Using Predictive Analytics to Design Investment Strategies for Fintech Lending

Constantin Romanescu

(500762092)

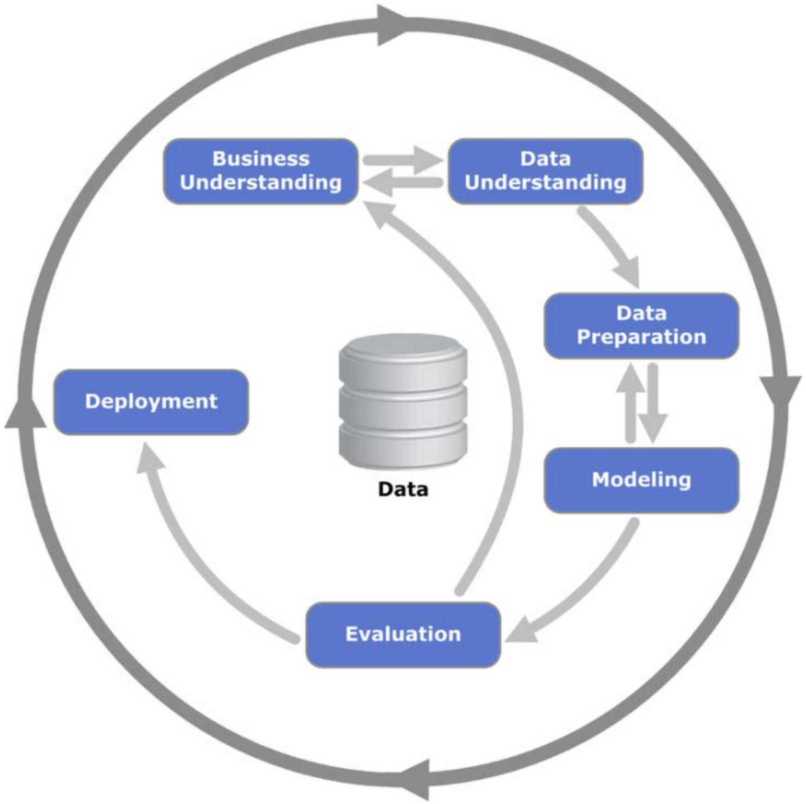
https://github.com/cromanes/Ryerson-Capstone

• founded in 2007, in San Francisco;
• the largest Peer-to-Peer (P2P) lending institution;
• first P2P IPO in December 2014;
• more than 41 billion USD in loan issuance (09/2018);
• makes its loan data publicly available.

