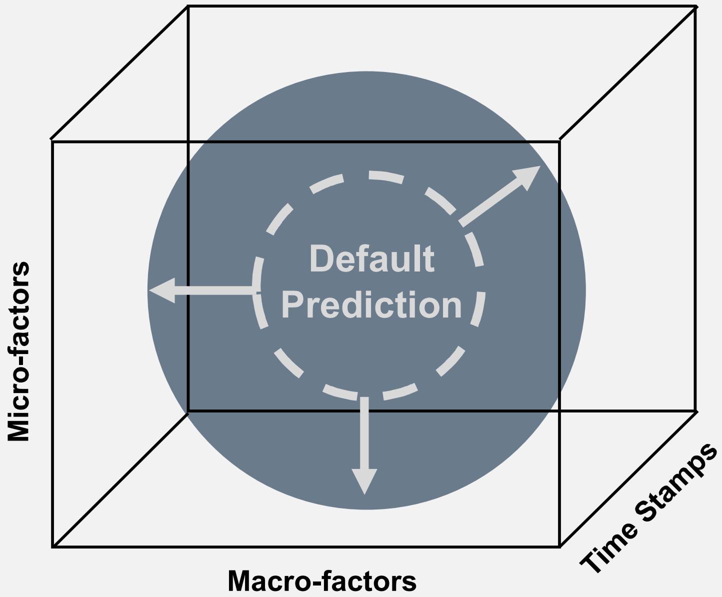
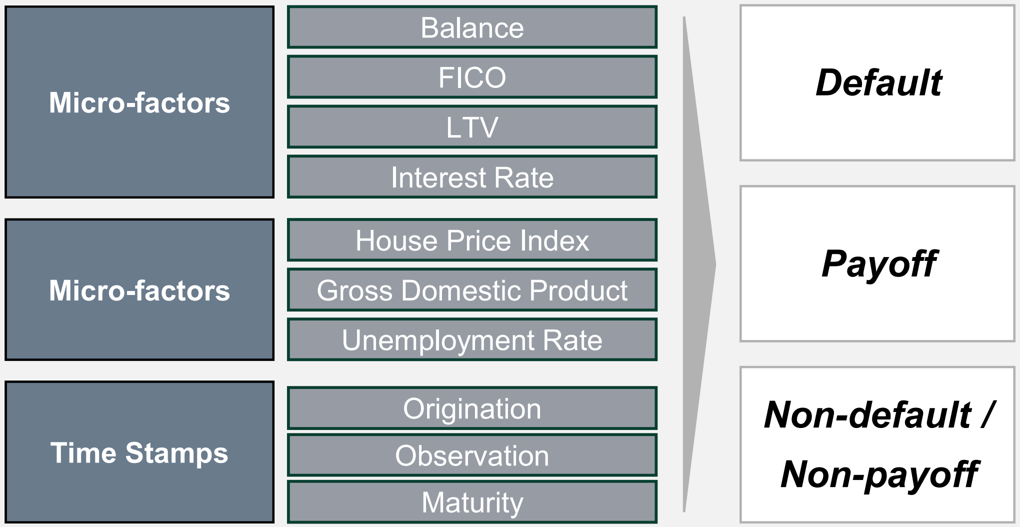
Prediction of Mortgage Loan Defaults

Table of Contents

Table of Contents

The purpose of this project is to understand the most important drivers of defaults and build classification models to investigate the effects on loan performance





Backgrounds

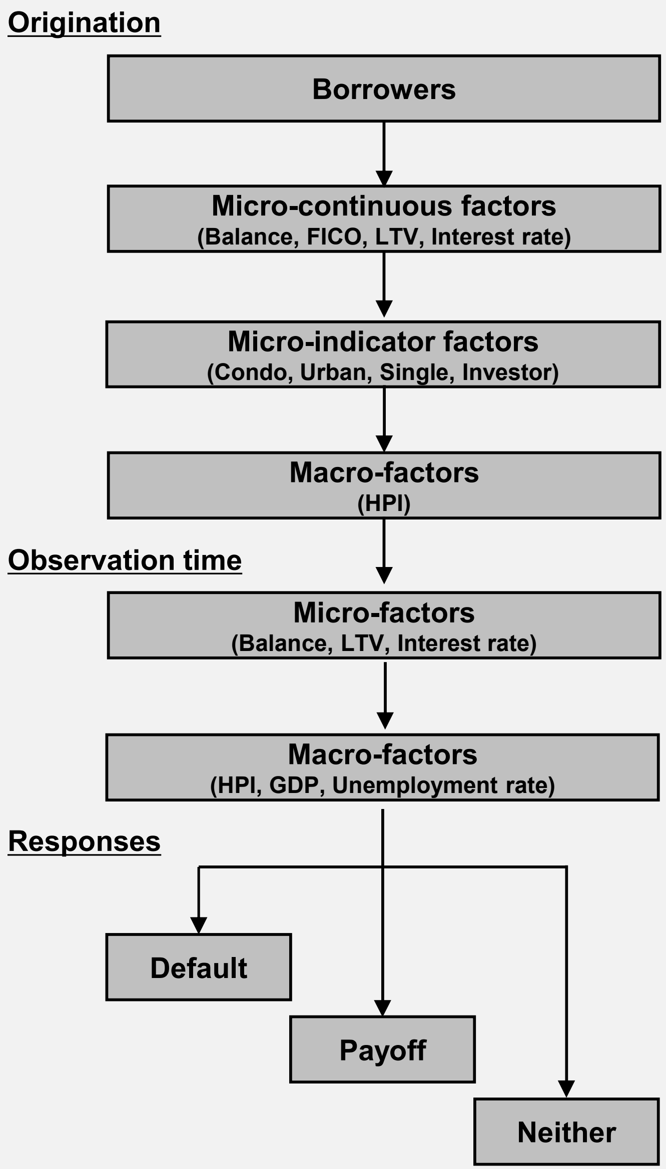
- The performance of a mortgage is likely driven by a wide range of factors including borrower information, loan characteristics and macroeconomic effects
- We would like to understand the most important drivers of defaults and predict which borrowers are likely to default at the time of mortgage origination

Project Descriptions

- We have built classification models to investigate the effects on loan performance of at least 12 variables
- This credit analysis is based on information specific to the mortgage or property, macro-economic variables, variables capturing borrower past credit history, and loan history.

Table of Contents

Data set mortgage is in panel form covering 50,000 residential mortgages over 60 periods, with 23 variables and 622,489 time stamped observations



Variables	Descriptions	Types
id	Borrower ID	int64
time	Time stamp of observation	int64
first_time	Time stamp for first observation	int64
orig_time	Time stamp for origination	int64
mat_time	Time stamp for maturity	int64
balance_orig_time	Outstanding balance at origination time	float64
FICO_orig_time	FICO score at origination time, in %	int64
LTV_orig_time	Loan-to-value ratio at origination time, in %	float64
interest_Rate_orig_time	Interest rate at origination time, in %	float64
hpi_orig_time	House price index at origination time, base year = 100	float64
balance_time	Outstanding balance at observations time	float64
LTV_time	Loan-to-value ration at observation time, in %	float64
interest_rate_time	Interest rate at observation time, in %	float64
hpi_time	House price index at observation time, base year = 100	float64
gdp_time	Gross domestic product growth at observation time, in %	float64
uer_time	Unemployment rate at observation time, in %	float64
REtype_CO_orig_time	Real estate type condominium = 1, otherwise = 0	int64
REtype_PU_orig_time	Real estate type planned urban development = 1, otherwise = 0	int64
REtype_SF_orig_time	Single-family home = 1, otherwise = 0	int64
investor_orig_time	Investor borrower = 1, otherwise = 0	int64
default_time	Default observation at observation time	int64
payoff_time	Payoff observation at observation time	int64
status_time	Default(1), payoff(2), and neither(0) observation at observation time	int64

(Source: http://www.creditriskanalytics.net/datasets-private2.html)

We wrangled *Dataset 2* – cross sectional dataset of 17 variables (10 selected, 7 wrangled), taking time series elements of the raw panel dataset

