Loan Approval Prediction using Random Forests

Fidelis Prasetyo, Sheldin Lau, Yao Lin, Andrew Lau, Arze Lu

Introduction

- Analyze loan dataset
- Simulate loan approval
- Build machine learning model for prediction
- Evaluation methods such as precision, F1-score, confusion matrix, etc.

Dataset

- Loan approval, synthetic data
- 45000 records and 14 features
- Extended by SMOTENC
- Class distribution - loan declined 78%, loan being approved 22%

rson_age	person_gender	person_education	person_income	person_emp_exp	person_home_ownership	loan_amnt	loan_intent	loan_int_rate	loan_percent_income	cb_person	credit_sco	previous_l	loan_st
22	female	Master	71948	0	RENT	35000	PERSONAL	16.02	0.49	3	561	No	
21	female	High School	12282	0	OWN	1000	EDUCATION	11.14	0.08	2	504	Yes	
25	female	High School	12438	3	MORTGAGE	5500	MEDICAL	12.87	0.44	3	635	No	
23	female	Bachelor	79753	0	RENT	35000	MEDICAL	15.23	0.44	2	675	No	
24	male	Master	66135	1	RENT	35000	MEDICAL	14.27	0.53	4	586	No	
21	female	High School	12951	0	OWN	2500	VENTURE	7.14	0.19	2	532	No	
26	female	Bachelor	93471	1	RENT	35000	EDUCATION	12.42	0.37	3	701	No	
24	female	High School	95550	5	RENT	35000	MEDICAL	11.11	0.37	4	585	No	
24	female	Associate	100684	3	RENT	35000	PERSONAL	8.9	0.35	2	544	No	
21	female	High School	12739	0	OWN	1600	VENTURE	14.74	0.13	3	640	No	
22	female	High School	102985	0	RENT	35000	VENTURE	10.37	0.34	4	621	No	
21	female	Associate	13113	0	OWN	4500	HOMEIMPROVEMENT	8.63	0.34	2	651	No	
23	male	Bachelor	114860	3	RENT	35000	VENTURE	7.9	0.3	2	573	No	
26	male	Master	130713	0	RENT	35000	EDUCATION	18.39	0.27	4	708	No	
23	female	Associate	138998	0	RENT	35000	EDUCATION	7.9	0.25	4	583	No	
23	female	Master	600891	5	MORTGAGE	30000	DEBTCONSOLIDATION	10.65	0.05	3	670	Yes	
23	male	Bachelor	144943	0	RENT	35000	EDUCATION	7.9	0.24	4	663	No	

Methodology

- Datasets heavily imbalanced.
- Random Forest chosen to reduce overfitting.
- 70-15-15 split.
- Fine tune by adjusting depth of tree

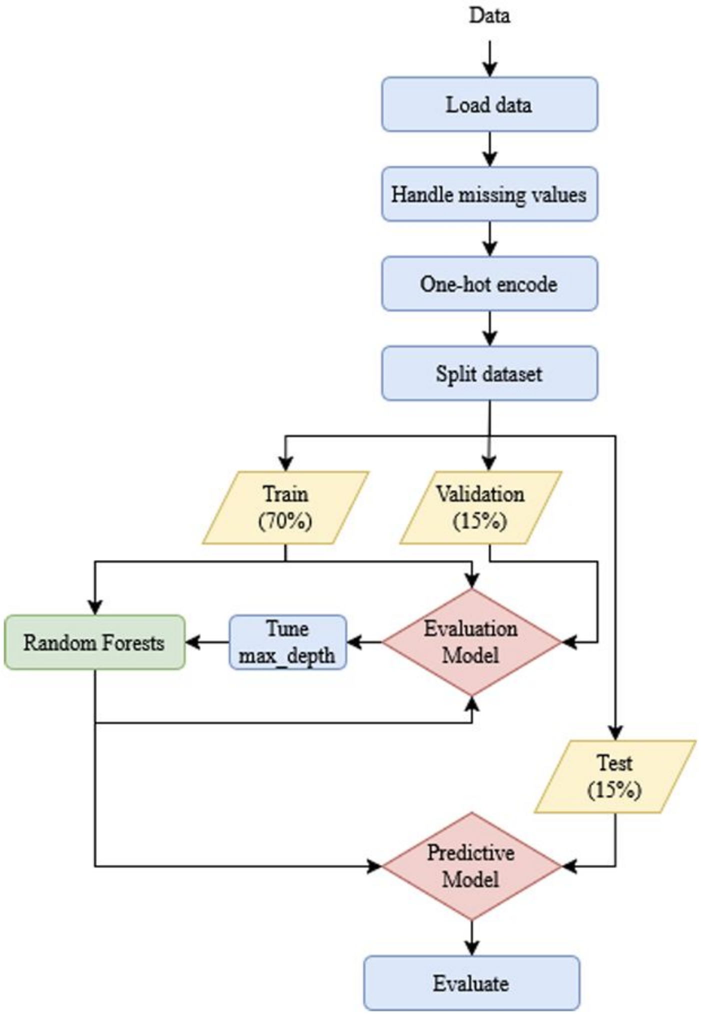
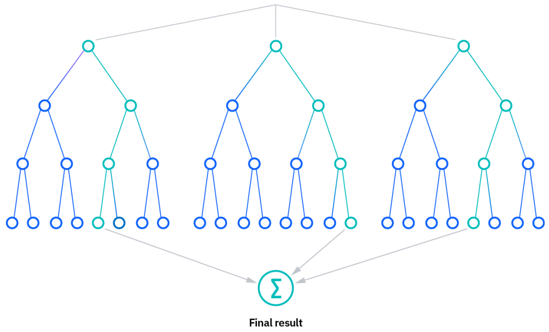


