Abstract

This research focuses on predicting default risk in home equity loans (HLE) using three classification models: logistic regression, decision trees, and random forests. Analyzing a dataset from a local bank comprising 5,960 recent HLEs, we address key questions regarding loan approval, default prediction, and feature importance.

Data preprocessing involves handling missing values, outliers, and feature engineering. Exploratory data analysis identifies critical predictors, including debt-to-income ratio, credit history, and job stability.

Logistic regression highlights significant predictors such as derogatory records and debt-to-income ratio. Decision tree models emphasize the importance of financial disclosure and debt-to-income ratio. Random forest models achieve high accuracy and recall rates.

In conclusion, decision tree models strike a balance between comprehensiveness and precision, while random forest models offer high accuracy. Future research may explore hyperparameter tuning and alternative models to enhance prediction accuracy. Adopting a tuned decision tree model can effectively identify risky clients and aid banks in mitigating default risks.

Keywords: Home Equity Loans (HLE), Logistic Regression, Decision Trees, Random Forest

Home Equity Loans Default Prediction

Home equity loans (HLE), a loan which allows costumer to borrow money using the equity in their home as collateral, has become popular in recent high-interest rate environment[1]. A major proportion of HLE profit comes from interests paid to bank. Non-Performing Assets (NPA), or bad debts usually eat up a major chunk of a bank's profit as they not only yield no interest, but eventually give no principal back. In this research, we aim to simplify the decision-making process for home equity lines of credit. We will utilize a dataset provided by a local bank to build three classification models and then make prediction on a test dataset to assess model performance.

Introduction

We want to answer three key questions: Should we approve a client based on the information he/she provided? What kinds of clients are likely to default on their loan? What are important features to consider while approving a loan? The bank's consumer credit department can utilize our models to answer these questions, predict clients who are likely to default on their loan, and understand important features to consider while approving a loan. Due to regulation requirement[2], the model created must be interpretable enough to provide a justification for any adverse behavior (rejections). Deep learning models like neural networks will be not used due to their lack of interpretability.

Our dataset is given from a local bank internally and, to protect trade secret, we will not disclose the bank's name. The data contains existing loan underwriting records. The dataset contains baseline and loan performance information for 5,960 recent home equity loans. The target, a binary flag call "BAD", is a binary variable that indicates whether an applicant has ultimately defaulted or has been severely delinquent. This adverse outcome occurred in 1,189 cases (20 percent). 12 input variables were registered for each applicant. A simple dictionary is provided below.

- **BAD:** 1 = Client defaulted on loan, 0 = loan repaid
- **LOAN:** Amount of loan approved.
- **MORTDUE:** Amount due on the existing mortgage.
- **VALUE:** Current value of the property.
- **REASON:** Reason for the loan request. (HomeImp = home improvement, DebtCon= debt consolidation which means taking out a new loan to pay off other liabilities and consumer debts)
- **JOB:** The type of job that loan applicant has such as manager, self, etc.
- **YOJ:** Years at present job.
- **DEROG:** Number of major derogatory reports (which indicates a serious delinquency or late payments).
- **DELINQ:** Number of delinquent credit lines (a line of credit becomes delinquent when a borrower does not make the minimum required payments 30 to 60 days past the day on which the payments were due).
- **CLAGE:** Age of the oldest credit line in months.
- **NINQ:** Number of recent credit inquiries.
- **CLNO:** Number of existing credit lines.
- **DEBTINC:** Debt-to-income ratio (all your monthly debt payments divided by your gross monthly income. This number is one way lenders measure your ability to manage the monthly payments to repay the money you plan to borrow.