#	Selected Variables	Wrangled Variables	Wrangled Descriptions	Expected Relationship
	id			
1	time	**seasoning_orig_x**	Seasoning = No. months loan serviced at x = first_time – orig_time	Negative
	first_time			
2	orig_time			Neutral (Control)
3	mat_time	**Term_orig**	Loan term at origination (e.g. 20, 25, 30 yr loans) = mat_time – orig_time	Positive
4	balance_orig			Positive
5	FICO_orig			Positive
6	LTV_orig			Positive
7	Interest_Rate_orig			Positive
8	hpi_orig			Pos/Neg
9		**uer_time_x**	Unemployment rate at observation time, in %	Pos/Neg
	balance_time			
	LTV_time			
10	interest_rate_time	**IR_change_orig_y**	Change in interest rate from origination to last observation (%)	Positive
11	hpi_time	**HPI_change_orig_y**	Change in house price index from origination to last observation	Negative
	gdp_time			
12	uer_time	**uer_change_xy**	Change in unemployment rate from first to last observation (%)	Positive
13	REtype_CO_orig			Positive
14	REtype_PU_orig			Positive
15	REtype_SF_orig			Negative
16	investor_orig			Positive
17	default_time	**default_time_y**	Default outcome at last observation (Default = 1, Repaid = 0)	**Response**
	payoff_time			
	status_time			

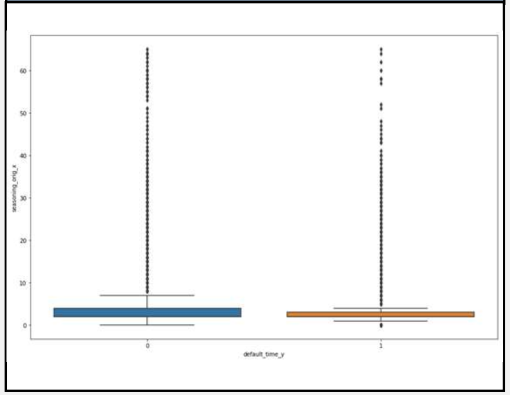
- Wrangled variables reflect change in raw time series variables from origination time to first observation (time **x**) or last observation (time **y**)
- Last observation (time **y**) gives the outcome on whether the loan defaulted or was repaid
- The selected variable “orig_time” is a **Control** variable for potential external factors
- The “neither” (0) observations from the raw data set were removed (reducing the 50,000 mortgages down to **41,736**) then repaid (2) replaced with (0) to create binary default response variable
- We also suggest expected relationship with **Default** (Response), noting **Pos/Neg** means both directions justifiable

EDA was performed on *Dataset 2* starting with summary statistics

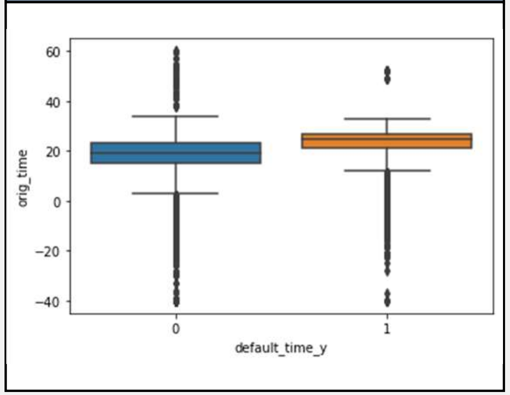
Group	#	Variables	Count	Mean	Std	min	25%	50%	75%	max	Types
Time	1	seasoning_orig_x	41736.0	3.53	5.22	0.00	2.00	2.00	3.00	65.00	int64
	2	orig_time	41736.0	20.02	7.35	-40.00	17.00	21.00	25.00	60.00	int64
	3	Term_orig	41736.0	117.58	13.61	5.00	120.00	120.00	121.00	201.00	int64
Micro-Macro	4	balance_orig	41736.0	253827.35	208366.23	0.00	117146.25	196750.00	336000.00	8000000.00	float64
	5	FICO_orig	41736.0	659.73	72.34	400.00	610.00	661.00	713.00	834.00	int64
	6	LTV_orig	41736.0	79.87	9.80	50.10	75.00	80.00	85.00	218.50	float64
	7	Interest_Rate_orig	41736.0	5.34	3.44	0.00	1.25	6.39	7.74	18.00	float64
	8	hpi_orig	41736.0	195.42	34.52	75.71	179.45	208.86	222.39	226.29	float64
	9	uer_time_x	41736.0	4.97	0.56	3.80	4.70	4.70	5.30	9.50	float64
Change in	10	IR_change_orig_y	41736.0	2.02	3.61	-11.88	0.00	0.00	4.50	37.50	float64
	11	HPI_change_orig_y	41736.0	-5.01	46.20	-79.84	-45.28	-4.04	30.15	149.61	float64
	12	uer_change_xy	41736.0	1.08	1.94	-0.30	-0.30	0.00	2.50	6.20	float64
Indicator	13	REtype_CO_orig	41736.0	0.07	0.25	0.00	0.00	0.00	0.00	1.00	int64
	14	REtype_PU_orig	41736.0	0.12	0.32	0.00	0.00	0.00	0.00	1.00	int64
	15	REtype_SF_orig	41736.0	0.62	0.48	0.00	0.00	1.00	1.00	1.00	int64
	16	investor_orig	41736.0	0.11	0.32	0.00	0.00	0.00	0.00	1.00	int64
Response	17	default_time_y	41736.0	0.36	0.48	0.00	0.00	0.00	1.00	1.00	int64

Boxplots show differences between the *Repaid* (default=0) and *Default* (default=1) groups for *hpi_orig* in particular then possibly *FICO_orig* and *Interest_Rate_orig*

