Revamping Mortgage Approval Operation

Johnny Chan Fan Hui

Recap of Bank’s Predicament

- Mortgage Default rate of 15% (bad debt of established banks are at 4%)
- Affecting bottom line margin and hurting shareholders’ value
- At the brink of security downgrade by Moody’s

Scope

01 Data Cleaning and Feature Engineering

02 Feature Selection

03 Data Preprocessing

04 Models training, predictive performance and selection

05 Discussion

01

Data Cleaning and Feature Engineering

The Data was joint from multiple sources

- Borrower's personal information (retain)
- Geospatial data of work and home address (dropped)
- Demographic statistic of borrower's origin (dropped)

Column Name	Missing Data and column deletion	Feature Engineering	Drop non-useful columns
gender	Drop missing row		
birthyear		Convert to Age	
maritalstatus	Drop missing row		
numofdependence			
education		Covert to 4 ordinal levels	
professionid			Dropped column
homestatus			Dropped column
staysinceyear		Convert to years of ownership	
EmploymentSinceYear	Back-fill, drop missing row and drop latter column	Covert to years of work experience	
MainBusinessSinceYear			
jobtypeid	Forward-fill into jobpos and drop jobtypeid	Convert Entrepreneur, others and education into self employed. Retain only self employed, staff, manager, supervisor and director	
jobpos			
monthlyfixedincome		Combine into household income	
monthlyvariableincome			
spouseincome	Fill missing value with 0		
MaxOverDueDays		Default (> 90 day overdue) = 1, otherwise = 0	

02

Feature Selection

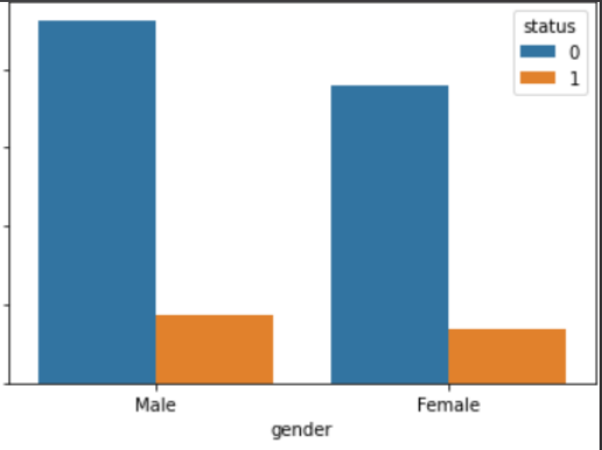
Numeric Variables

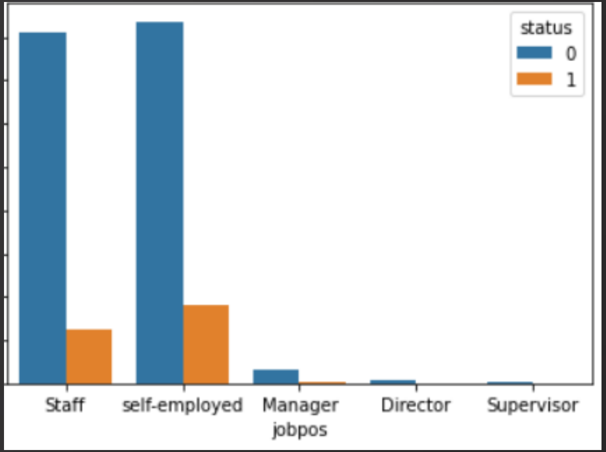
No two columns were strongly correlated. All columns were used

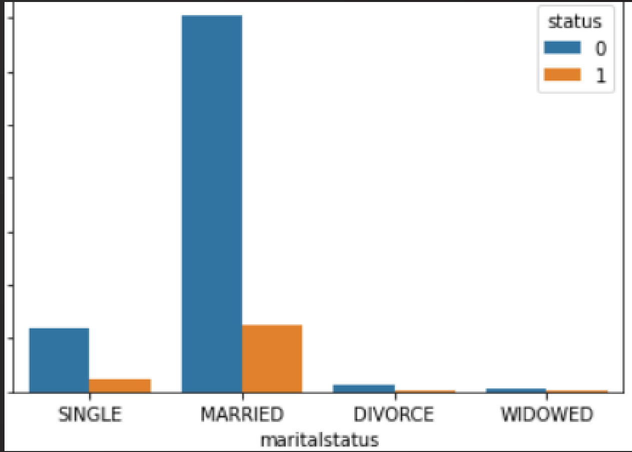
	numofdependence	education	age	household_income	work_experience	years_home_owned
numofdependence	1.000000	-0.059467	0.393454	0.058754	0.024034	-0.006137
education	-0.059467	1.000000	-0.004440	0.116505	0.045915	-0.068272
age	0.393454	-0.004440	1.000000	0.083122	0.076135	0.210013
household_income	0.058754	0.116505	0.083122	1.000000	-0.000815	-0.043121
work_experience	0.024034	0.045915	0.076135	-0.000815	1.000000	0.022823
years_home_owned	-0.006137	-0.068272	0.210013	-0.043121	0.022823	1.000000

Categorical Variables

No single level dominated all default cases. Very little chance of model using the level as key predictor. Hence, all categorical variables were retained







03

Data Preprocessing

Pre-processing Steps

1. **Balance dataset**
   Defaulted client is only 15% of dataset. Bootstrapping is done to increase defaulters to 50% of dataset
2. **One hot encoding categorical variables**
3. **Train-Test split**
4. **Feature scaling of numerical variables**

04

Models training and predictive performance

Model Performance

Ensemble of all 4 models: 96.1% AUC

Logistic Regression
(56.8% AUC)

Navie Bayes Model
(51.8% AUC)

Decision Tree
(89.9% AUC)

Random Forest
(94.4% AUC)
(201 trees)

Evaluvation of Models and Model Selection

Random Forest		Predicted Class	
		0	*1*
Actual Class	0	30494	3210
	1	547	33093

AUC: 94.4%

Ensemble		Predicted Class	
		0	*1*
Actual Class	0	28539	5165
	1	525	33115

AUC: 96.1%

Despite having a better AUC score, the ensemble of 4 models only captured 22 more defaulters while predicting ~2k more false positivity cases, resulting in a substantial lost of revenue for the bank. Therefore, the random forest was selected for deployment

Additionally, with the balancing of dataset, there is no **overfitting of model**

05

Discussion

1. Drawbacks

1) We assume that the clients have consistent income. Loan default due to loss of job is not captured

2) There is no information on the loan such as loan amount, tenure and interest rate. Such information, coupled with income, is pertinent in assessing client's ability to service the loan

2. Area of Improvement (time permitting)

1) Explore other method of ensemble such as bagging, stacking and boosting

2) With the geospatial data of home address, conduct web scrapping of the median home price in the region to proxy for loan amount

3) We can explore creating another model which aims to tune the interest rate to the client profile to minimize default rate and maximize revenue

Thank you