Data Cleaning and Exploring

First step is data cleaning and feature engineering. We identify missing value in almost every variable, and therefore data cleaning is required. We also understand some variables are numeric while the others are categorical, which should be treated separately regarding missing value and feature engineering. We further find outliers in many variables and needs to decide drop them or not. Filling missing value with median (numeric), mode (categorical), etc. could retain important information but could also create bias for our analysis. The actual data

for the missing value could be very different from the value we filled. Our final treatment is to keep outliers since they occur naturally and contains meaningful information. We decide to fill missing value with median/mode, and prepared two train datasets: For each column we create a binary flag, if there are missing values for a row in this column, then 1 else 0. The other dataset does not have this missing value flag. (Missing value flag is a kind of feature engineering)

Second step is exploratory data analysis (EDA). We did some univariate (bar plots, histograms, boxplots, etc.), bivariate (two-way scatterplots), and multi-variate (pair plots, heat maps) to identify correlated variables. EDA pave the road for successful modeling. Third step is modeling, we applied logistic regression, decision trees, random forest classifiers on two datasets, one with missing value flag, the other without. For this research paper, we will focus on the one with missing value flag. We split trained dataset with 30% test and 70% trained data, using trained data to train the model, and test data to evaluate model performance.

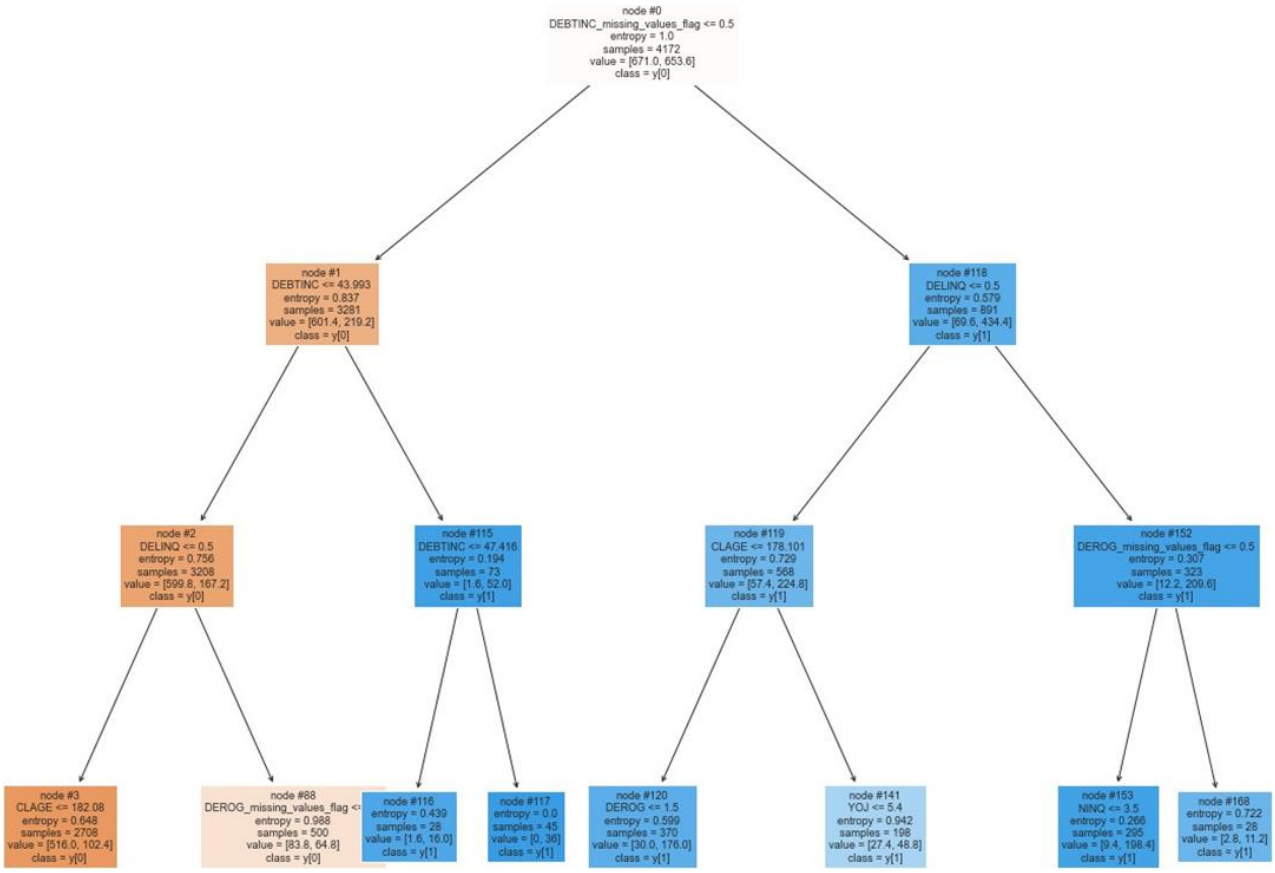
Modeling: Logistic Regression