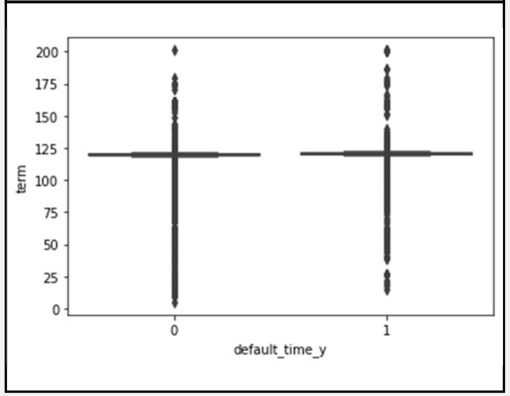
1 **seasoning_orig_x**



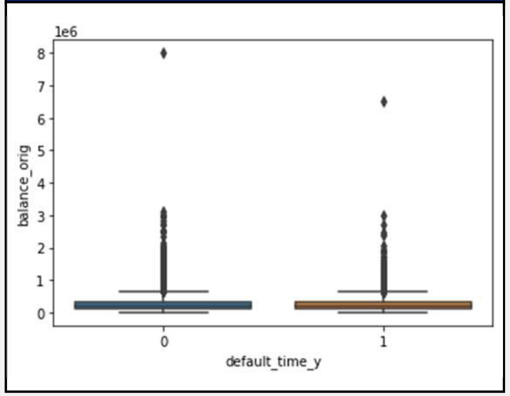
2 **orig_time**



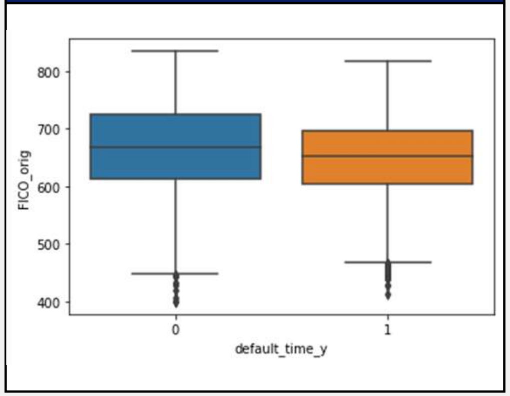
3 **term**



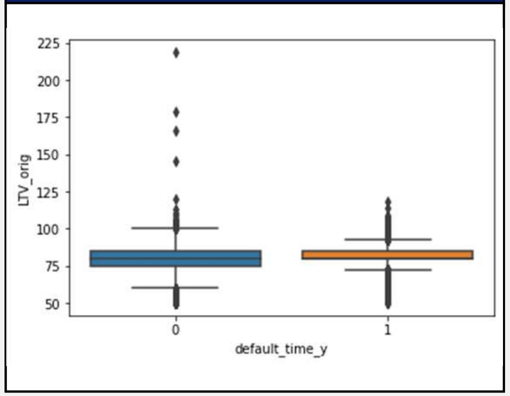
4 **balance_orig**



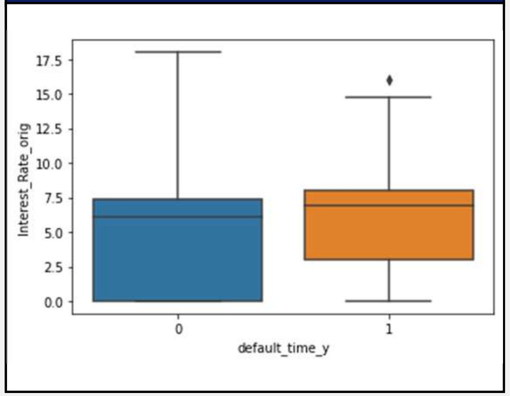
5 **FICO_orig**



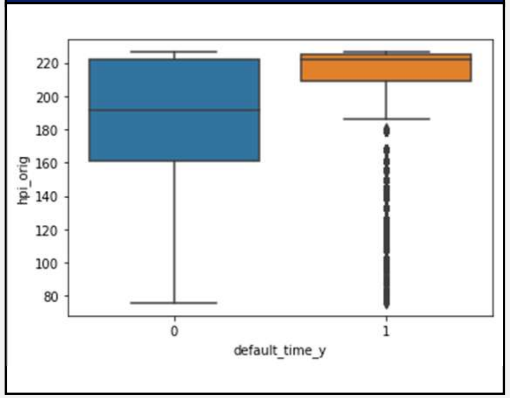
6 **LTV_orig**



7 **Interest_Rate_orig**



8 **hpi_orig**

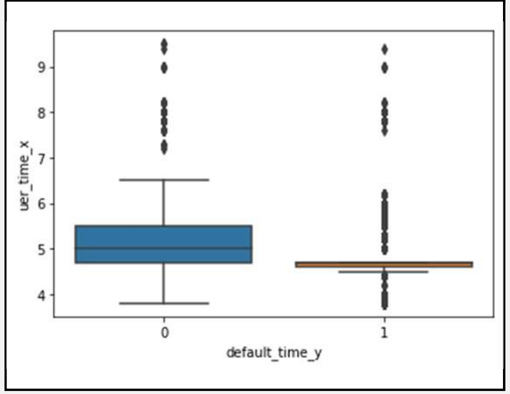


Time based variables

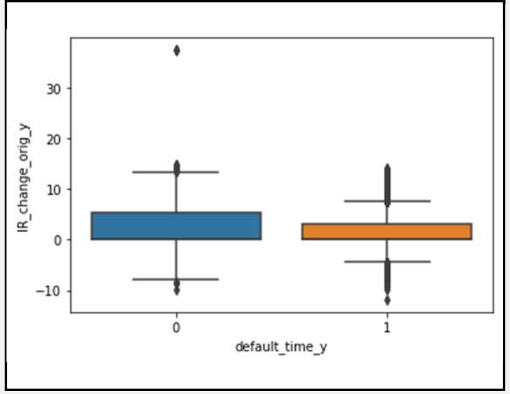
Macro-Micro based variables

Uer_time_x also likely important… "Change in" variables show strong explanatory potential (particularly *HPI_change* & *uer_change*), while Indicator variables appear to yield limited information

