UNIVERSITY of NICOSIA

Department of Digital Innovation

MSc in Blockchain and Digital Currency

BLOC 526 – Emerging topics in fintech

Session 10 - Machine learning case studies

Dr. Periklis Thivaios CFA, FRM, BTRM

Learning objectives and expected learning outcomes

Summary

Session 10 builds up on the theoretical foundations outlined during the previous session in order to illustrate two practical case studies on the use of machine learning techniques for credit modelling in the banking sector and fraud identification in the insurance sector. The benefits and downsides of such techniques will be illustrated and compared to more traditional statistical analyses.

Learning objectives

- Machine learning techniques that can assist with real life fintech challenges
- Benefits and downsides of such techniques and the challenges that fintech firms need to overcome in using them

Expected learning outcomes

- Understand the potential use of techniques such as decision trees, random forest, gradient boosting and neural networks in fintech applications
- Develop a critical evaluation of the benefits and downsides of machine learning techniques for credit modelling and insurance fraud identification

Agenda

- Introduction and recap
- Machine learning in credit analysis
- Machine learning in insurance fraud identification
- Concluding remarks

Agenda

- **Introduction and recap**
- Machine learning in credit analysis
- Machine learning in insurance fraud identification
- Concluding remarks

Summary of takeaways from previous session

1. Machine learning is a 'new' toolkit for predictive analytics	▪ Enhanced methodologies, greater processing power and big data allow us to solve problems previously too cumbersome to do so
2. Machine learning opens up opportunities for fintechs	▪ The fintech applications of machine learning are numerous ▪ Fintechs can become enablers for incumbent institutions, or disruptors where appropriate
3. Machine learning should be applied with caution!	▪ Despite the possibilities, machines may not be as intelligent as we think ▪ Or, even worse, machines may be more honest than humans…

You don't need to know the techniques, the math or the software (well, ideally you do…)
But, if you want to make it in the fintech world, you should at least use the term somewhere! (and ideally understand what it means…)

Question: What are some practical applications of machine learning to banking and insurance?

Banking	Insurance
Credit analysis	**Fraud identification**

You don't need to know the techniques, the math or the software (well, ideally you do…)
But you should be able to question the grand statements offered by marketing decks and aspiring millionaires

Agenda

- Introduction and recap
- **Machine learning in credit analysis**
- Machine learning in insurance fraud identification
- Concluding remarks

This section is based on the Masters Dissertation of Natalie van Niekerk, from North West University, supervised by Periklis Thivaios

Question: What is credit scoring?

A credit score is a numerical expression based on a level analysis of a person's credit files, to represent the creditworthiness of an individual

1900s

- Systematic data on borrowers barely existed
- Local credit bureaus scoured newspapers for notices of arrests, marriages, promotions etc

1956

- William Fair and Earl Isaac started using statistical techniques to predict the probability that a borrower could default
- The first scorecard was made of cardboard and shared with Loan Officers who could fill them in with applicants' information to see if they exceeded an acceptable level of risk

1970s

- The Fair Credit Reporting Act (FCRA) required credit bureaus to report information only to those with a legitimate purpose, obliged them to ensure accuracy, and gave consumers the right to see and correct their files
- The Equal Credit Opportunity Act (ECOA) of 1974 made it unlawful for lenders to discriminate on the basis of sex or marital status
- In 1976 it was amended to outlaw the consideration of race, religion and several other characteristics.

1989

- Working with Equifax, Experian and TransUnion (three credit bureaus that had come to dominate the market), Fair Isaac unveiled the first consumer-credit score: a number between 300 and 850
- Higher scores indicate a better credit rating
- Known as the FICO (for Fair Isaac Corporation) score, it rapidly became the standard for American lenders.

Source: The Economist

Applying machine learning to credit analysis

We compared machine learning approaches to traditional logistic regression to evaluate strengths and weaknesses

1. Data exploration
2. Feature engineering
3. Check redundancy in data and remove irrelevant variables
4. Split data into train, validation and out-of-time data sets
5. Oversample
6. Implement variable selection and transformation
7. Model fitting
8. Model validation and evaluation
9. Comparative evaluation of results

The analysis applied traditional and machine learning approaches to data, variables and models

	Data	Variable selection	Model fitting
'Traditional' approaches	Raw data Manually engineered features	Stepwise selection	Logistic regression
Machine learning approaches	Machine engineered features	Decision tree Random forest Gradient boosting	Decision tree Random forest Neural network

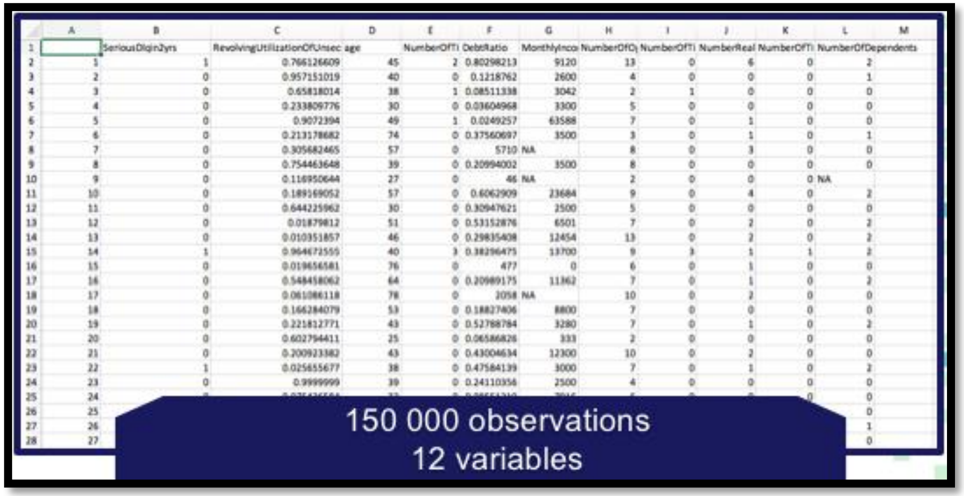
Which approach yields better discriminatory results?

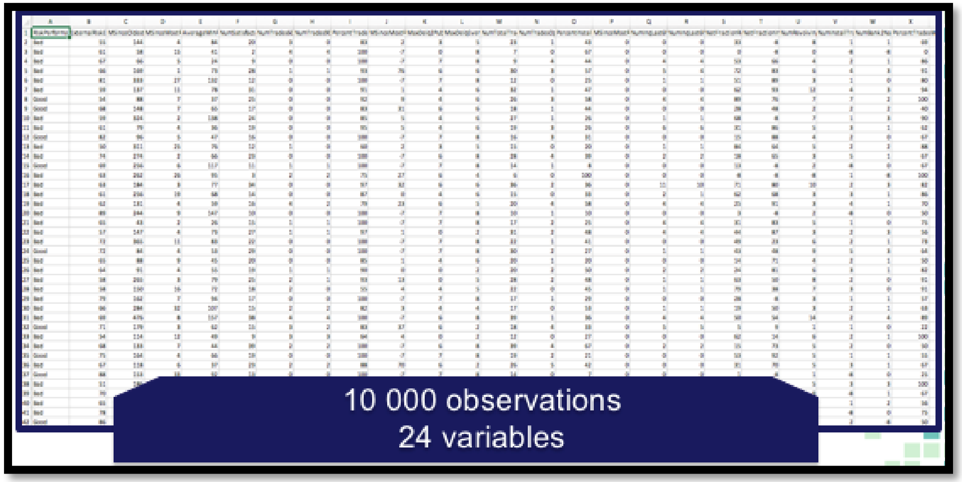
1. Data exploration

- Essential first element in all data analyses!

- Before we delve into quantitative techniques, we need to get a good feeling about the data

  - Are there any observable patterns?
  - Are there any extreme outliers?
  - Does the data make sense?





