values.Exploratory analysis will done to find reasonable values. Outliers should been elimated in dataset.It should be noted, that if unreasonable values were , for example age is 1000 then it also should be elimaneted.

Completing: There are null values or missing data in dataset. Missing values can be bad, because some algorithms don't know how-to handle null values and will fail. While others, like decision trees, can handle null values. Thus, it's important to fix before modeling will started because several models will have compared. There are two common methods, either delete the record or populate the missing value using a reasonable input. It is not recommended to delete the record, especially a large percentage of records, unless it truly represents an incomplete record. Instead, it's best to impute missing values. A basic methodology for qualitative data is impute using mode. A basic methodology for quantitative data is impute using mean, median, or mean + randomized standard deviation.

Creating: Feature engineering is when we use existing features to create new features to determine if they provide new signals to predict our outcome.

Converting: Last, but certainly not least, we'll deal with formatting. There are no date or currency formats, but datatype formats. Our categorical data imported as objects, which makes it difficult for mathematical calculations. For this dataset, we will convert object datatypes to categorical dummy variables

6. Correlation Elimination

All variable analyze the correlation of target. We will choose higher than 0.05 or lower than -0.005. Correlations are very useful in many applications, especially when conducting regression analysis. However, it should not be mixed with causality and misinterpreted in any way. I should also always check the correlation between different variables in our dataset and gather some insights as part of my exploration and analysis.

6.1 Final Model Variable

- AGE_CAL
- CODE_GENDER_F
- DAYS_LAST_PHONE_CHANGE

7. Results Summary

Model ranking based on test data accuracy and AUC

- LogisticRegression model alternative 1 accuracy: 0.920
- LogisticRegression model alternative 2 accuracy: 0.920
- LogisticRegression model alternative 1 AUC: 0.626
- LogisticRegression model alternative 2 AUC: 0.608

Model ranking based on cross validation data F1-score

- LogisticRegression model alternative 1 accuracy: 0.919
- LogisticRegression model alternative 2 accuracy: 0.919
- LogisticRegression model alternative 1 AUC: 0.621
- LogisticRegression model alternative 2 AUC: 0.609

Model altenative 2 is better than first model. Beucause of Cross Validation AUC value is higer than train dataset. In addition to between Model alternative 1 and Model alternative 2 Auc is similiar so I choose the Model Alternative 2.

8. Conclusion

My expectation would be credit type,credit amount or income type for final variable modelling. But these variable were eliminated correlation step. I am surprised for this happen. This proeject aimed to end to end data processing and data modelling in credit risk data. I enjoyed to analyze and create this project

9. References

- Udacity Data Scientist Nanodegree Program
- Kaggle's Home Credit Defa   ult Risk
- Source Data Dictionary
- Creating New Column
- Correlation Elimination
- Train Test Cross Validation
- Credit Risk Modelling
- Lgistic Regression
- Gradient Boosting
- LightGBM’s documentation
- Pandas Data Frame Describe
- Pandas Missing Data