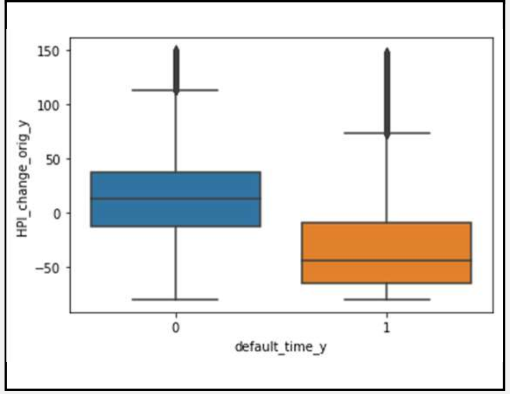
9 **uer_time_x**



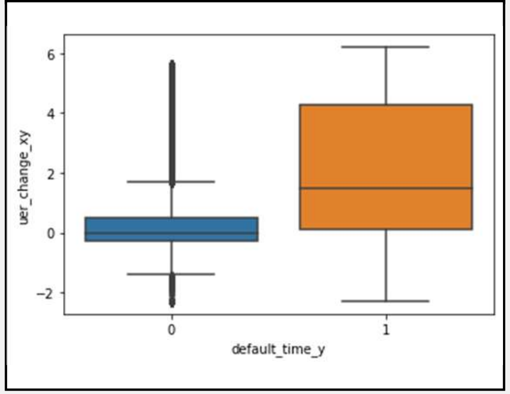
10 **IR_change_orig_y**



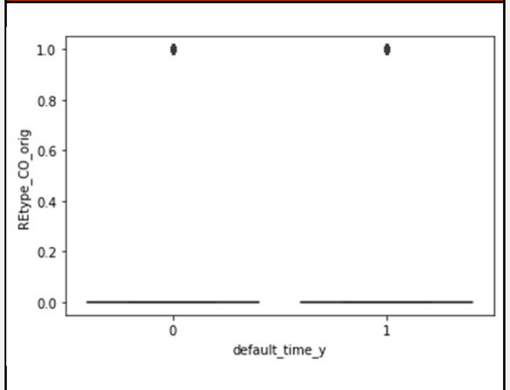
11 **HPI_change_orig_y**



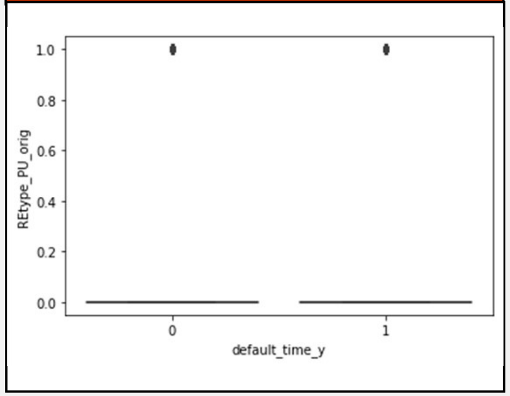
12 **uer_change_xy**



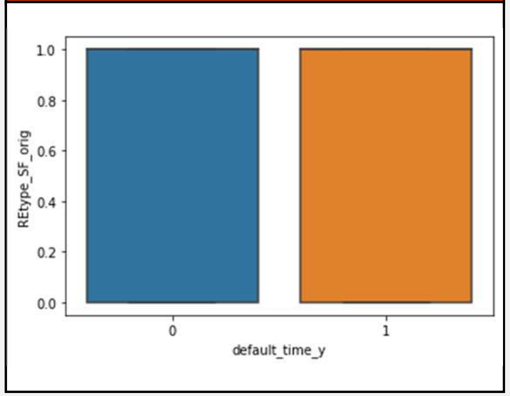
13 **REtype_CO_orig**



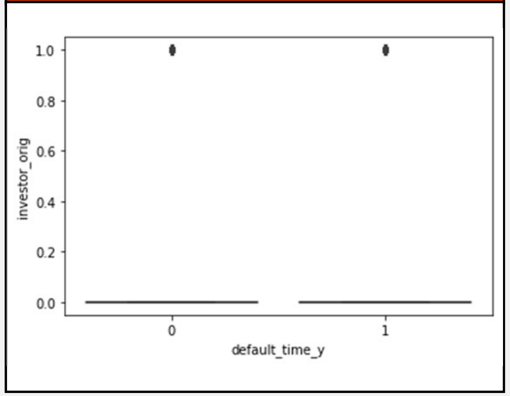
14 **REtype_PU_orig**



15 **REtype_SF_orig**



16 **investor_orig**



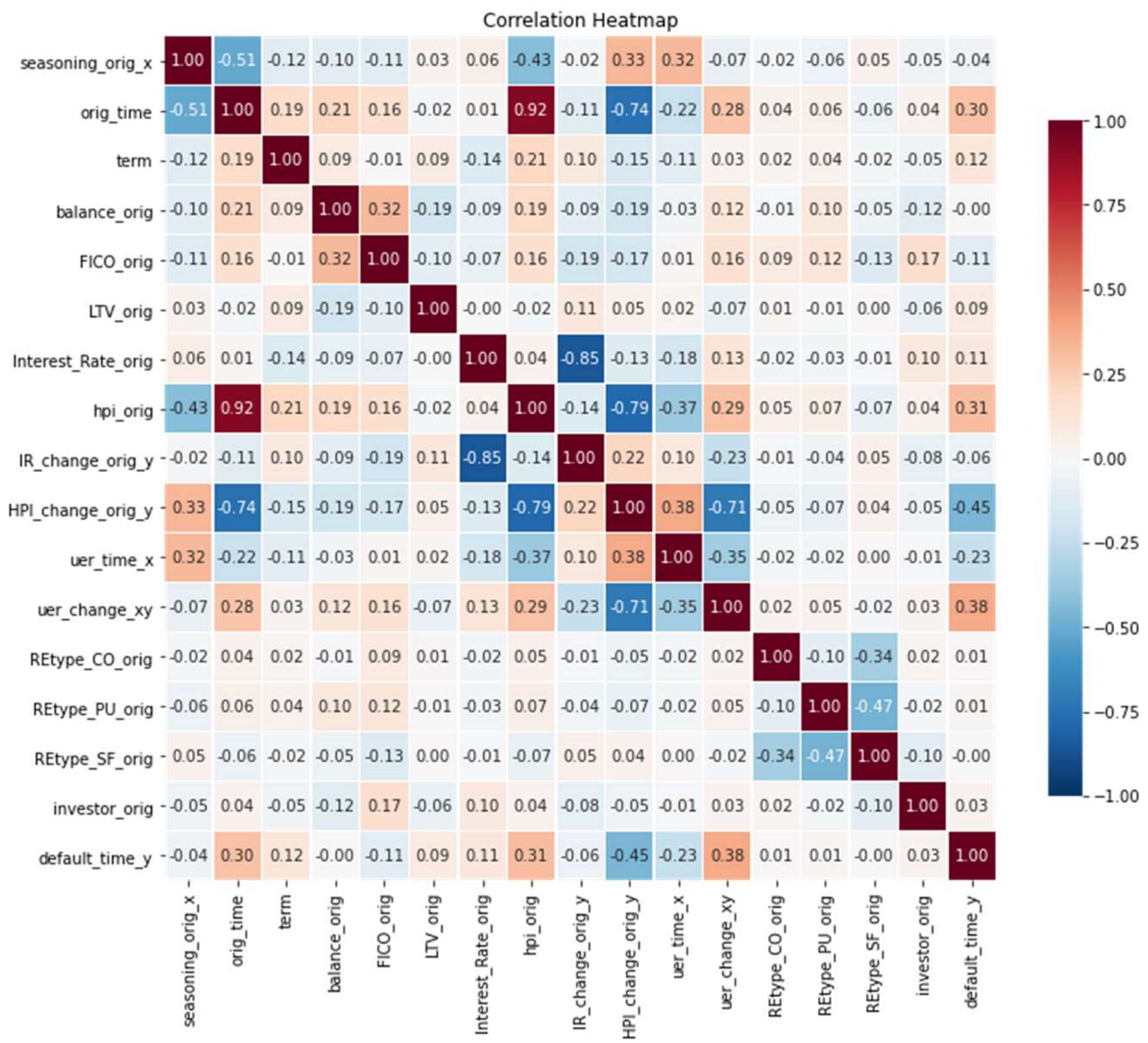
Macro-Micro based variables (cont.)

"Change in" variables

Indicator variables

We demonstrate the correlation of 17 variables through a heatmap and validate previous relationship expectation between response and independent variables

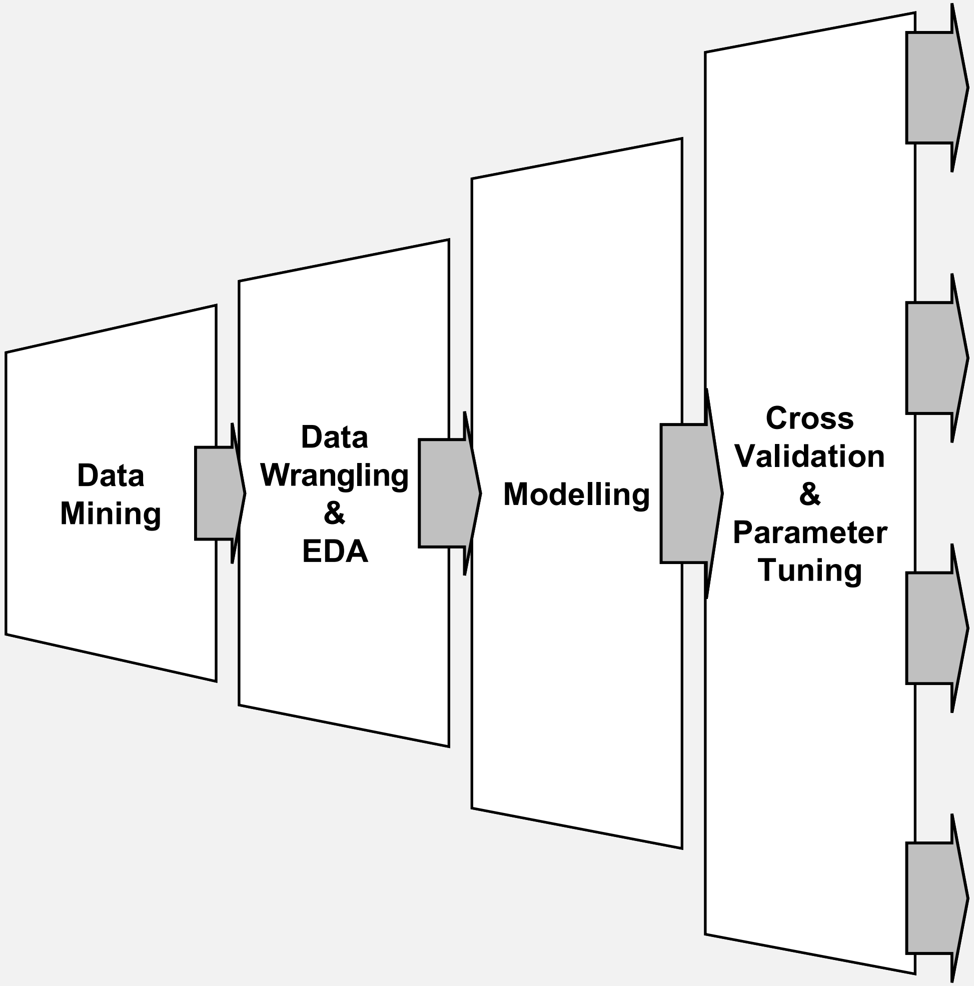
Heatmap of 17 variables



	Variables	Correlation	Expected Relationship
1	seasoning_orig_x	-0.04	Negative
2	orig_time	0.30	Neutral (Control)
3	term	0.12	Positive
4	balance_orig	-0.00	Positive
5	FICO_orig	-0.11	Positive
6	LTV_orig	0.09	Positive
7	Interest_Rate_orig	0.11	Positive
8	hpi_orig	0.31	Positive/Negative
9	uer_time_x	-0.23	Positive/Negative
10	IR_change_orig_y	-0.06	Positive
11	HPI_change_orig_y	-0.45	Negative
12	uer_change_xy	0.38	Positive
13	REtype_CO_orig	0.01	Positive
14	REtype_PU_orig	0.01	Positive
15	REtype_SF_orig	-0.00	Negative
16	investor_orig	0.03	Positive
17	default_time_y	1.00	Response

Table of Contents

We expect to answer the following scientific questions through the framework of our analysis



Primary Question

- What variables can be used in predicting mortgage defaults and what classification models provide the best prediction?