Data analytics pipeline

Cross-Industry Standard Process for Data Mining (CRISP – DM)*

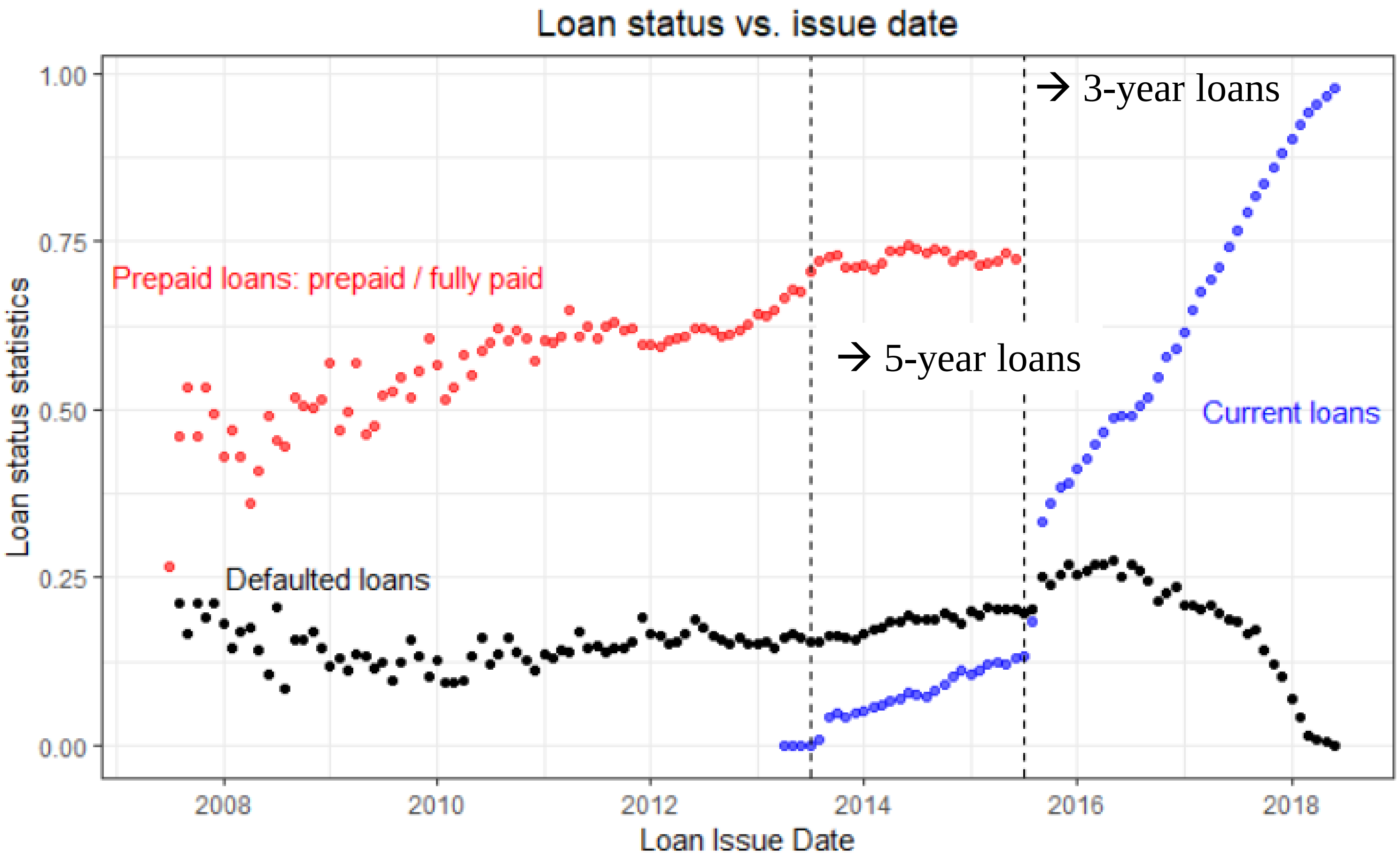
(Chapman et al. (2000). CRISP-DM 1.0. Step-by-step data mining guide. Chicago, Il.)



Tools:

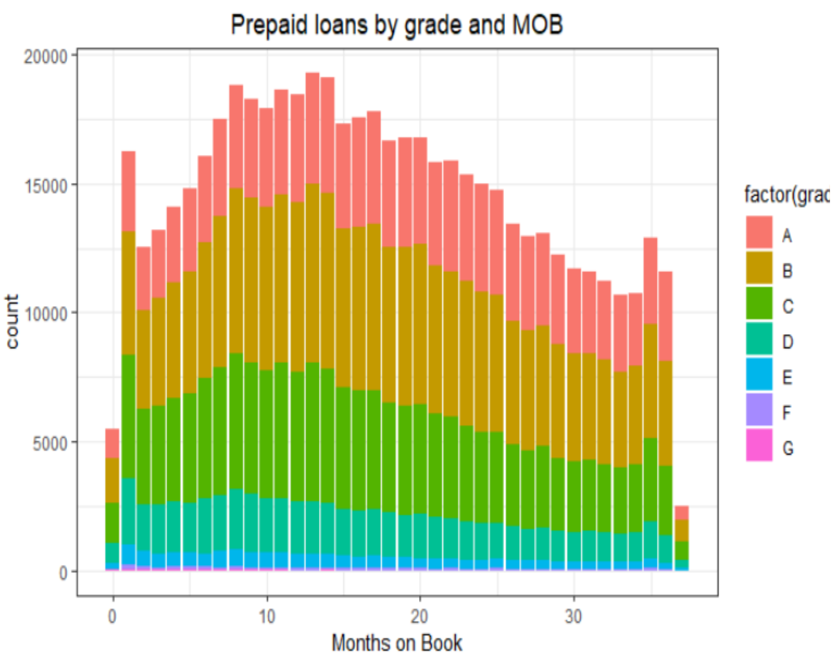
- R, SQL;
- Cloud computing (Azure ML, h2o.ai)