Fig. 1. Proposed model development pipeline

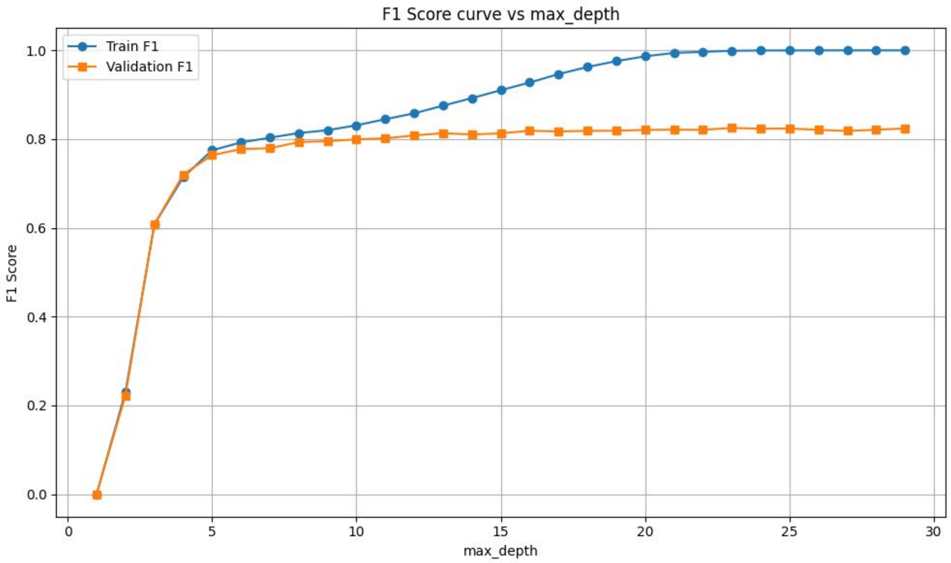
Why Random Forest

- It's an ensemble learning method (set of classifiers) as it's made up of multiple decision trees.
    - **Bagging (boostrap aggregation)**: random sample of data is training set with replacement.
    - **Feature randomness:** random subset of features.
- Prediction is made by aggregating individual tree's predictions.
- **Hyperparameters:** node size, number of trees, number of features.
- Benefit:
    - Reduced risk of overfitting
    - Flexibility
    - Feature importance
- Challenges:
    - Time consuming
    - Needs larger data sets
    - Harder to interpret



Results—fine tune

- Shows f1-scores of training and validation datasets
- Overfit when max tree depth around 10 to 15
- Selected 10 as max_depth final value



Results—sampling

- Use different sampling strategies - original, weighted, and SMOTE
- Original model performs the best based on accuracy and F1-Score
- Significantly worse recall in the original then SMOTE, matches the dataset
- Prioritize precision over recall to fit real world scenario better

TABLE I
PERFORMANCE OF THE MODEL USING DIFFERENT SAMPLING METHODS

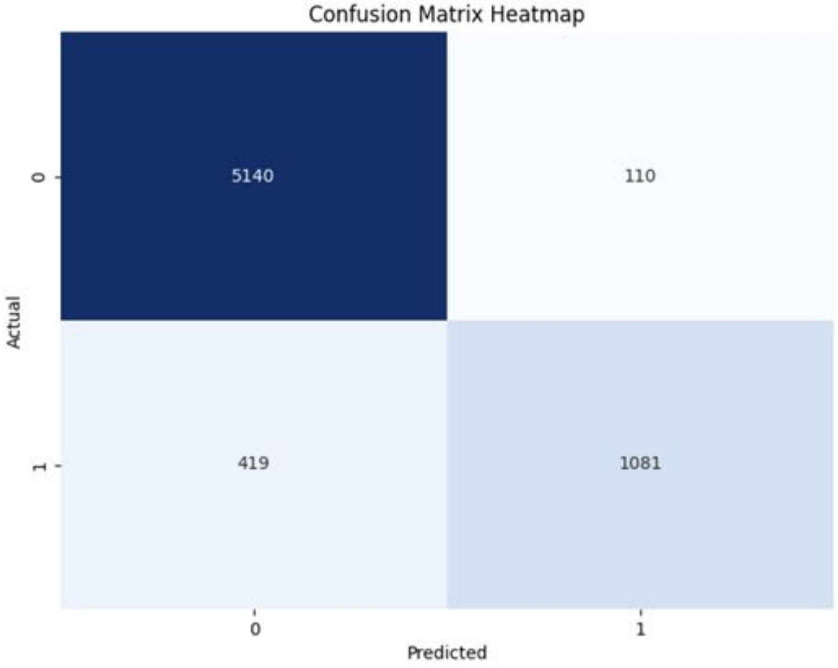
	Original		Weighted		With SMOTE	
Class	**0**	**1**	**0**	**1**	**0**	**1**
Precision	0.92	0.91	0.97	0.7	0.96	0.71
Recall	0.98	0.72	0.89	0.89	0.9	0.85
F1-Score	0.95	0.8	0.93	0.78	0.93	0.78
Accuracy	0.92		0.89		0.89	