Supporting Question 1

- How does the economic environment, or change in macro-environment over time, impact mortgage performance?

Supporting Question 2

- Can accurate default predictions be achieved without expectations of a future change in the macro-economic environment?

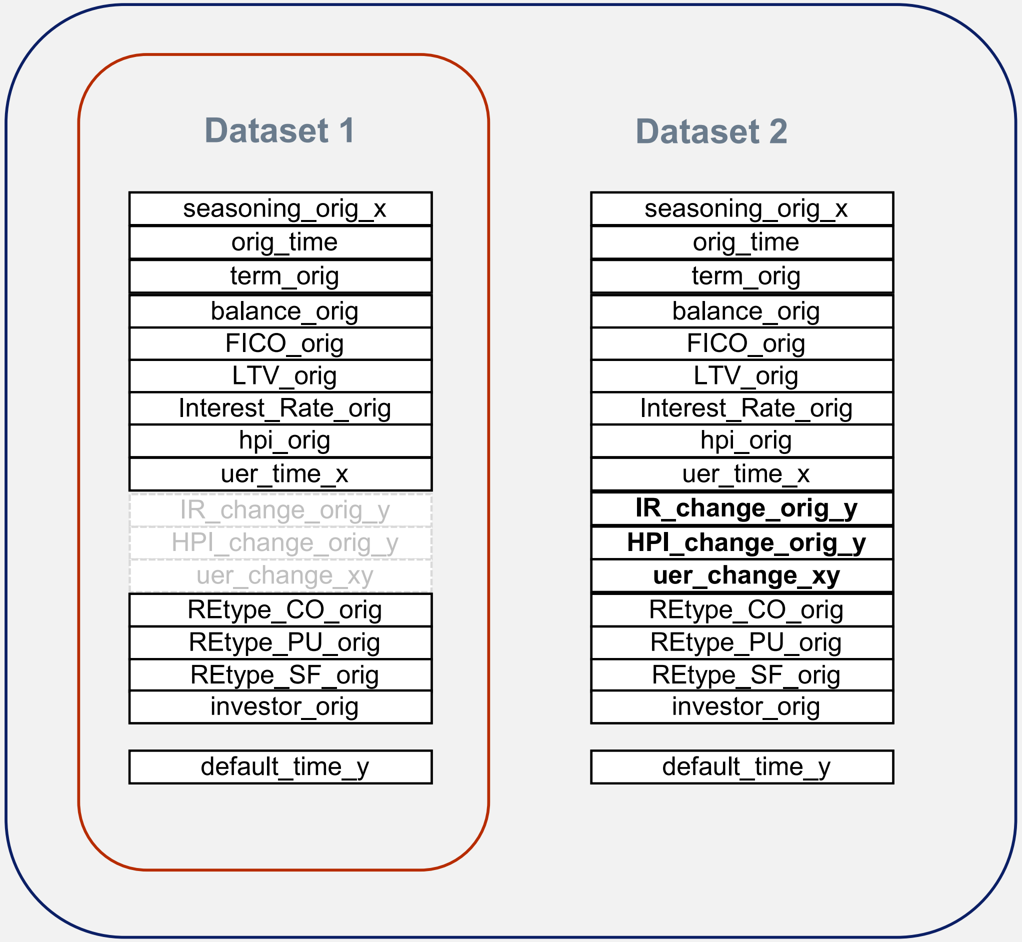
Supporting Question 3

- To what extent does additional loan and borrower information assist in identifying potential defaults?

  (i.e. wide range variables more useful than small key variables)

Table of Contents

We utilise both subset *Dataset 1* and *Dataset 2* to perform analytical methods to observe effect of newly created predictors and answer the research questions



- *Dataset 1* is a subset of *Dataset 2* to test for the predictive power of modelling without the macroeconomic related "change in" variables, noting these variables would require forecasting of future values when being used for prediction in the real world, introducing additional variance

We have built eight classification models suggested below, which will be tuned further

Decision Trees

- Progressively split data based on values of key features (at nodes) until and an endpoint (leaf) is reached
- Leaf provides the class of the datapoint

Random Forest

- Machine learning technique that simulates large number of decision trees during training
- Output is the class selected by most trees

Boosting

- Similar to random forest in producing many decision trees
- Additional step of using learnings from one tree in growing the next through accounting for the error prediction of the previous tree

Logistic Regression

- Models log odds as a linear function of X variables
- Log odds translated into probabilities for classification relative to determined threshold
- Makes no assumptions on probability distribution of predictors

Linear Discriminant Analysis (LDA)

- Finds linear combinations of features that best separate classes in a data set
- Relies on predictor variables being normally distributed and classes having common variance, estimated by within-sample covariance

Quadratic Discriminant Analysis (QDA)

- Operates similarly to LDA but is more flexible given variance is estimated by the sample covariance of each class
- Again, assumes X variables are normally distributed

Naïve Bayes

- Conditional probability classification model that (naively) assumes independence between X variables
- Often performs well in real world despite the above
- Also relies on X variables being normally distributed

K-th Nearest Neighbor (KNN)

- Non-parametric supervised learning technique (no assumptions on distributions or variances of features)
- Simply classifies a new datapoint based on distance from its k closest datapoints

Tree-based / Ensemble models

Linear models

Probabilistic / Neighbor-base models

After building models, we are to perform Cross Validation and Tuning Parameters to enhance modeling performances

Process	Methods
Cross Validation	▸ With exception of the tree-based models, models are cross validated using 100 iterations of Monte Carlo Cross Validation ▸ Within each iteration, data is split 80% training to 20% test ▸ Due to random forest and boosting being computationally expensive, 10-fold cross validation is performed on these models
Variable Selection & Threshold Probability	▸ Variables have been selected using domain knowledge. For logistic regression, a subset of variables have been refined using stepwise regression ▸ Parametric models may require further transformations of independent variables (e.g. log transformation) to satisfy assumed distributions ▸ For models requiring specification of threshold probability for classification, the proportion of defaults in the training set (~36%) has been taken on base models ▸ Other levels of threshold probability to be iterated through for tuning
Tuning Parameters (Random Forest)	▸ 10 iterations of each parameter will be allowed for in the following values: - Number of trees (ntree): 100, 200, …., 900, 1000 - Number of variables randomly sampled as candidates at each split (mtry): 1, 2, …, 9, 10 - Minimum size of terminal nodes (nodesize) : 1, 2, …, 9, 10
Tuning Parameters (Boosting)	▸ 10 iterations of each parameter will be allowed for in the following values: - Number of trees (ntrees): 100, 200, …., 900, 1000 - Learning rate (shrinkage): 0.01, 0.02, …., 0.09, 0.1 - Maximum depth of each tree (interaction.depth): 1, 2, …, 9, 10

Modeling Performances

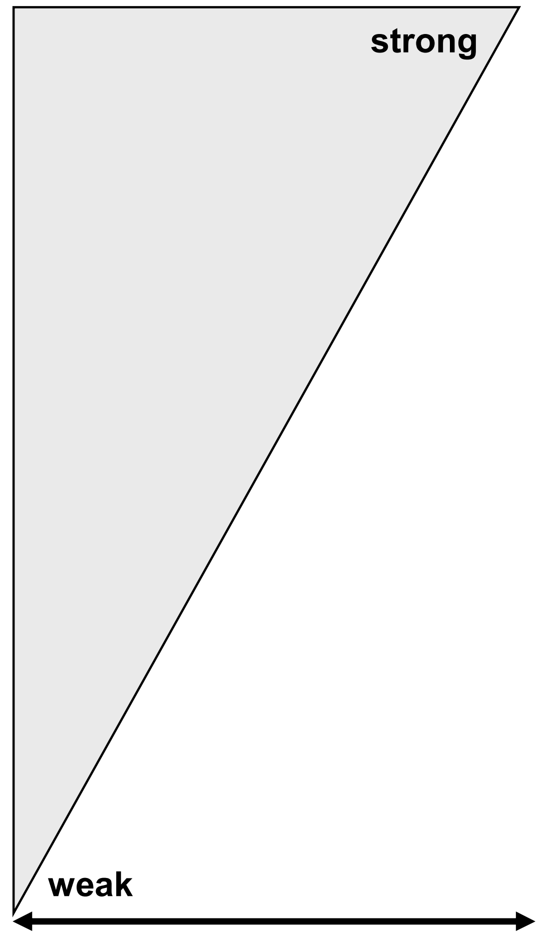


Table of Contents

In models using *Dataset 1* (12 predictors, without any "change in" variables), we found that testing errors are generally moderately high, with Boosting showing the lowest test error value