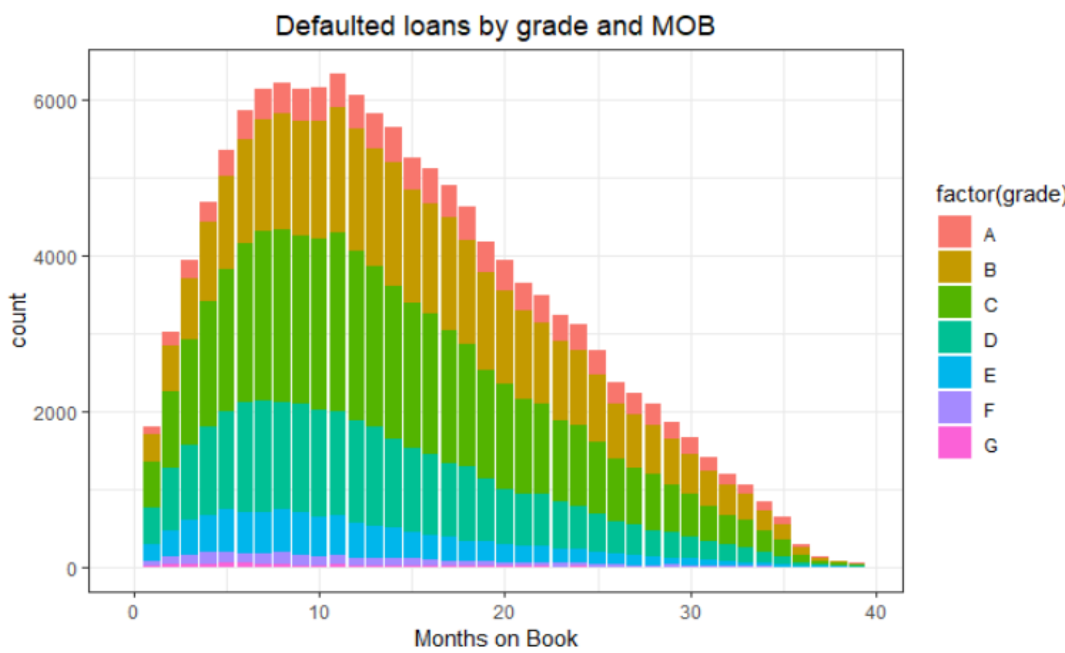
Gaining Insights from Data



- Loan defaults and prepayment lead to a decrease in investor earnings.
- There is a slight increase in the prepayment rate.

Gaining Insights from Data





- Both loan prepayments and loan defaults peak at approximately 12 months;
- Borrowers with high grade loans (low interest rate) are more likely to prepay;
- Borrowers with lower grade loans are more likely to default on their loans.

Data preparation

Impute missing values

Feature Selection based on correlation between independent features (Pearson, χ^2 tests)

Feature Selection based on correlation between loan_status and the independent features (Wilcox, χ^2 tests)