2. Feature engineering

Feature engineering refers to creating new features from the existing raw variables

Raw Data	Manual Feature Engineering	Automated Feature Engineering
The original dataset	The creation of new features for modelling, based on raw features Involves combining or splitting existing features into new with greater predictive power	The creation of new features using algorithmic aggregation and transformation Employed Python's 'featuretools' library

We built three models based on 1. raw data; 2. manually engineered features; and 3. automatically engineered features

3. Data redundancy and remove irrelevant variables

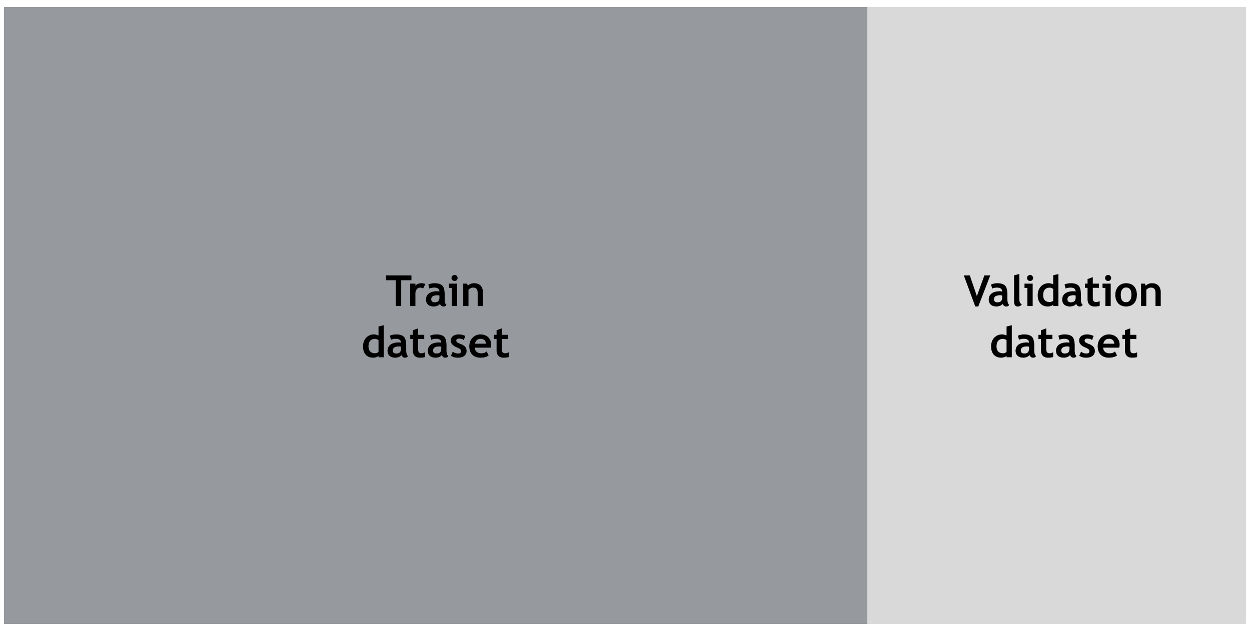
- Redundant variables are variables whose values can be determined by other variables. For example, when two variables are highly correlated, one of the variables is redundant

- Irrelevant variables are variables which cannot be used to predict the target variable. For example, variables with many missing values or with only one level

	Raw Data	Manual Feature Engineering	Automated Feature Engineering
# variables removed	0	922	1046
Final # of variables	82	273	290

4. Train, validation and out-of-time data sets

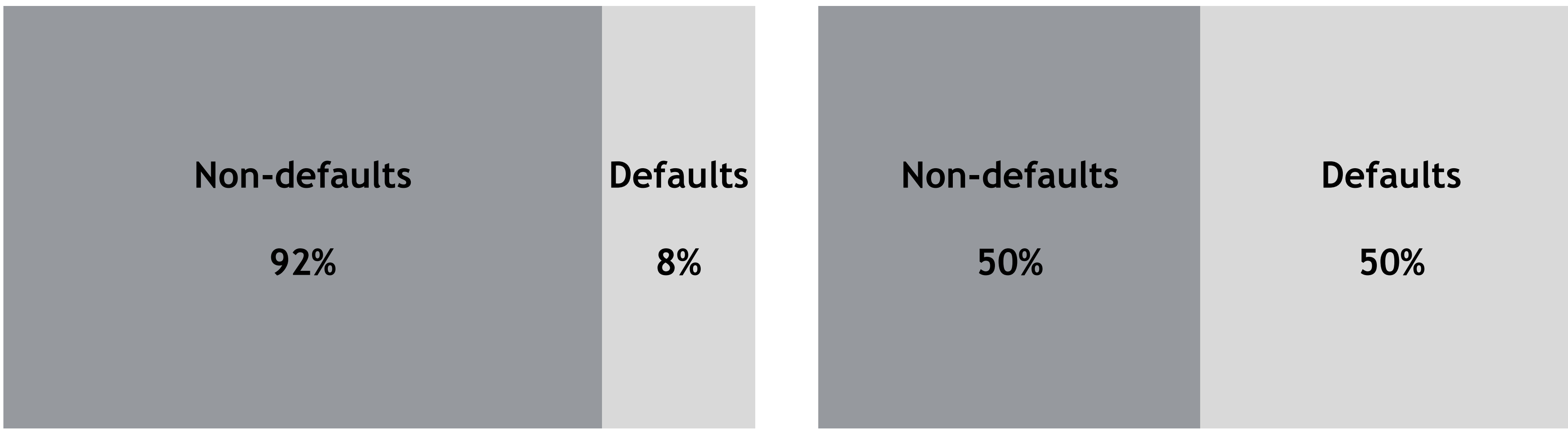
- The training data is the sample of data on which the model is developed and the validation data is the sample used to independently test the built model
- According to Siddiqi (2017), the dataset should be split into 70 to 80 percent training data and 20 to 30 percent validation data.



5. Oversample

- The target variable was imbalanced, meaning that there was approximately 92% of applicants who managed to repay the loan on time and approximately 8% of applicants who did not manage to repay the loan on time
- According to Siddiqi (2017), a 50 percent bad rate is necessary for good scorecard development

Original dataset **Oversampled dataset**



6. Variable selection and transformation

- Variable selection was done using four different methods, stepwise selection (traditional), decision tree, random forest and gradient boosting
- The variables which were selected by each method was used to build a model to compare which variable selection results in the best model

Stepwise selection	Decision tree	Random forest	Gradient boosting
Includes and eliminates variables based on how significant the variable is. If a variable is added, but it does not contribute to the model, then it is eliminated in the next step. The process ends when all the variables selected are significant and all variables eliminated are not significant	The process of growing a decision tree is the variable selection method. Each tree has several nodes, and each node includes one selected variable. Only the variables used in the decision tree were used for modelling.	Combines many independent decision trees built using random samples. After all the trees are built, the random forest combines the outcomes to obtain the final model by ranking the variables. Only important variables were used and variables with zero importance were eliminated	Builds multiple decision trees by using multiple random samples and building trees independently. The gradient boosting algorithm ranks the variables in order of importance. Only the important variables were used for modelling and the variables with zero importance were eliminated

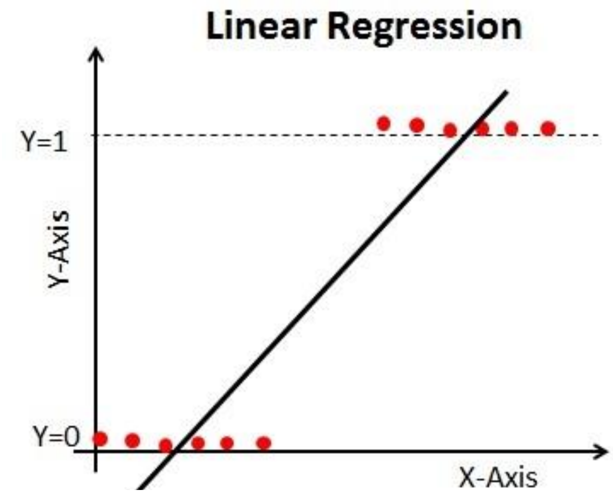
7. Model fitting – logistic regression

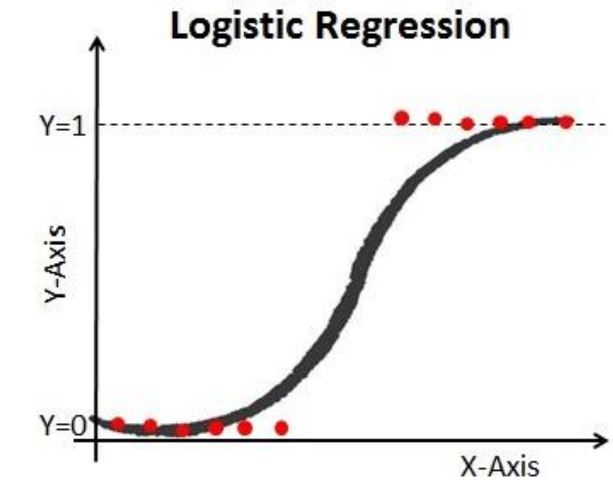
- Logistic regression belongs to the group called generalised linear models
- Logistic regression is related to linear regression, but where linear regression estimates the outcome of an event, logistic regression estimates the *odds* of an event occurring
- In logistic regression, the logistic function is used