Single Decision Tree

- Test Error = 0.2773
- Tree pruning was conducted with optimal complexity parameter (cp) value of 0.01
- **orig_time** or loan origination time was the key variable

Random Forest

- Test Error = 0.2651
- Optimal values for mtry was 3, which was the same as the default value that R calculates
- **FICO_orig** and **balance_orig** were the two most important predictors., i.e. the FICO score and loan balance at origination.

Boosting

- Test Error = 0.2587
- Learning rate was set to 0.05 and number of trees to 5000.
- **orig_time** was the key variable
- Further parameter tuning too computationally expensive at time of presentation production

Logistic Regression

- Test Error = 0.2949
- Stepwise Logistic Regression was also conducted, but it produced the same optimal model in terms of test error

LDA

- Test Error = 0.3011
- Columns had to be converted back to numerical values instead of factors

QDA

- Test Error = 0.3248
- Columns had to be converted back to numerical values instead of factors

Naïve Bayes

- Test Error = 0.3355
- While tuning of the Laplace smoother is possible, it has not been attempted at this stage

KNN

- Test Error = 0.3670
- KNN with CV for k values of 1-15 was attempted, with k=15 producing the lowest test error

Tree-based / Ensemble models

Linear models

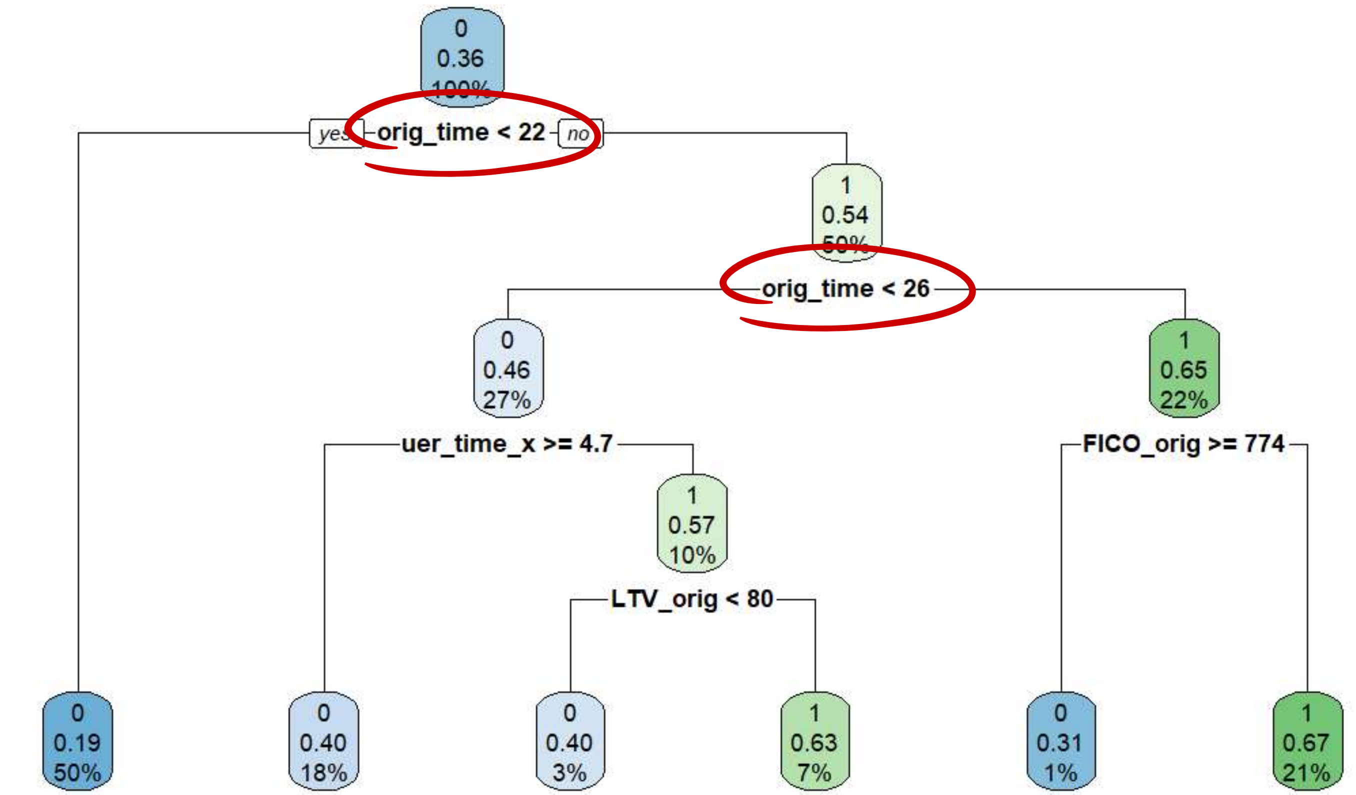
Probabilistic / Neighbor-based models

*Results are preliminary and subject to further cross validation and tuning

In Single Decision Tree modelling, the first two nodes are divided from the *orig_time,* highlighting the importance of this variable to *Dataset 1*

- The interpretation of control variable *orig_time* will be discussed in under Conclusions (section 6)

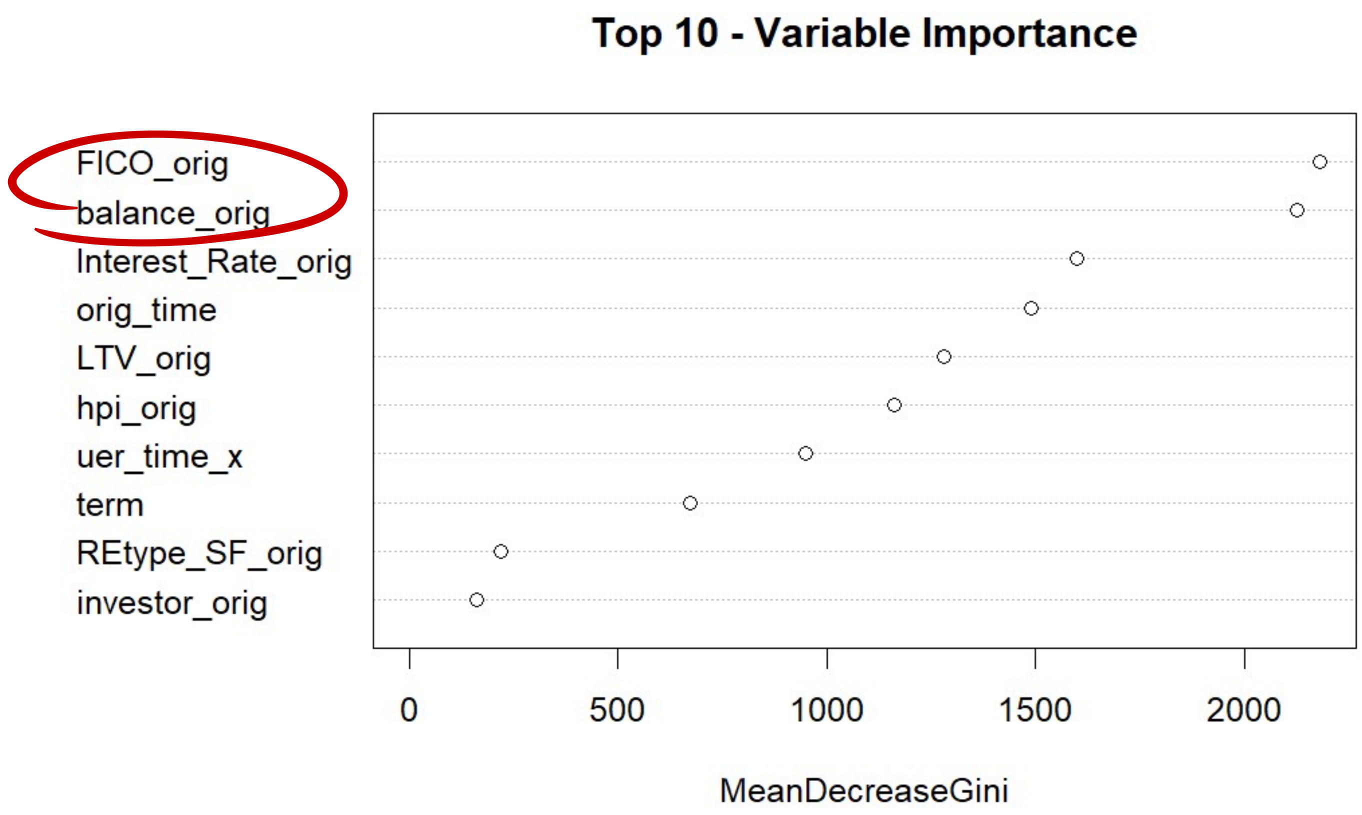
Single Decision Tree (Dataset 1)