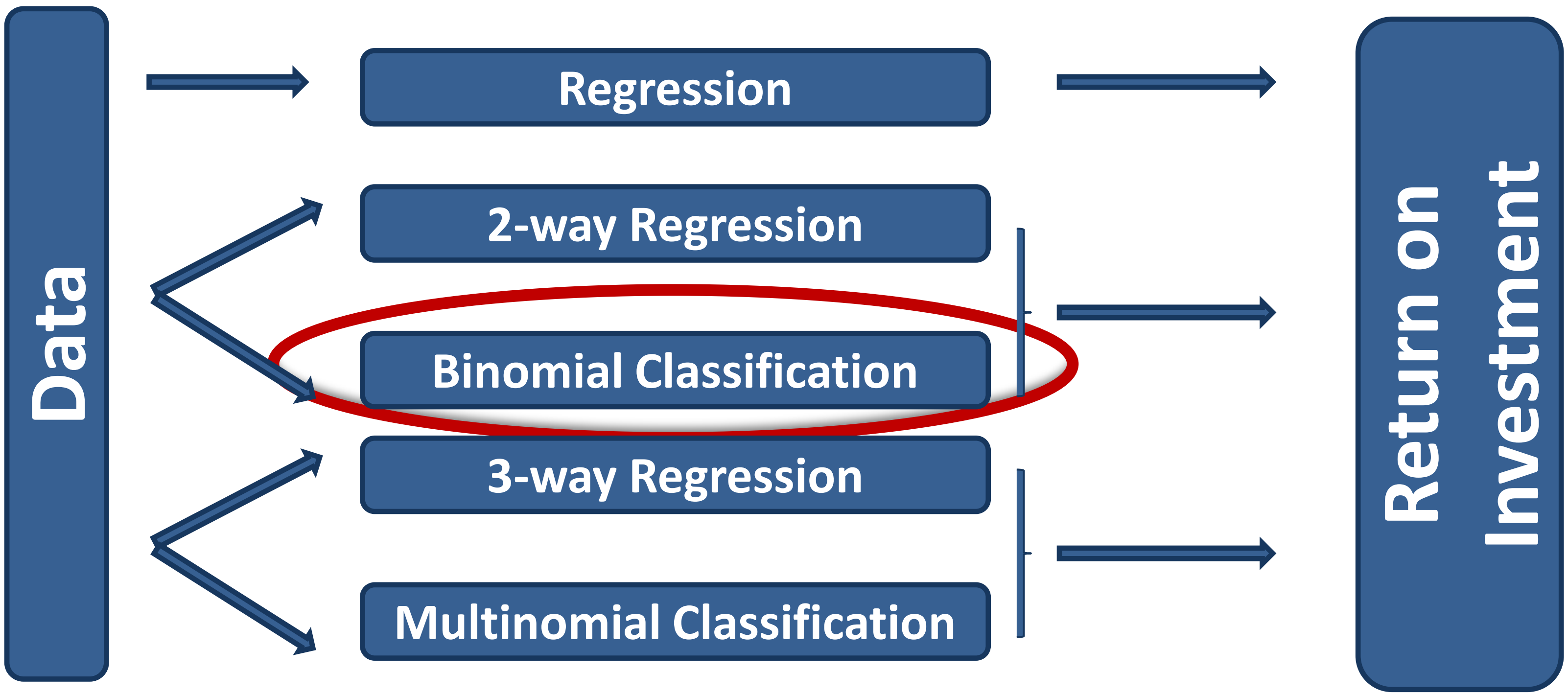
Data transformation of the numerical variables using basic transformations (log, sqrt, 1/sqrt) to reduce skew

Data transformation of categorical features to reduce the imbalance between levels

Feature Engineering

Apply all transformations to the validation, test and new datasets

Modelling Approach



$$ROI = \frac{total\_payment - loan\_amount}{loan\_amount} \times \frac{12}{36}$$

$$ROI_i = P_i(loan_{status} = charged\_off) * roi_i(loan_{status} = charged\_off) + P_i(loan_{status} = fully\_paid) * roi_i(loan_{status} = fully\_paid)$$

