Credit Card Approval With Machine Learning

A Model You Can Take to the Bank

By The Data Dudes: Orr Shalev, Antoine Nadaud,
Noam Kleinman, and Jacob Salomon

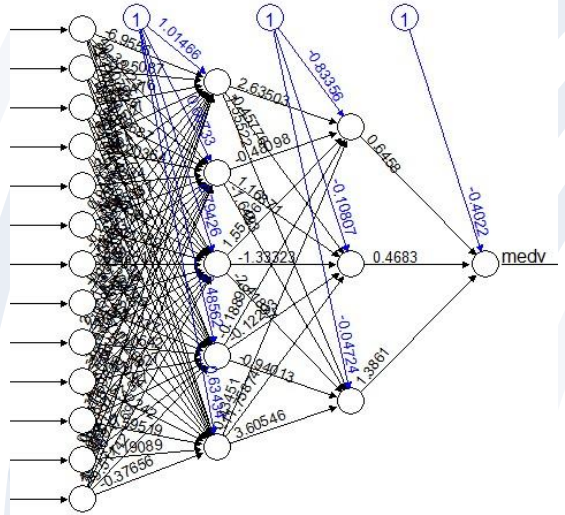
First Steps

- Is ML appropriate?
- Domain knowledge
  - Application process
- Model selection
  - Predictive, transparent

Simple Regression

$$y = \alpha + \beta x_i + \varepsilon_i$$

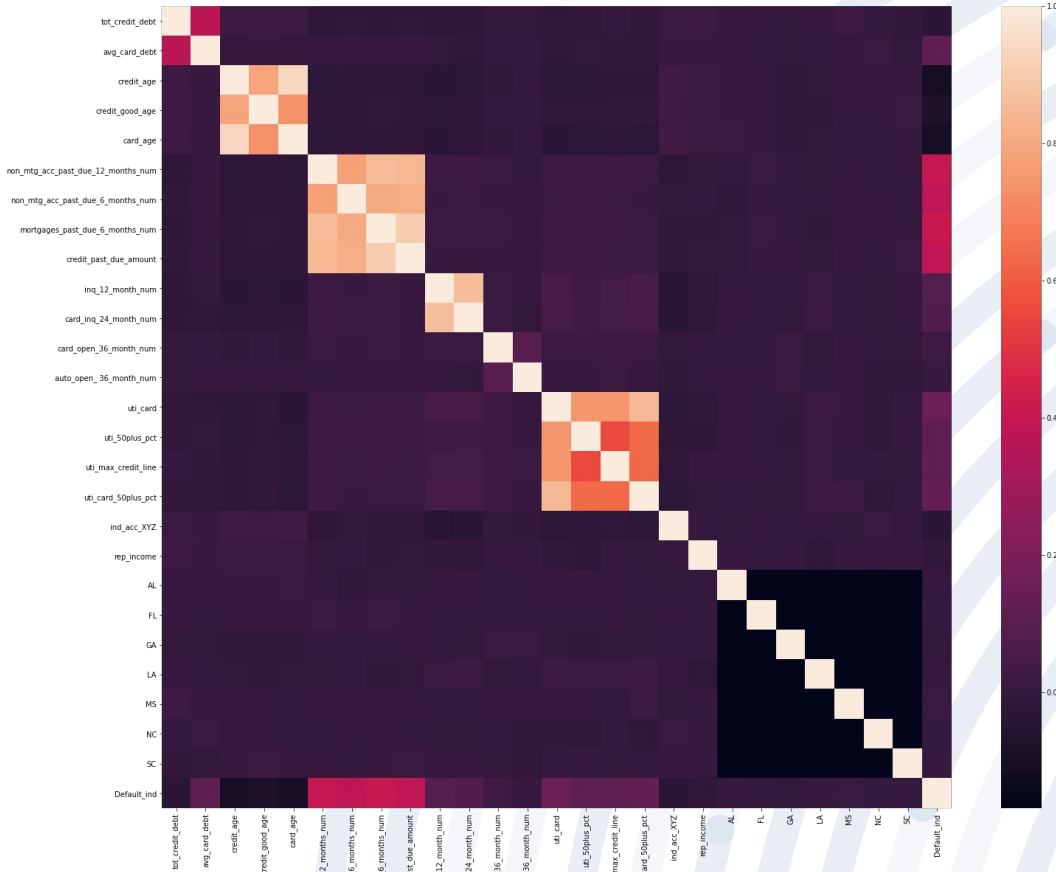
Response Explanatory Noise



Exploratory Data Analysis

- Problems with reported income
- Disparities between defaulting and non-defaulting accounts
  - Defaulting Variance
- Collinearity and multicollinearity
  - Variance inflation factor
  - Correlation matrix