*Results are preliminary and subject to further cross validation and tuning

In Random Forest modelling using *Dataset 1*, *FICO_orig* and *balance_orig* are the most important predictors

Top 10 Variable Importance of Random Forest (Dataset 1)



*Results are preliminary and subject to further cross validation and tuning

In Gradient Boosting modelling using *Dataset 1*, *orig_time* is the most important predictor, with *balance_orig* and *FICO_orig* also of importance

Variable Importance of Boosting (Dataset 1)

	var <chr>	rel.inf <dbl>
orig_time	orig_time	22.3824792
balance_orig	balance_orig	16.5357246
FICO_orig	FICO_orig	15.8755718
hpi_orig	hpi_orig	10.7199370
Interest_Rate_orig	Interest_Rate_orig	10.6748320
LTV_orig	LTV_orig	10.3636928
uer_time_x	uer_time_x	6.0150319
term	term	5.3063772
investor_orig	investor_orig	0.8336410
REtype_SF_orig	REtype_SF_orig	0.5160390

*Results are preliminary and subject to further cross validation and tuning

In the second round of modelling using *Dataset 2* (16 predictors, including "change in" variables), we found improved model outcomes overall and Boosting again the top performer

Single Decision Tree

- Test Error = 0.2308
- Tree pruning was conducted with optimal complexity parameter (cp) value of 0.01
- **HPI_change_orig_y** or change in HPI was the key variable

Random Forest

- Initial Test Error = 0.2115, Tuned Test Error = 0.2103
- Optimal values for mtry=2, and ntree=500 were found through tuning
- **HPI_change_orig_y** or change in HPI was the most important predictor by far

Boosting

- Initial Test Error = 0.2115, Tuned Test Error = 0.2086
- Learning rate (shrinkage) parameter values of 0.3, 0.1, 0.05, 0.01, 0.005 were attempted, with 0.05 performing the best and number of iterations = 2124. Other parameter tuning pending final report
- **HPI_change_orig_y** was again the most important variable

Logistic Regression

- Test Error = 0.2378
- Stepwise Logistic Regression was also conducted, but the test error turned out slightly higher than the default model.

LDA

- Test Error = 0.2375
- Columns had to be converted back to numerical values instead of factors.

QDA

- Test Error = 0.2335
- Columns had to be converted back to numerical values instead of factors.

Naïve Bayes

- Test Error = 0.2781
- While tuning of the Laplace smoother is possible, it was not attempted.

KNN

- Test Error = 0.3173
- KNN with CV for k values of 1-15 was attempted, with k=5 producing the lowest test error.

Tree-based / Ensemble models

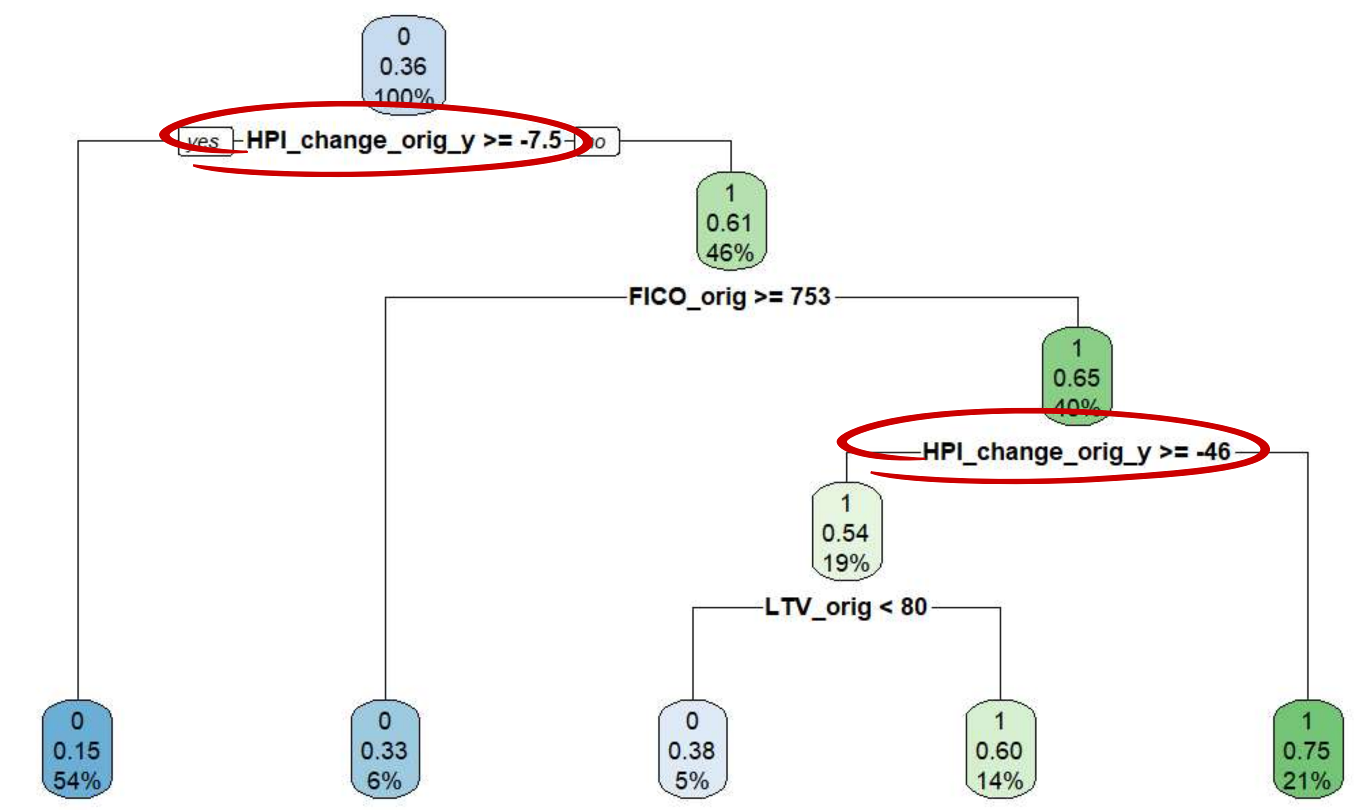
Linear models

Probabilistic / Neighbor-based models

*Results are preliminary and subject to further cross validation and tuning

In Single Decision Tree modelling using Dataset 2, the first and third nodes are divided on *HPI_change_orig_y*, *FICO_orig* also important