- Where:

$$p(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}},$$

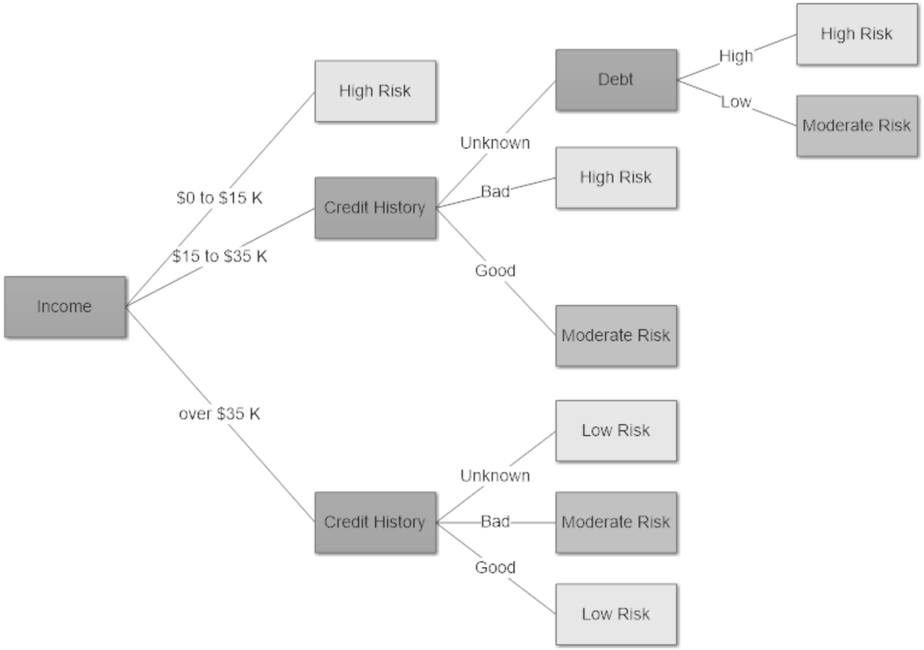
$$p(x) = \Pr(y = 1|x).$$





7. Model fitting – decision tree

- A decision tree is a supervised machine learning algorithm which can be divided into two categories: classification and regression
- A decision tree has a flow chart structure where each internal node represents an attribute, each branch represents a decision and each leaf represents a categorical or continuous outcome



Advantages

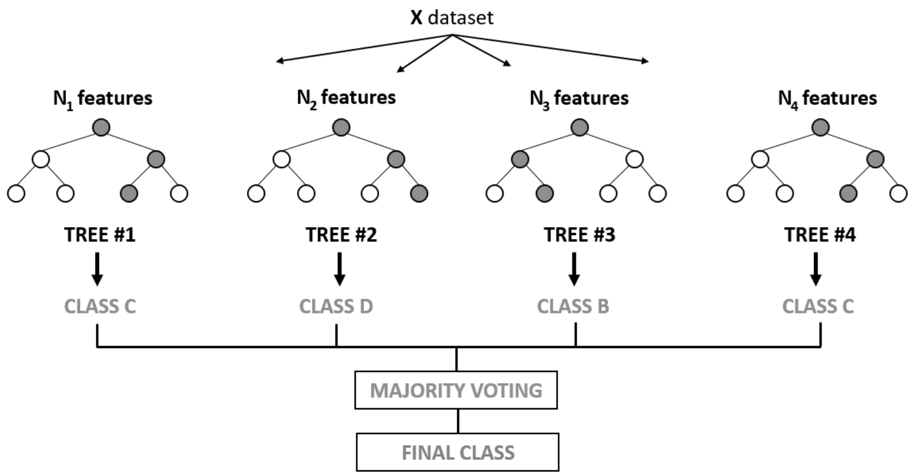
- Simple to understand, interpret and visualise
- Implicitly perform variable screening or feature selection
- Handles both numerical and categorical variables
- Require little effort for data preparation; and
- Non-linear relationships don't affect the performance of the model

Disadvantages

- Can create complex trees which do not generalise well – overfitting;
- Can be unstable because of small variations in the data; and
- In general trees do not have the same accuracy as other predictive models

7. Model fitting – Random forest

- The algorithm creates a forest with several decision trees. In general, the more trees in the forest, the more robust the prediction and the higher the accuracy
- The main difference between the decision tree and random forest algorithms, is that the process of finding the root node and splitting the internal nodes runs randomly in the random forest (Polamuri (2017))



Advantages

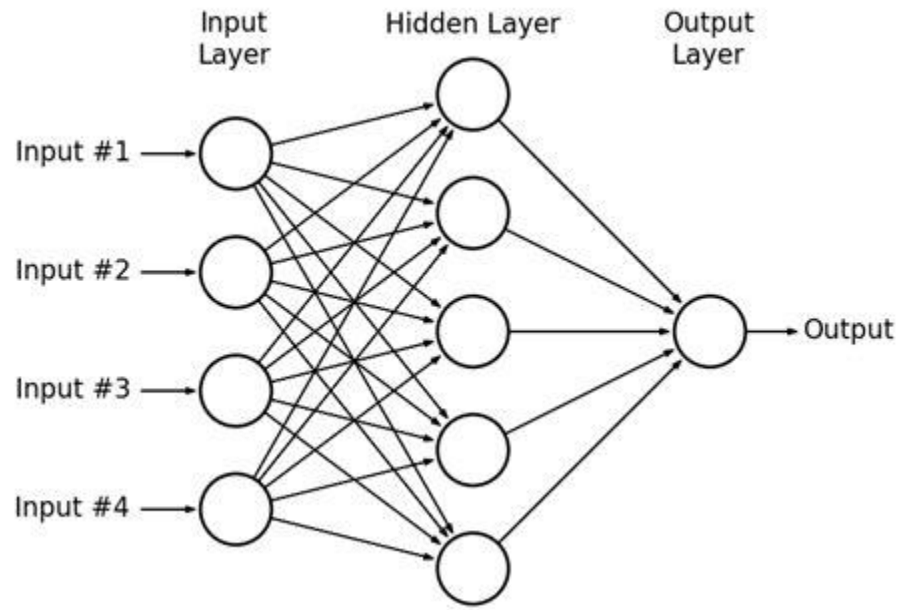
- Handles missing values and maintains accuracy regardless of missing data
- Does not overfit
- Power to handle large datasets with high dimensionality

Disadvantages

- Difficult to interpret
- Biased for attributes with more levels than other attributes

7. Model fitting – neural network

- A neural network involves many processors functioning analogous and ordered in steps (Rouse (2018))
- The first step receives the raw input information, like optic nerves in the human brain
- Each consecutive step receives the output from the step before it, rather than from the raw input. The last step creates the output of the model



Advantages

- Can handle a wide range of problems
- Give good results in complex domains
- Can handle all variable types
- Does not impose any restrictions on input data

Disadvantages

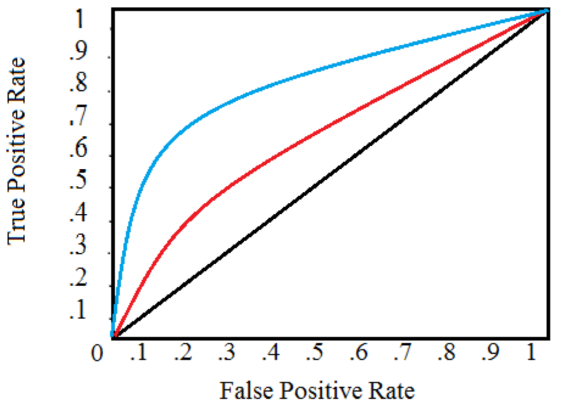
- Convergence to local minima
- Difficult to interpret
- All inputs and outputs should be in the (0,1) range