Binomial Classification - Approach

Choose an algorithm

↓

Run feature selection

↓

Tune algorithm parameters using the train and validation sets

↓

Measure performance metrics of the test and new datasets

↓

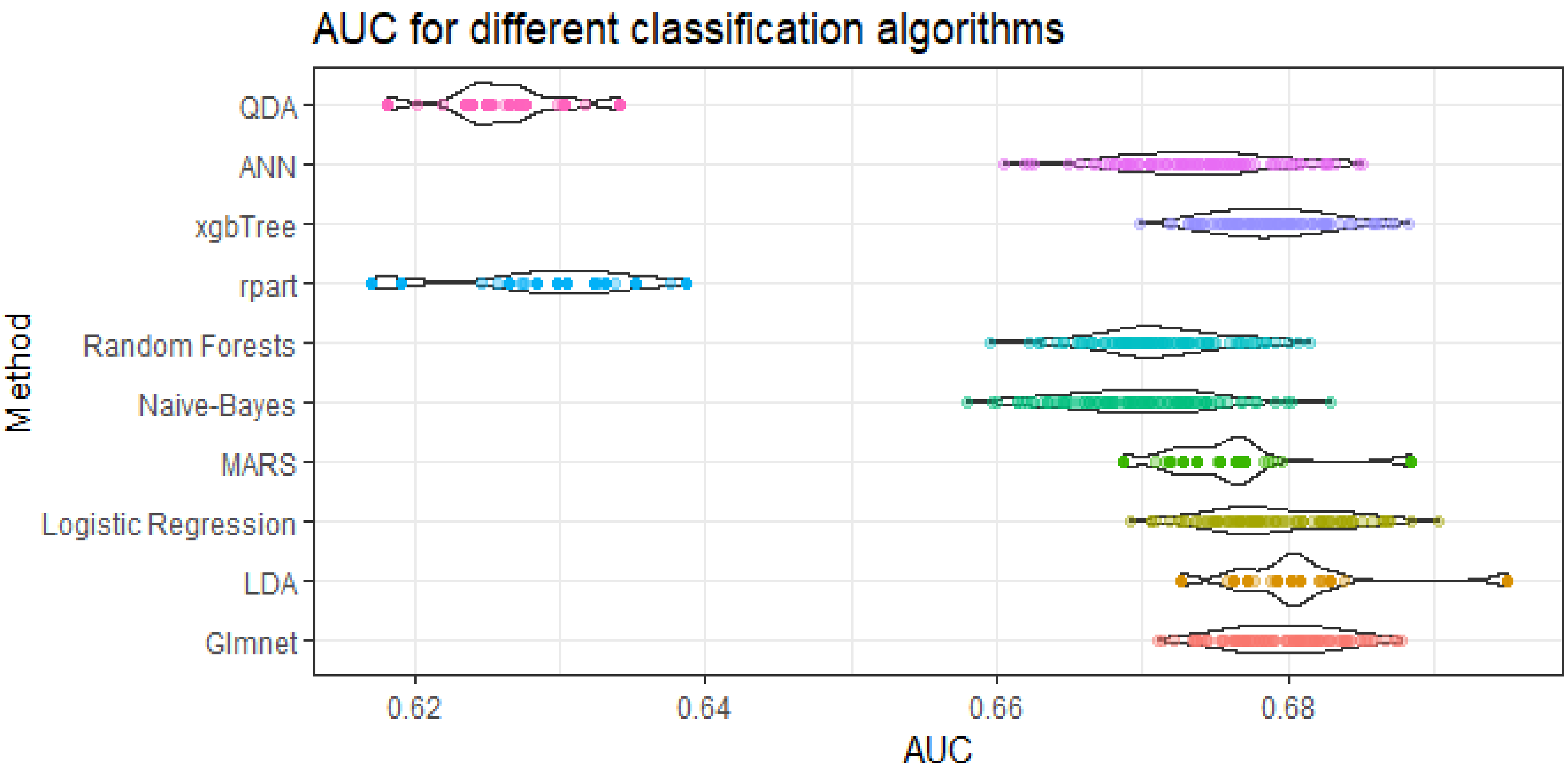
Measure the variance of performance metrics using repeated cross-validation (10 x 10)