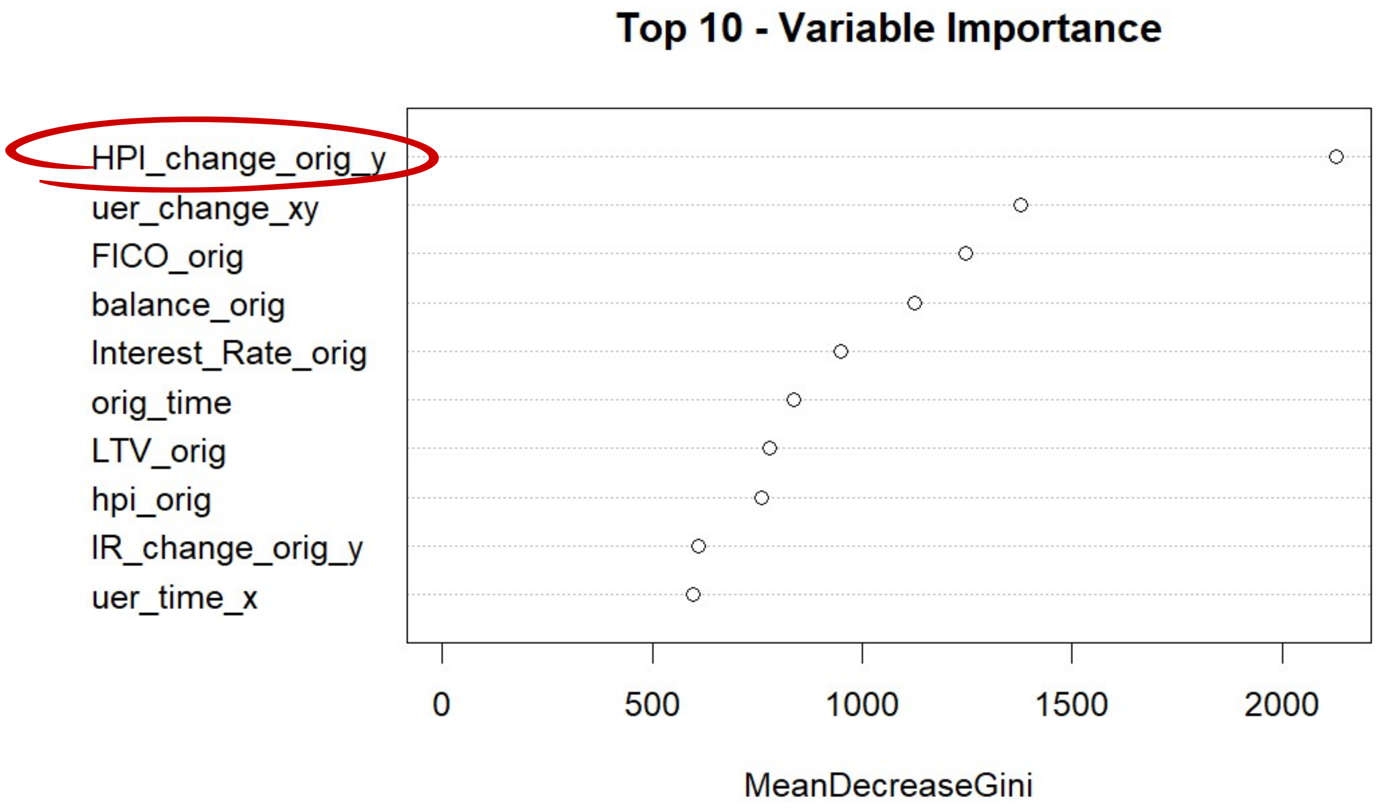
Single Decision Tree (Dataset 2)



*Results are preliminary and subject to further cross validation and tuning

In Random Forest modelling *HPI_change_orig_y* is the most important predictor by far to *Dataset 2*, followed by *uer_change_xy and FICO_orig*

Top 10 Variable Importance of Random Forest (Dataset 2)



*Results are preliminary and subject to further cross validation and tuning

In Gradient Boosting modelling using *Dataset 2*, *HPI_change_orig_y* is again the most important predictor by a substantial margin, followed by *FICO_orig* and *balance_orig*

Variable Importance of Boosting (Dataset 2)

	var <chr>	rel.inf <dbl>
HPI_change_orig_y	HPI_change_orig_y	39.7311245
FICO_orig	FICO_orig	12.1817687
balance_orig	balance_orig	9.5086875
Interest_Rate_orig	Interest_Rate_orig	6.8440288
LTV_orig	LTV_orig	6.6726356
IR_change_orig_y	IR_change_orig_y	6.6163561
uer_change_xy	uer_change_xy	6.5549772
uer_time_x	uer_time_x	2.6314375
orig_time	orig_time	2.4996181
seasoning_orig_x	seasoning_orig_x	2.1814978

*Results are preliminary and subject to further cross validation and tuning

Table of Contents

Modelling using the additional variables in *Dataset 2* outperformed *Dataset 1*, with Boosting using Dataset 2 achieving the least test error value

	Single Decision Tree	Random Forest	Boosting	Logistic Regression	LDA	QDA	Naïve Bayes	KNN
Dataset 1 (12 predictors)	**0.2773**	**0.2651**	**0.2587**	**0.2949**	**0.3011**	**0.3248**	**0.3355**	**0.3670**
Dataset 2 (16 predictors)	**0.2308**	**0.2103**	**0.2086**	**0.2378**	**0.2375**	**0.2335**	**0.2781**	**0.3173**

Observations

- For both datasets, Boosting performed best in terms of the classification error on the testing set; Random Forest was the second best in both cases
- *Dataset 2* with four additional predictors resulted in lower test error for all eight models. This means that the inclusion of change variables was likely useful for capturing more information with predictive power
- The *change in the House Price Index* from Dataset 2 was the most important variable for all tree-based models
- For *Dataset 1*, *Loan Origination time* was consistently important (but not always the most important)

Interpretation

- Because orig_time was highly correlated with *HPI_change_orig_y* (-0.74) and *HPI_orig* (0.92), it is possible that it is capturing similar macroeconomic information, making it an important variable to *Dataset 1*. Yet, it is evident that the change in HPI during the life of the loan carries additional information given its importance in modelling *Dataset 2*.
- Despite the boost in performance for all models using the additional variables in *Dataset 2*, when making predictions in the real world, the future change in HPI will be a (potentially unreliable) forecast such that actual performance will likely suffer higher variance
- Reliance on *Loan Origination time* in *Dataset 1* for ***future*** prediction would also require caution as it will involve extrapolating into the future. This variable was included to control for ***historical*** factors. A scenario without this variable will be included in the final report.

*Conclusions are preliminary and subject to further model cross validation and tuning

Answering our research questions through data mining

	Scientific Questions	Answers
Primary Question	What variables can be used in predicting mortgage defaults and what classification models provide the best prediction?	Most of our variables selected through domain knowledge added some predictive power to the modelling, but the key variable was *change in house price index*, followed by *credit score (FICO score)* and *loan balance at origination*
Supporting Question 1	How does the economic environment, or change in macro-environment over time, impact mortgage performance?	With *change in house price index* the most important variable by far, change in macro-environment over time is the key determinant in mortgage performance. We see this variable as a macro proxy given *unemployment rate at origination* was not available in the dataset (only *unemployment rate at each time observation)*
Supporting Question 2	Can accurate default predictions be achieved without expectations of a future change in the macro-economic environment?	Circa 1 in 4 predictions on *Dataset1* were incorrect (and these predictions benefited from inclusion of *orig_time*). We expect that a bank would like higher accuracy than this. Nonetheless a number of variables that do not rely on foresight (eg FICO score, loan balance, LTV) proved useful in constructing a base model.
Supporting Question 3	To what extent does additional loan and borrower information assist in identifying potential defaults?	Models with more variables (*Dataset 2*) performed better, though this was due to the change in macro variables per above. Incremental loan and borrower information such as the *REtype* variables actually yielded limited value. We suspect however that more variables would be available to a bank such that a clear answer cannot be provided to this question based on our dataset.

*Conclusions are preliminary and subject to further model cross validation and tuning

The End of The Presentation