8. Model validation and evaluation

The scorecards were validated and evaluated using the following methods

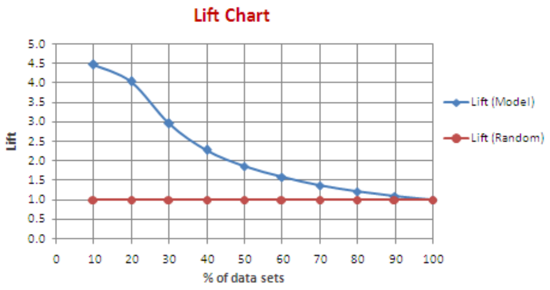
ROC curves

- Receiver Operating Characteristic (ROC) curves show the relationship between the false positive rate and true positive rate
- The greater the area under the ROC curve (AUROC), the better the model
- If the AUROC is 0.5, the model is just as good as a random selection



Lift charts

- The lift chart is a measure used to evaluate the effectiveness of a model
- The random selection model is represented by the horizontal line intersecting the y-axis at 1. The most effective model will have the greatest cumulative lift



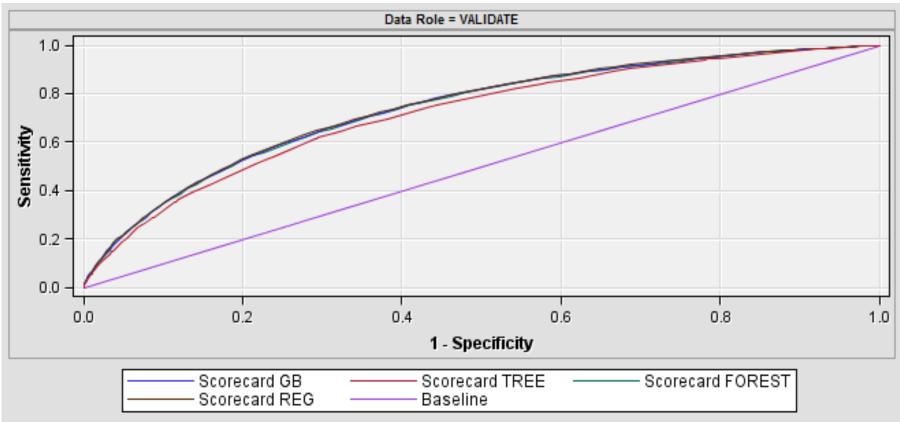
Confusion matrices

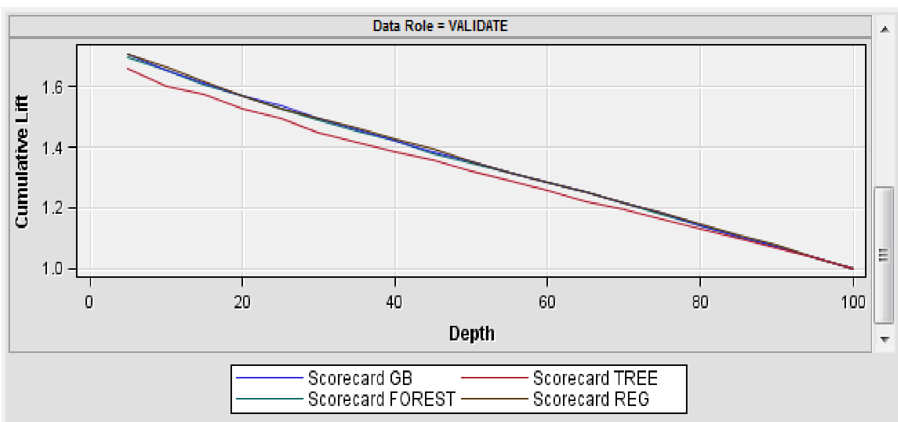
- A confusion matrix is a summary of model predictions and the type of errors the model makes. The terms used are true positive, false positive, true negative and false negative
- A perfect model has no false positives and false negatives. Practically, it is preferable to minimise false negatives

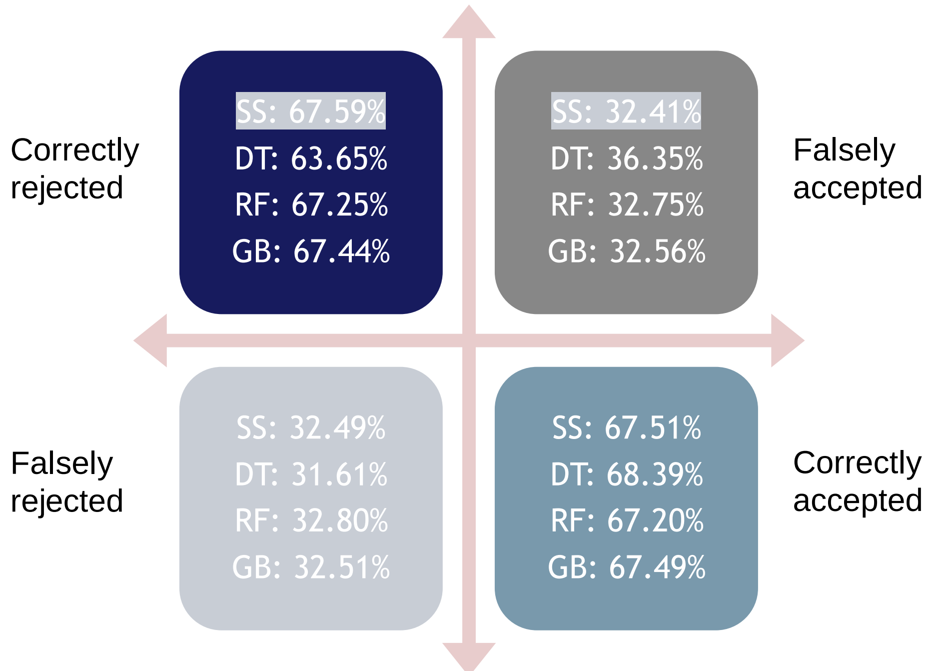
		Predicted Class	
		No	Yes
Observed Class	No	TN	FP
	Yes	FN	TP