Logistic regression estimates the probability of an event occurring, such as default or not default, based on a set of independent variables. In logistic regression, a logit transformation is applied on the odds—that is, the probability of success divided by the probability of failure.[3]

We tried two kinds of logistic regression models. When applied simple logistic model directly on the dataset, only 3% of bad loan (default) cases are captured. After using Lasso L1 regularized regression, standard scaling and 0.5 threshold, we got a much better model capturing 62% of bad loan cases. Important features we identified are “Ever Has Derogatory Records “, “Missing number of credit lines data flag “, “Missing applicant job type data“, “Missing derogatory records flag“, “Debt-to-Income Ratio“, “Ever Delinquent“. It can be seen that if a person has been seriously delinquent and has high debts to income ratio are very likely to default on their home loan.

Modeling: Decision Tree

Decision tree utilize quantitative selection criteria like Entropy to identify the optimal split points within a tree hat can classify data into separated groups. This process of splitting is then repeated in a top-down, recursive manner until all, or the majority of records have been classified under specific class labels.[4] Decision Tree model is relatively easy to overfit, performing almost perfect on trained dataset but poorly on the test dataset. We tried three trees: one simple balanced tree and overfitted: one with imbalance weights, 0.8 toward bad loan case and 0.2 toward good loan, and still overfitted; finally, a hyperparameter tuned tree with imbalance weight, max depth 12, minimum leaf sample of 24, and not overfit. Our model can capture 77% of default case in test dataset.