Binomial Classification - Results

	Naive Bayes		Logistic Regression		RF		ANN		rpart		MARS		xgbtree		LDA	
	Test	New	Test	New	Test	New	Test	New	Test	New	Test	New	Test	New	Test	New
Accuracy	0.82	0.82	0.85	0.84	0.85	0.84	0.85	0.85	0.85	0.84	0.85	0.84	0.85	0.85	0.85	0.84
Sensitivity	0.15	0.14	0.00	0.00	0.00	0.00	0	0	0.02	0.03	0.00	0.00	0	0	0.01	0.02
Specificity	0.93	0.94	1	1	1	1	1	1	0.99	0.99	1	1	1	1	1	1
Precision	0.29	0.33	0.48	0.51	0.44	0.42	NA	NA	0.33	0.35	0.39	0.48	NA	NA	0.43	0.47
F1	0.2	0.20	0.00	0.02	0.01	0.01	NA	NA	0.04	0.06	0.00	0.01	NA	NA	0.02	0.04
AUC	0.66	0.69	0.67	0.70	0.66	0.68	0.66	0.7	0.62	0.62	0.67	0.69	0.67	0.70	0.68	0.7
Brier score	0.13	0.13	0.12	0.12	0.12	0.12	0.12	0.12	0.12	0.13	0.12	0.12	0.12	0.12	0.12	0.12

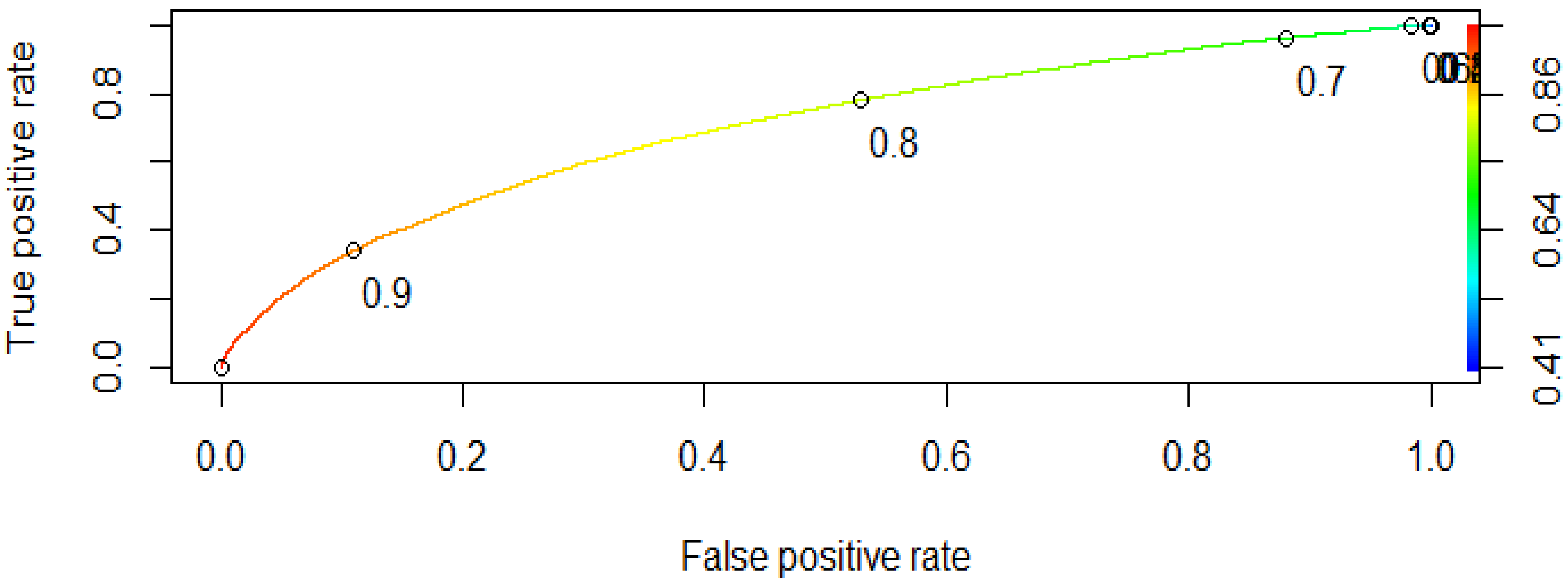
- Default is the positive class (except for the AUC calculations);
- Most algorithms fail to identify a single default case;
- AUC and Brier coefficient calculations indicate a rather good separation between default and non-default cases;
- Needs to look more closely at the ROC.