TN	True Negative
FP	False Positive
FN	False Negative
TP	True Positive

9. Comparative evaluation of results
Raw dataset - logistic regression



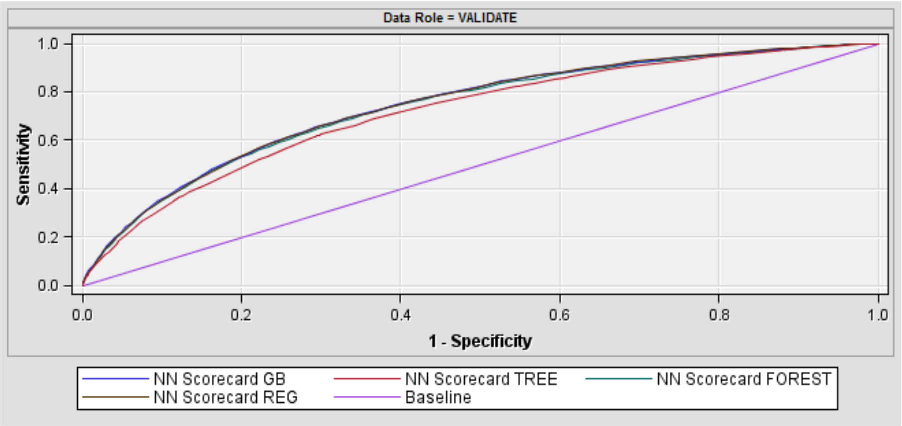


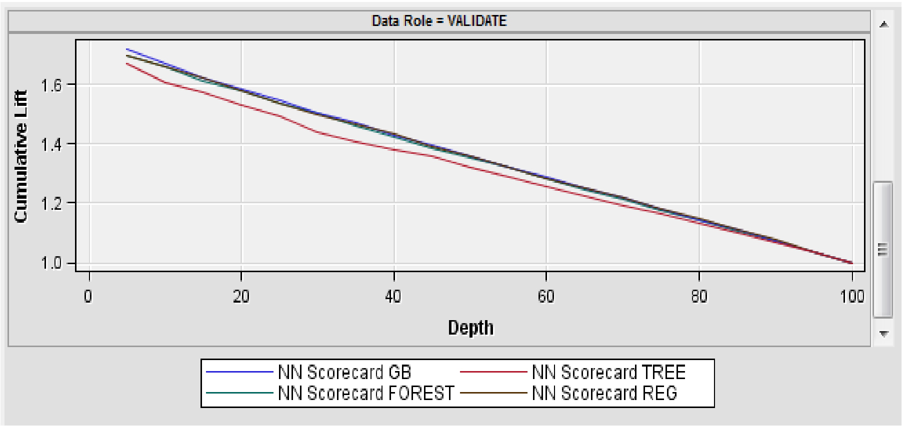


Best selection method

SS: Stepwise Selection
DT: Decision Tree
RF: Random Forest
GB: Gradient Boosting

9. Comparative evaluation of results
Raw dataset - neural network



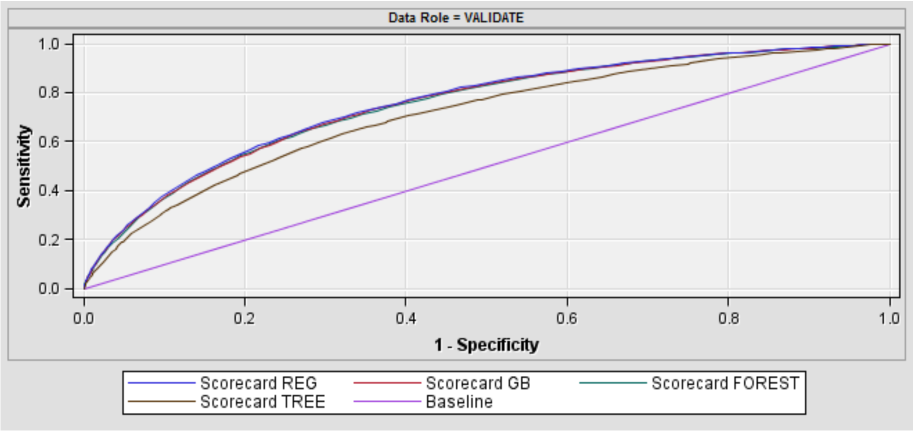


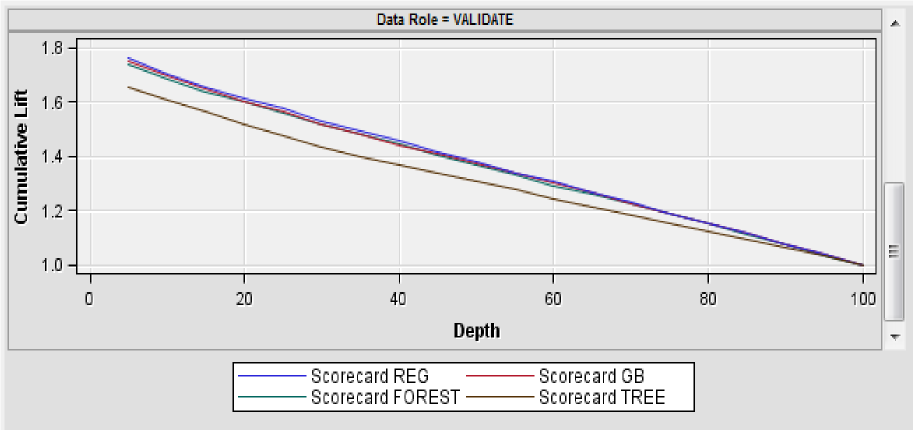
Correctly rejected	SS: 67.21% DT: 62.93% RF: 66.18% GB: 67.96%	SS: 32.79% DT: 37.07% RF: 33.82% GB: 32.04%	Falsely accepted
Falsely rejected	SS: 31.52% DT: 30.67% RF: 31.00% GB: 32.09%	SS: 68.48% DT: 69.33% RF: 69.00% GB: 67.91%	Correctly accepted

Best selection method

SS: Stepwise Selection
DT: Decision Tree
RF: Random Forest
GB: Gradient Boosting

9. Comparative evaluation of results
Manually engineered data - logistic regression



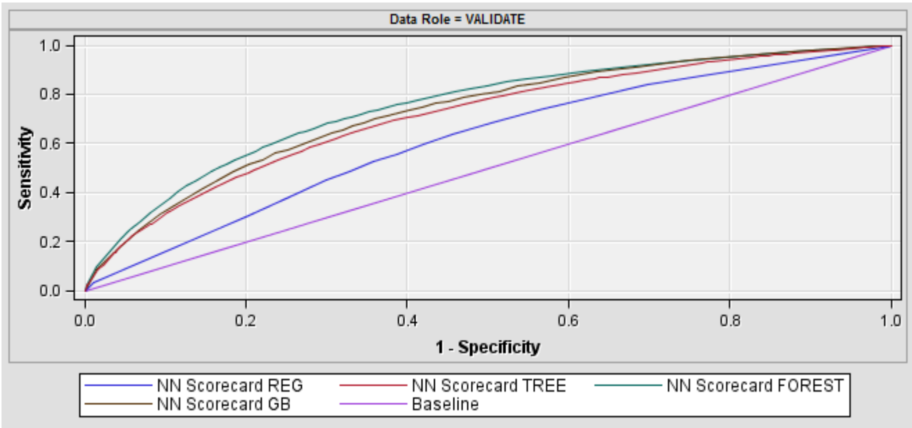


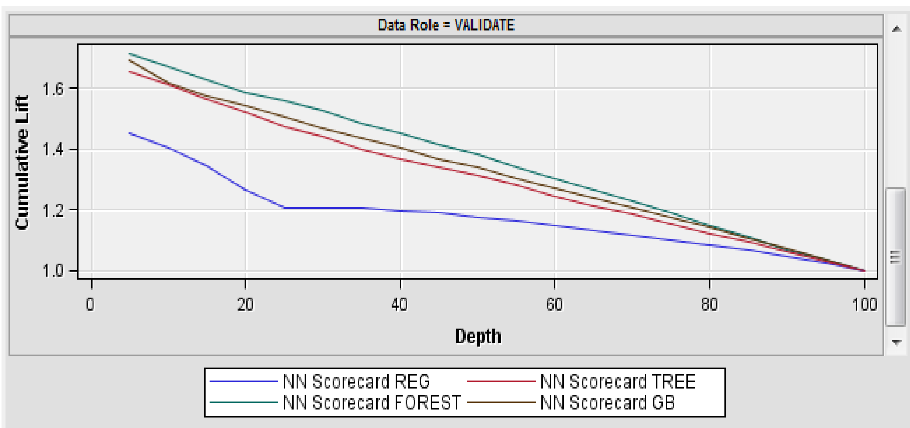
Correctly rejected	SS: 69.01% DT: 66.03% RF: 68.23% GB: 68.39%	SS: 30.99% DT: 33.97% RF: 31.77% GB: 31.61%	Falsely accepted
Falsely rejected	SS: 31.01% DT: 35.09% RF: 31.50% GB: 31.28%	SS: 68.99% DT: 64.91% RF: 68.50% GB: 68.72%	Correctly accepted