*0 = Rejected, 1 = Approved

Results—evaluation

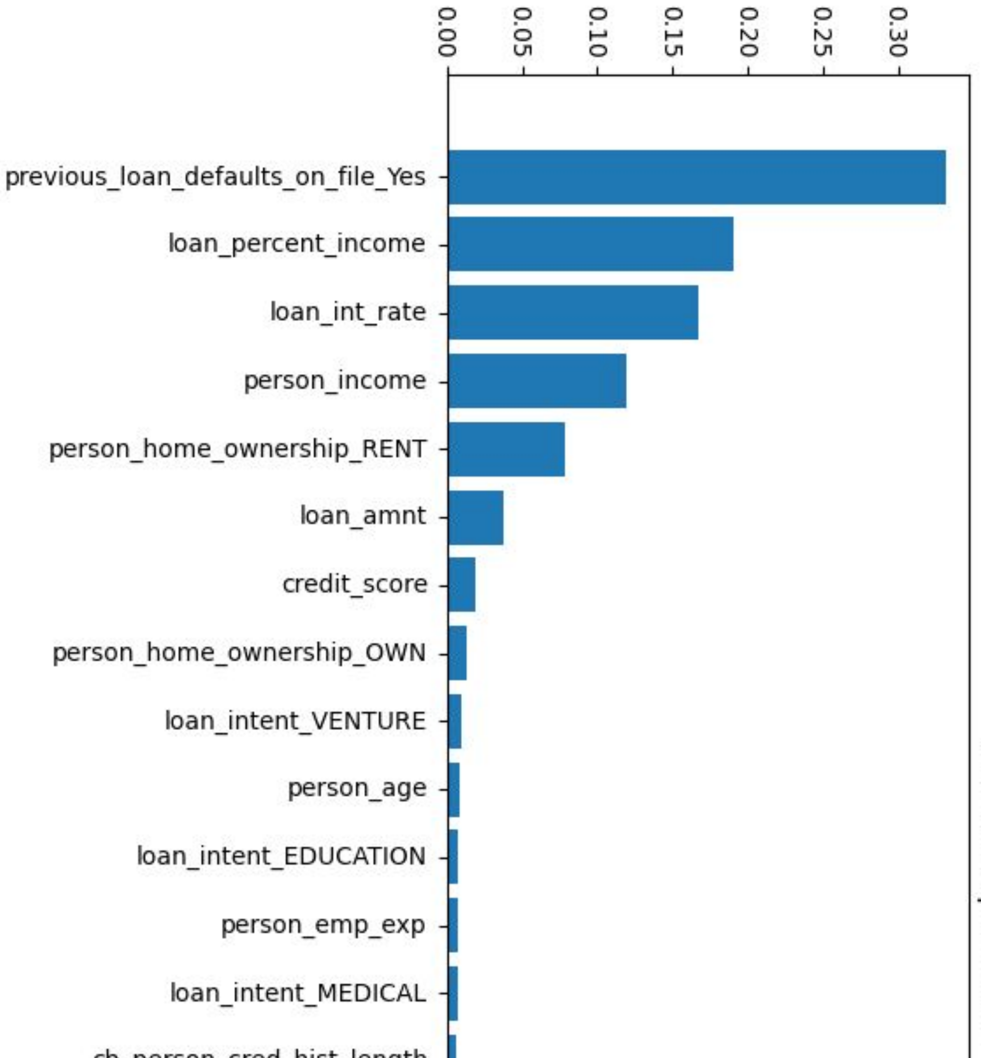
- Performs well in predicting true positives
- Slightly better at predicting true negatives
- 28% actual approvals missed
- 9% predicted approvals incorrect

	Original	
Class	0	1
Precision	0.92	0.91
Recall	0.98	0.72
F1-Score	0.95	0.8
Accuracy	0.92	



Results

- Applicant's loan default history the most determining factor
- Demographics of applicant (gender, age, education) barely affect decision



Conclusion

- Most important features - previous loans being defaulted, loan to income percentage and loan interest rate
- Can overfit very easily if max depth is too high
- Weighted random forest does not perform well