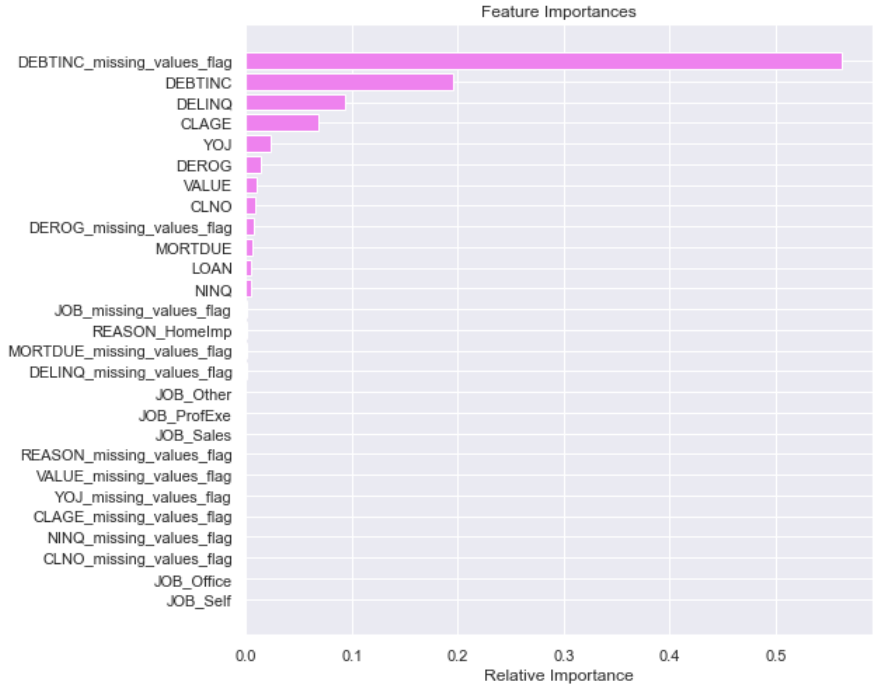


By visualizing the decision tree, we can see if clients do not disclose their Debt-to-Income ratio when applying, they are more likely to default. If clients have >47 Debt-to-income ratio, they are very likely to default. If clients have a debt-to-income ratio < 44, no delinquent account, and have account age > 182 months, they are less likely to default.

Modeling: Random Forest

Random Forest constructs many decision trees using resampling techniques like bootstrapping and random selection of features Less interpretable than decision tree.[5] Random Forest model is even easier to overfitting and hyperparameter tuning requires a long time. we carefully choose an unbalanced parameters

(class_weight={0: 0.2, 1: 0.8}, criterion='entropy', max_depth=4, max_features=0.7, min_samples_leaf=30, n_estimators=90, random_state=7) and got a decent model that can capture 74% default case (recall) with 87% accuracy (precision). Some important features includes Debt-to-Income ratio, ever delinquent, Age of the oldest credit line in months, etc.



Conclusion

In conclusion, all models indicates that the most important factors to consider when determine the default risk of a clients are: Debt-to-income Ratio, Number of Delinquent Accounts, Age of the oldest credit line in months, whether there is a derogatory report, years of job experience. Decision Tree seems to perform better overall on to capture as many default cases as possible. Although the random forest model capture 5% fewer actual default client than the decision tree model, the

precision is much higher, meaning higher accuracy and fewer false positive cases. Depends on the bank's preference, it can sacrifice some prediction comprehensiveness in exchange for higher precision and cost-saving.

	Model	Train_Accuracy	Test_Accuracy	Train_Recall	Test_Recall	Train_Precision	Test_Precision
0	LogisticRegression(No Penalty)	0.810000	0.790000	0.040000	0.010000	0.740000	0.500000
1	LogisticRegression('l1')	0.890000	0.890000	0.630000	0.620000	0.780000	0.820000
2	DecisionTree	1.000000	0.865772	1.000000	0.618280	1.000000	0.701220
3	DecisionTree_Tuned	0.863135	0.838367	0.884945	0.776882	0.602500	0.583838
4	Random Forest(Default)	1.000000	0.914989	1.000000	0.706989	1.000000	0.859477
5	Random Forest(class_weight)	1.000000	0.907159	1.000000	0.666667	1.000000	0.855172
6	Random Forest_Tuned	0.870566	0.864653	0.787026	0.744624	0.637265	0.653302

Limitations and Calls for Action

Due to the calculation power limitation, we did not tune a lot of hyperparameters with random forest model and we can in the future find the best hyperparameters. We can also use other models like random forest, SVM, KNN to improve the chance of finding the best model. We purpose to adopt tuned decision tree model. This model can capture ~80% clients that actually default. Banks can use this model as a good starting point to identify risky clients/applications and conduct further research.

Reference

1. Artificial Intelligence and Machine Learning in Financial Services. (n.d.). https://crsreports.congress.gov/product/pdf/R/R47997

2. *What is a home equity loan?*. Consumer Financial Protection Bureau. (n.d.). https://www.consumerfinance.gov/ask-cfpb/what-is-a-home-equity-loan-en-106/

3. *What is logistic regression?*. IBM. (2024b, March 18). https://www.ibm.com/topics/logistic-regression

4. *What is a decision tree?*. IBM. (2024a, May 10). https://www.ibm.com/topics/decision-trees

5. *What is Random Forest?*. IBM. (2024c, April 2). https://www.ibm.com/topics/random-forest