Best selection method

SS: Stepwise Selection
DT: Decision Tree
RF: Random Forest
GB: Gradient Boosting

9. Comparative evaluation of results
Manually engineered data - neural network



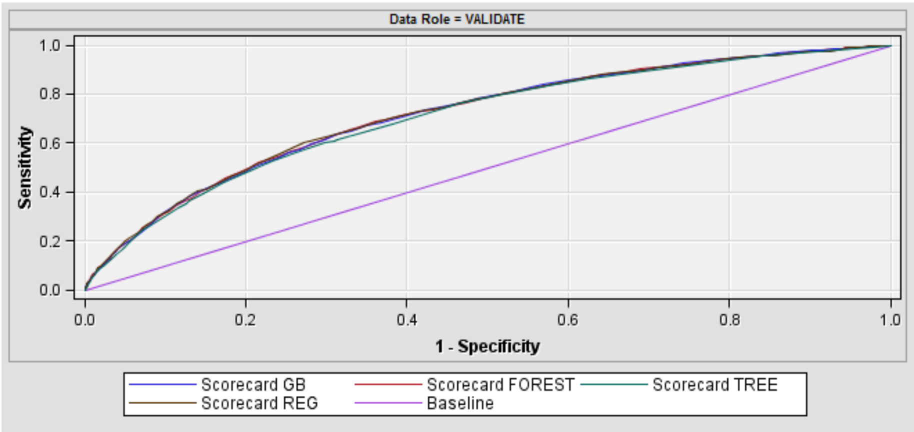


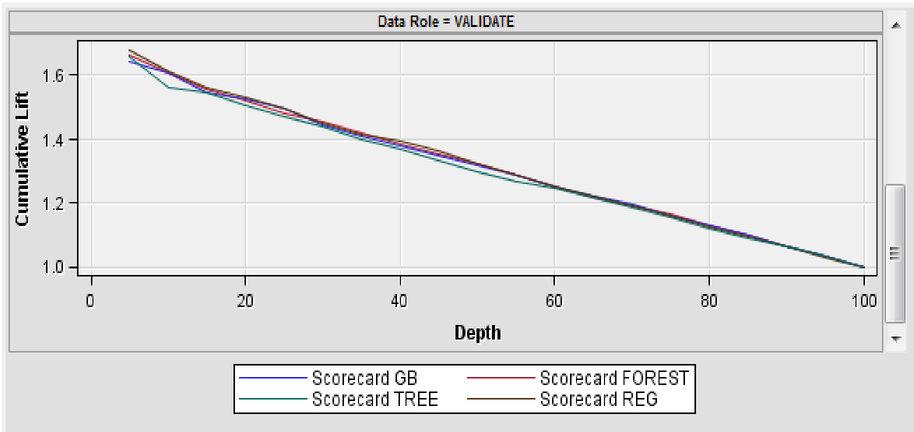
Correctly rejected	SS: 59.45% DT: 62.16% RF: 68.25% GB: 69.82%	SS: 40.55% DT: 37.84% RF: 31.75% GB: 30.18%	Falsely accepted
Falsely rejected	SS: 41.99% DT: 31.18% RF: 30.23% GB: 36.06%	SS: 58.01% DT: 68.82% RF: 69.77% GB: 63.94%	Correctly accepted

Best selection method

SS: Stepwise Selection
DT: Decision Tree
RF: Random Forest
GB: Gradient Boosting

9. Comparative evaluation of results
Automatically engineered data - logistic regression





Correctly rejected	SS: 64.81% DT: 61.23% RF: 65.14% GB: 65.06%	
		SS: 35.19% DT: 38.77% RF: 34.86% GB: 34.94%	Falsely accepted

Correctly rejected
- SS: 64.81%
- DT: 61.23%
- RF: 65.14%
- GB: 65.06%

Falsely accepted
- SS: 35.19%
- DT: 38.77%
- RF: 34.86%
- GB: 34.94%

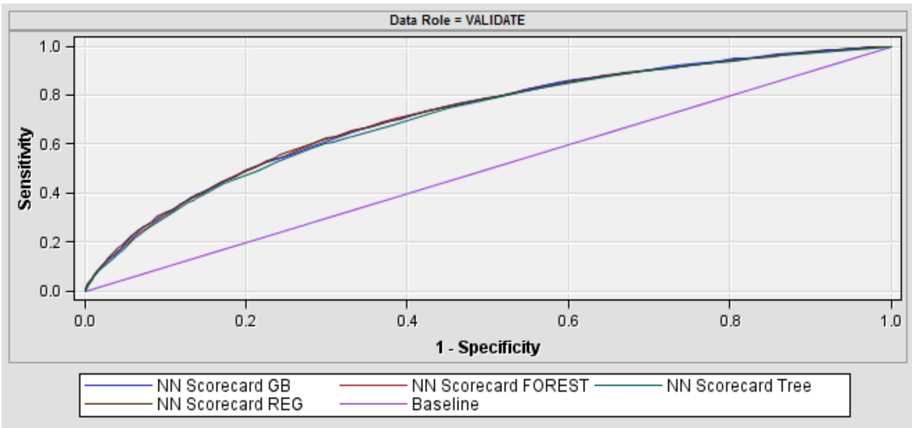
Falsely rejected
- SS: 32.53%
- DT: 31.37%
- RF: 32.92%
- GB: 33.00%

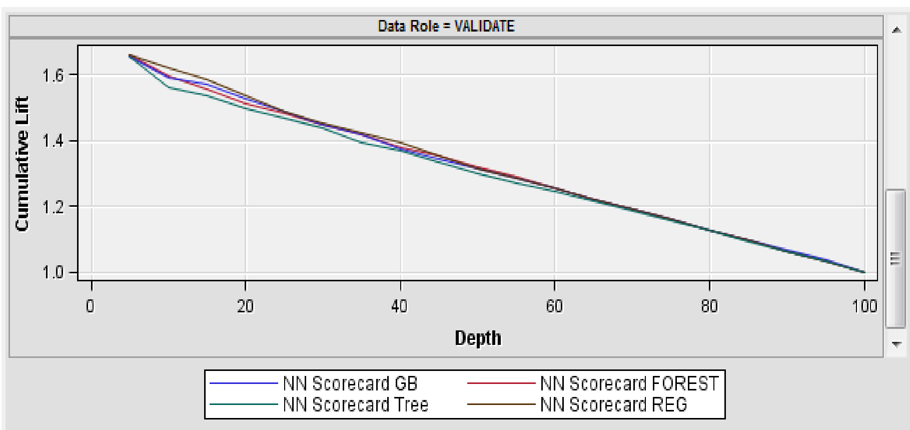
Correctly accepted
- SS: 67.47%
- DT: 68.63%
- RF: 67.08%
- GB: 67.00%

Best selection method

SS: Stepwise Selection
DT: Decision Tree
RF: Random Forest
GB: Gradient Boosting

9. Comparative evaluation of results
Automatically engineered data - neural network





Correctly rejected

SS: 63.33%
DT: 61.23%
RF: 63.16%
GB: 65.80%

SS: 36.67%
DT: 38.77%
RF: 36.84%
GB: 34.20%

Falsely accepted

Falsely rejected

SS: 31.38%
DT: 31.28%
RF: 31.13%
GB: 34.34%

SS: 68.62%
DT: 68.72%
RF: 68.87%
GB: 65.66%

Correctly accepted

Best selection method

SS: Stepwise Selection
DT: Decision Tree
RF: Random Forest
GB: Gradient Boosting

Conclusions

Data	Variable selection	Model fitting
Raw	Stepwise selection	Logistic regression
	Decision tree	Decision tree
Manual engineering		
	Random forest	Random forest
Automated engineering	Gradient boosting	Neural network

The model which performed best (correctly rejected the highest percentage of applicants and falsely accepted the lowest percentage of applicants) was the model built on the manually feature engineered dataset, using gradient boosting as variable selection method and which was fitted using a neural network algorithm.

?

Are the benefits of the best model worth it?

Conclusions

Data: Kaggle competition data - real world data from Home Credit, a company that strives to give loans to people with insufficient credit history

Results:

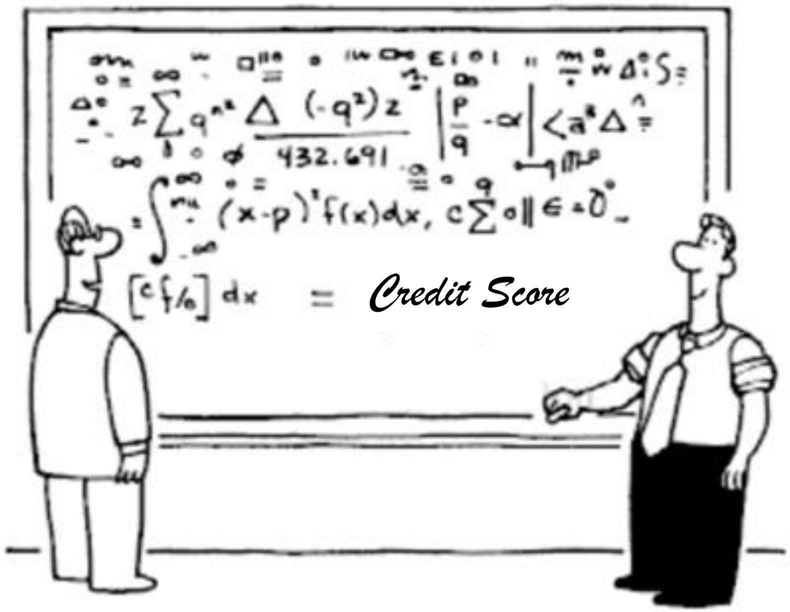
	Modelling fitting method								
	Logistic regression				Neural Network				
Variable selection method:	Stepwise Selection	Decision Tree	Random Forest	Gradient Boosting	Stepwise Selection	Decision Tree	Random Forest	Gradient Boosting	Measure
Raw Data (no feature engineering)	0,6755	0,6602	0,6723	0,6747	0,6785	0,6613	0,6759	**0,6794**	Accuracy
	0,6756	0,6530	0,6723	0,6746	0,6762	0,6516	0,6711	**0,6794**	Precision
	0,6751	0,6839	0,6720	0,6749	0,6848	**0,6933**	0,6900	0,6791	Recall
	0,6754	0,6681	0,6722	0,6747	**0,6805**	0,6718	0,6804	0,6793	F1 Score
Manual Feature Engineering (expert judgement)	0,6900	0,6547	0,6837	0,6856	0,5873	0,6549	**0,6901**	0,6688	Accuracy
	0,6900	0,6565	0,6832	0,6849	0,5886	0,6452	0,6873	0,6793	Precision
	0,6899	0,6491	0,6850	0,6872	0,5801	0,6882	**0,6977**	0,6394	Recall
	0,6900	0,6528	0,6841	0,6861	0,5843	0,6660	**0,6924**	0,6588	F1 Score
Automated Feature Engineering (Python FeatureTools)	**0,6614**	0,6493	0,6611	0,6603	0,6598	0,6498	0,6602	0,6573	Accuracy
	0,6572	0,6390	**0,6580**	0,6572	0,6517	0,6393	0,6515	0,6575	Precision
	0,6747	0,6863	0,6708	0,6700	0,6862	0,6872	**0,6887**	0,6566	Recall
	0,6658	0,6618	0,6644	0,6636	0,6685	0,6624	**0,6696**	0,6571	F1 Score