Binomial Classification - Results



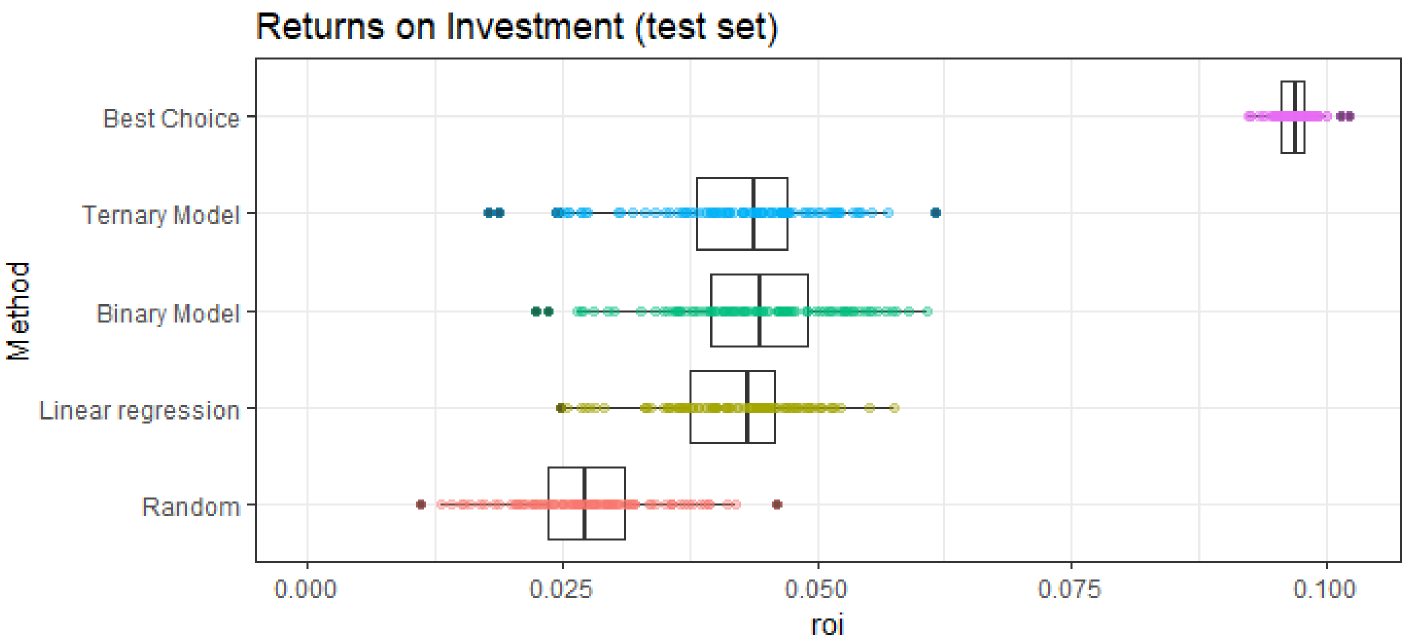
One-way Anova → TuckeyHSD → Occam's razor → Logistic Regression