Learnings

- ***Manual feature engineering*** has at this stage ***not been fully replaced*** by automated feature engineering. ***Human intuition*** in this phase is still a valuable asset for enhancing a scoring model
- ***Machine learning feature selection outperformed*** stepwise selection in almost all the cases, especially when ***combined*** with Neural Network for model fitting
- Machine learning for ***model fitting*** presented ***slight improvements*** for all approaches

Takeaways

1. Machine learning can improve credit decisions	▪ The careful application of machine learning techniques can improve the discriminatory power of credit models ▪ However, the increase in accuracy comes at the cost of lower interpretability of results



It is *obvious* this is a great model!

Agenda

- Introduction and recap
- Machine learning in credit analysis
- **Machine learning in insurance fraud identification**
- Concluding remarks

This section is based on the Masters Dissertation of Jason la Cock and Jonathan Lombard, from the University of Cape Town, supervised by Periklis Thivaios

Question: What are potentially fraudulent activities in insurance?

Insurance fraud refers to the intentional, illegal manipulation of the insurance process with the objective of financial gain. This may include the exaggeration of losses or even the artificial manufacturing of the entire claim

- In the United States, the total cost of (non health) insurance fraud is estimated to be more than $40 billion per year
- Insurance Fraud costs the average U.S. family between $400 and $700 per year in the form of increased premiums

Premium diversion

- The embezzlement of insurance premiums
- An insurance agent fails to send premiums to the underwriter and instead keeps the money for personal use
- Also involves selling insurance without a license, collecting premiums and then not paying claims

Fee churning

- A series of intermediaries take commissions through reinsurance agreements
- The initial premium is reduced by repeated commissions until there is no longer money to pay claims
- The company left to pay the claims is often a business the conspirators have set up to fail

Asset diversion

- The theft of insurance company assets, occurring almost exclusively in the context of an acquisition or merger of an existing insurance company
- Often involves acquiring control of an insurance company with borrowed funds. After making the purchase, the subject uses the assets of the acquired company to pay off the debt

Workers' compensation fraud

- Some entities purport to provide workers' compensation insurance at a reduced cost and then misappropriate premium funds without ever providing insurance

Source: FBI (https://www.fbi.gov/stats-services/publications/insurance-fraud)

Applying machine learning to insurance fraud analysis (automobile insurance)

We compared machine learning approaches to traditional logistic regression to evaluate strengths and weaknesses

1. Data exploration
2. Modelling
3. Machine learning environment and algorithms
4. Performance metrics
5. Observations and results

1. Data exploration and minimisation of misclassification cost

- The data corresponds to a sample of 11565 automobile insurance claims recorded between 1994 and 1996
- In order to address the issues pertaining to imbalanced data learning (poor classification accuracy of the minority class, credibility of classification accuracy, and over-fitting on the training data), we implemented a cost-sensitive method, determining the misclassification costs for false positives and false negatives.
- We optimised classification performance by retraining classifiers on feedback received on false positives, therefore minimising the misclassification cost on the training data set
- Thereafter the weightings assigned to both classes were optimised to minimise the misclassification cost on the training data set
- This feedback was especially beneficial for classifiers such as gradient boosting classifiers and SVMs, which assign greater weight to difficult-to-classify observations

Variable	**Category**	**Variable**	**Category**
Accident Area	Rural	Month Claimed	August
Address Change	1 change	No. of Cars	2 and 3
Age of PH	Younger people	No Supplements	None
Age of Vehicle	0 years	Past No Claims	0 claims
Base Policy	All Perils	Police Report	No report
Claim Size	Larger Claims	Sex	Male
Day of Week	Sunday	Vehicle	Sedans
Day Claimed	Sunday	Vehicle Price	Most Expensive
Deductible	0	Week of Month	No Relation
Driver Rating	Best Drivers	Week Claimed	1st Week
Fault	Policyholder	Witness Present	No Witness
Marital Status	Widowed	Year	1994

Negligible ☐ Less Significant ☐ Significant ☐

2. We employed a number of different approaches to modelling fraud and compared the results

Logistic regression

- A regression technique used to regress a binary response variable against a set of explanatory variables
- Most insurance fraud data sets expressing fraud as a binary response vari-able

Random forest

- An ensemble of decision trees constructed to reduce the variances of estimates and eliminate the likelihood of over-fitting
- The non-parametric and simple nature of the underlying decision trees in RFs are beneficial to fraud detection

Gradient boosting

- A generalisation of boosting trees constructed sequentially from the residuals of trees grown previously
- Boosting trees combine several decision trees with poor estimation capabilities to create a boosted tree that can accurately predict the likelihood of a claim being fraudulent

Neural networks

- A multi-layer network constructed by determining the weights of the hidden nodes connecting the network
- Neural networks require long training times; however once trained, neural networks can produce rapid evaluations of the target output function

Support Vector Machines

- Used to classify an observation based on its characteristics
- Vectors situated close to the hyperplane are known as support vectors and are most significant in determining the position of the hyperplane

Naïve Bayes

- A classification methodology that makes use of Bayes theorem
- To estimate the conditional probability, Naive Bayes assumes that the set of predefined explanatory variables X1, ..., Xn are conditionally independent

3. Machine learning environment and algorithms

- The ML algorithms were implemented using the Scikit-learn open-source libraries for machine learning in the Python programming language (Pedregosa et al. 2011). Development was performed using Jupyter Python 3 note- books which are open-source interactive programming environments supporting code and markdown text

- The following generalised algorithm was applied for each of the identified ML techniques:
  1. The hyperparameters of each classifier were optimised using a grid search algorithm. A range of possible values was specified for each hyperparameter and the optimal set of hyperparameters was identified as the set which maximised the balanced accuracy score
  2. Stratified 10-fold cross validation was applied during hyperparameter tuning. This reduced the likelihood of overfitting the training data, hence improving the robustness of the trained model
  3. An estimate of the variance and mean of the classifier was obtained through a 100-fold cross validation of the trained classifier. Specifically, the variance and mean of the balanced accuracy score was identified
  4. The identified performance metrics were calculated for each ML technique for classification performance comparisons
  5. For each of the tree and boosting classifiers, a bar graph was constructed comparing the relative importance of the different features in the training data set
  6. An aggregate PRC and ROC plot comparing the classification performance of all classifiers was constructed

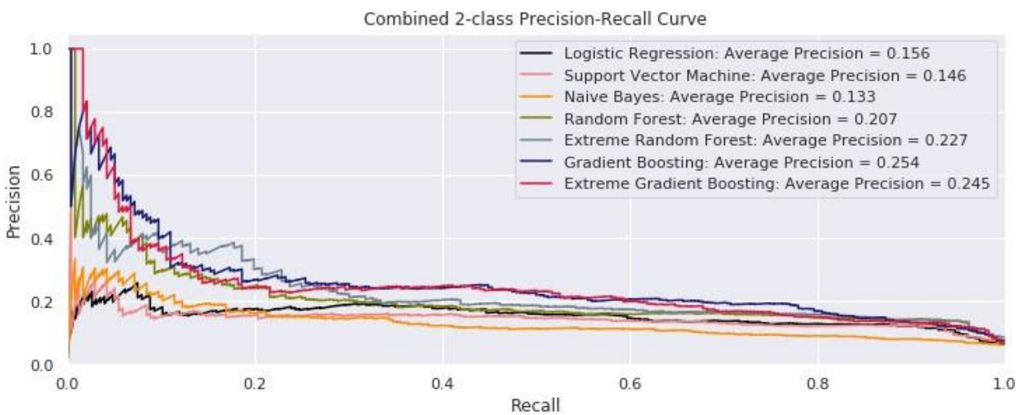
4. Performance metrics

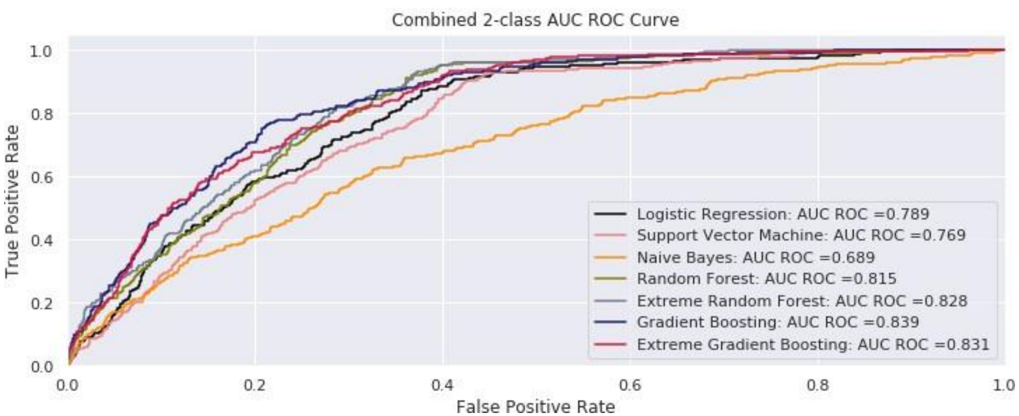
- Since no single metric can definitively detect the best ML classifier for identifying fraud (Wainer and Franceschinell 2018), we employed the following performance metrics for comparing machine learning classifiers:
  1. Matthews correlation coefficient (MCC): MCC was selected for its proven usefulness for imbalanced data. Luque et al. show that MCC is the best performance metric to be used on imbalanced data sets when both successes and errors are important, as they are for fraud identification
  2. Recall and balanced accuracy score (BAS): BAS is a performance metric calculated as the weighted average of recall per class. BAS is applicable to fraud identification as it avoids inflating accuracy achieved when dealing with imbalanced data. Recall is important as it indicates how efficiently the classifier correctly identifies claims which are truly fraudulent.
  3. Graphical performance metrics were used to visualise classification , namely:
     - Precision recall curves (PRC)
     - Receiving operating curves (ROC)

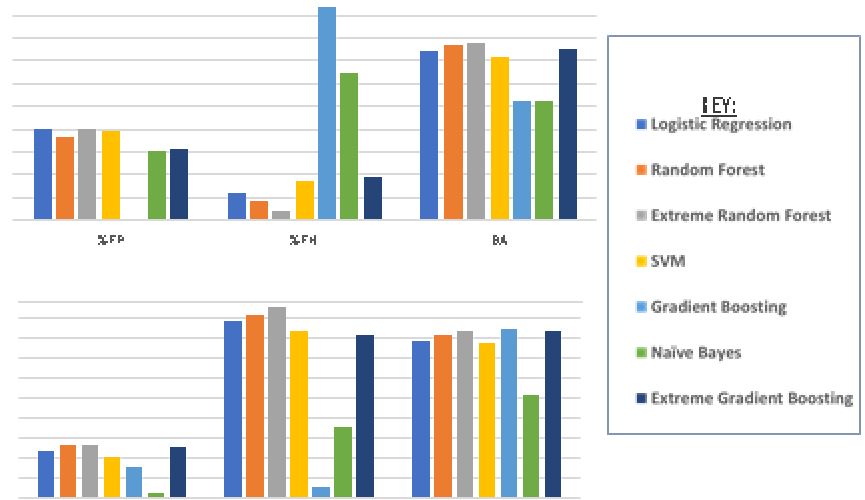
MCC	$\frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$
Recall	$\frac{TP}{TP+FN}$
AUC ROC	Area under the ROC curve.
Normalised False Positives	$\frac{FP}{FP+TP}$
Normalised False Negatives	$\frac{FN}{FN+TN}$
BAS	$\frac{TPR+TNR}{2}$

5. Observations and results

Extreme random forest and gradient boosting classifiers achieved the best performance in fraud identification, but processing was slow







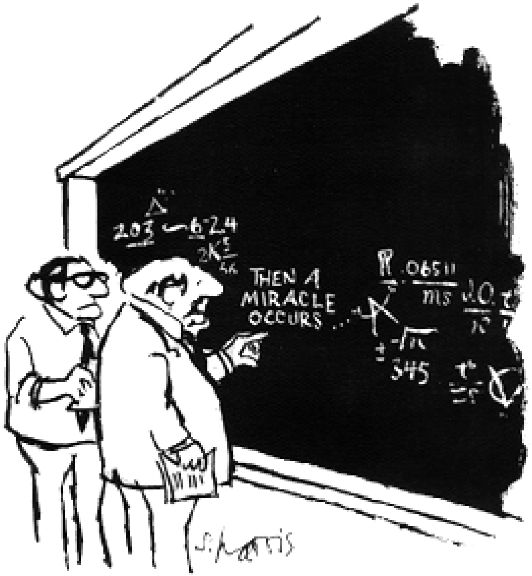
Classifier	MCC	Recall	AUC ROC	% FP	% FN	BAS
Logistic Regression	0,233	0,885	0,787	0,400	0,120	0,741
Random Forest	0,268	0,911	0,815	0,366	0,089	0,773
Extreme Random Forest	0,268	0,953	0,828	0,402	0,047	0,776
SVM	0,212	0,826	0,769	0,393	0,174	0,717
Gradient Boosting	0,160	0,059	0,839	0,004	0,941	0,528
Naïve Bayes	0,026	0,352	0,513	0,302	0,648	0,525
Extreme Gradient Boosting	0,254	0,814	0,831	0,314	0,186	0,750

5. Observations and results

Logistic regression	Random forest	Extreme random forest	Gradient boosting	Extreme gradient boosting	Support Vector Machines
▪ Logistic regression was found to exhibit faster model running times ▪ Practically this is advantageous for claim processors wishing to classify claims in real time	▪ The RF classifier achieved excellent performance in identifying actual fraudulent claims and therefore minimising false negatives, identifying 92.1% of fraudulent claims	▪ XRFs increase randomisation by randomly selecting nodes in tree construction where trees are split, rather than splitting at the best location as in standard RFs ▪ XRF classifiers successfully reduced variance	▪ The GB classifier did not adapt well to the imbalanced nature of data, only successfully classifying 14.41% of fraudulent claims ▪ Further, the GB classifier was extremely time-intensive	▪ Extreme GB classifiers were employed to tackle the imbalanced nature of the data set ▪ XGB classifier identified 81% of fraudulent claims	▪ The SVM classifier was found to be of particular significance in learning from new claim data ▪ It ran faster than other ML techniques such as decision trees and gradient boosting

Takeaways

2. Machine learning can improve insurance fraud identification	▪ Various machine learning techniques offer upsides and downsides in car insurance fraud identification ▪ No single model is universally better and careful consideration is required in their application



"I think you should be more explicit here in step two."

Agenda

- Introduction and recap
- Machine learning in credit analysis
- Machine learning in insurance fraud identification
- **Concluding remarks**

Summary of takeaways

1. Machine learning can improve credit decisions	▪ The careful application of machine learning techniques can improve the discriminatory power of credit models ▪ However, the increase in accuracy comes at the cost of lower interpretability of results
2. Machine learning can improve insurance fraud identification	▪ Various machine learning techniques offer upsides and downsides in car insurance fraud identification ▪ No single model is universally better and careful consideration is required in their application
3. Machine learning is a great, but risky tool	▪ The benefits of machine learning techniques are undeniable, but they don't come without downsides

Any questions?

UNIVERSITY *of* NICOSIA

Questions?

Contact Us:

Twitter: **@mscdigital**
Course Support: **digitalcurrency@unic.ac.cy**
IT & Live Session Support: **dl.it@unic.ac.cy**