Dealing with unbalanced data



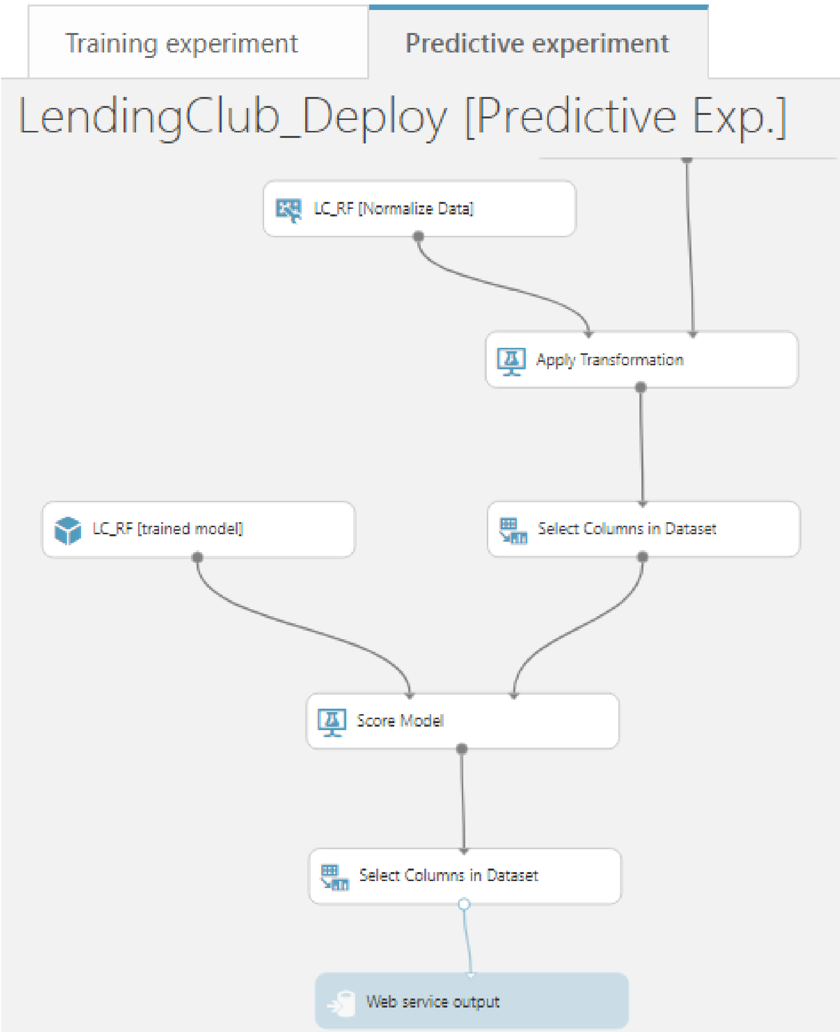
Threshold	True negatives	False negatives	Accuracy	Penalty
0.5	8	4	0.853	-
0.6	98	132	0.853	1.42
0.7	889	1760	0.842	1.98
0.8	4597	13149	0.745	2.86
0.9	9996	44650	0.416	4.47
1.0	11635	67678	0	Inf

Naïve Model Evaluation



- All three models outperform the random model by ~ 60 %;
- There are no statistically significant differences between the proposed models;
- There is plenty of room for model improvement.

Model Deployment



PARAMETERS				PREDICTED VALUES
_rate	grade	sub_grade	purpose	ScoredLabels
16.99	D	D3	debt_consolidation	0.0058
14.65	C	C5	credit_card	0.0079
12.29	C	C1	credit_card	0.0168
13.99	C	C4	debt_consolidation	0.0151
15.61	D	D1	other	0.0261
16.99	D	D3	credit_card	0.0043
13.99	C	C4	other	0.0242
13.99	C	C4	debt_consolidation	0.0155
17.86	D	D5	moving	0.0107
19.99	EFG	E4	debt_consolidation	0.0117

Conclusions

- used statistical and graphical methods using R and Azure ML to gain insights on the loans offered by Lending Club;
- used a wide range of machine learning algorithms to identify loans likely to default;
- designed and tested investment models that yield better annualized returns on investment;
- classification results are consistent with results from peer-reviewed publications;
- created a deployable model using Azure ML.

Borrowers

2M+
Mainstream Consumers[1]

600-850
FICO Range[2]

$62K
Median Income[2]

Investors

180K+
Self-directed Individuals

60+
Managed Accounts

40+
Banks

70+
Institutions

LendingClub

1) Across all products. 2) Personal loan borrowers in Q3 2017. FICO range reflects LendingClub's personal loans credit policy.