Correlation Heatmap

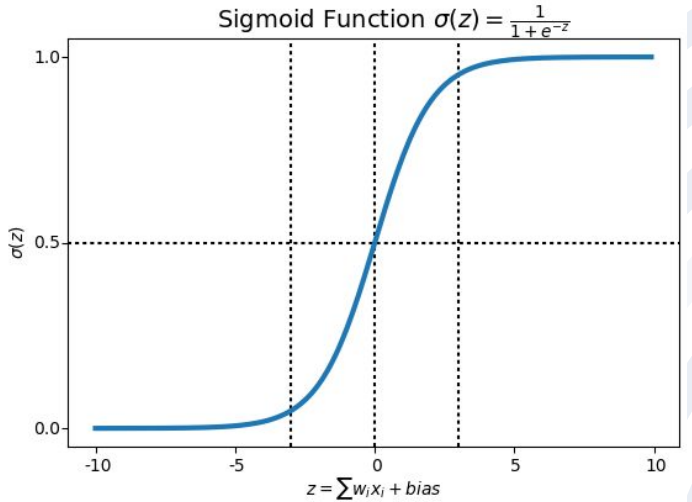


Logistic Regression

- Removed variables with high collinearity
- Dummy variables for states
- Mean utilization
- Past due months num addition
- Normalization

$$n = \frac{n - min}{max - min} * 2$$

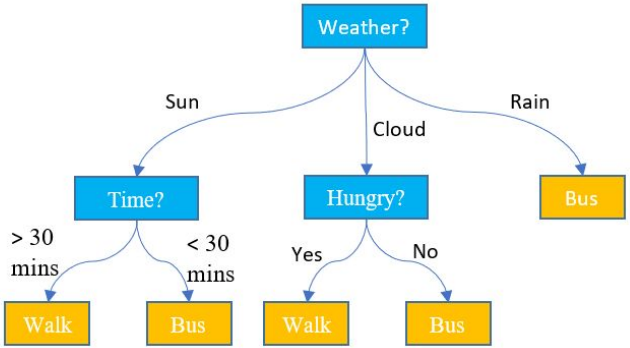
Sigmoid Function Basis



Random Forests

- Removed variables with low response linearity
- Dummy variables for states
- Imputation of missing values

Decision Tree Basis



Model Building and Results

Logistic Regression

- Hyperparameter tuning:
  - Liblinear solver
  - Penalty of L2
  - C value: 0.01
- Accuracy: 93.58%

Random Forests

- Hyperparameter tuning:
  - 100 trees
  - Max variables when splitting: Log_2
- Accuracy: 93.66%

Error Comparison

Logistic Regression

Actual Values

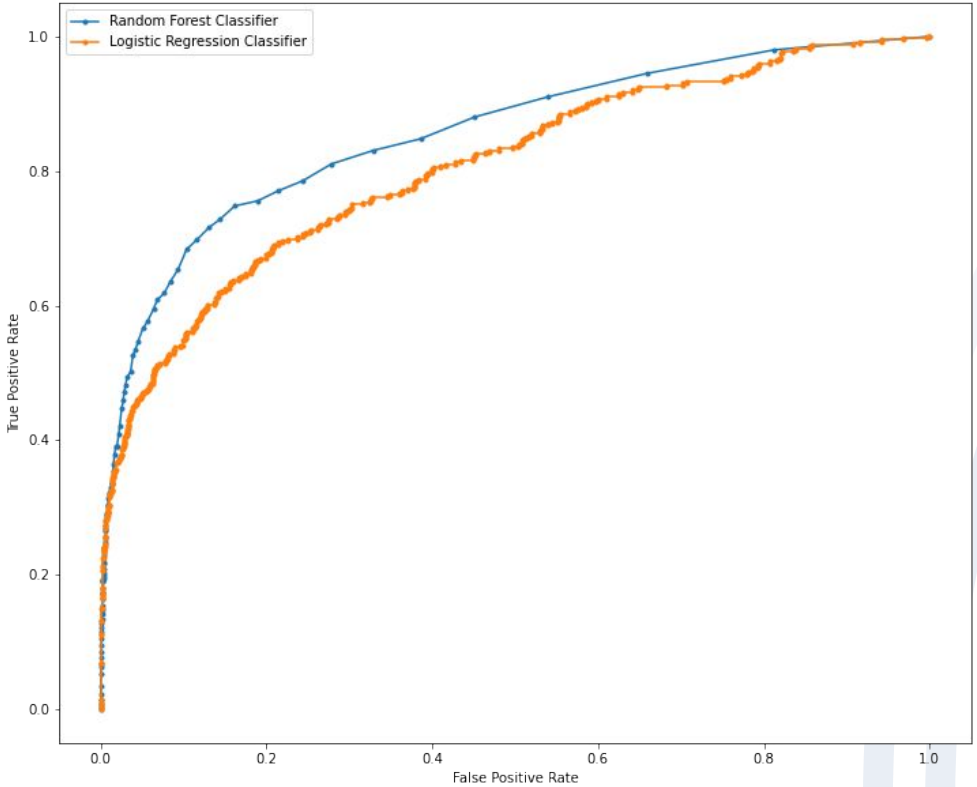
Predicted Values	Non-default (0)	Default (1)
Non-default (0)	TP 4552 0.9104	FP 47 0.0094
Default (1)	FN 274 0.0548	TN 127 0.0254

Random Forest

Actual Values

Predicted Values	Non-default (0)	Default (1)
Non-default (0)	TP 4568 0.9136	FP 31 0.0062
Default (1)	FN 286 0.0572	TN 115 0.0230

Plotting Receiver Operating Characteristic (ROC) Curve



Area Under Curve:

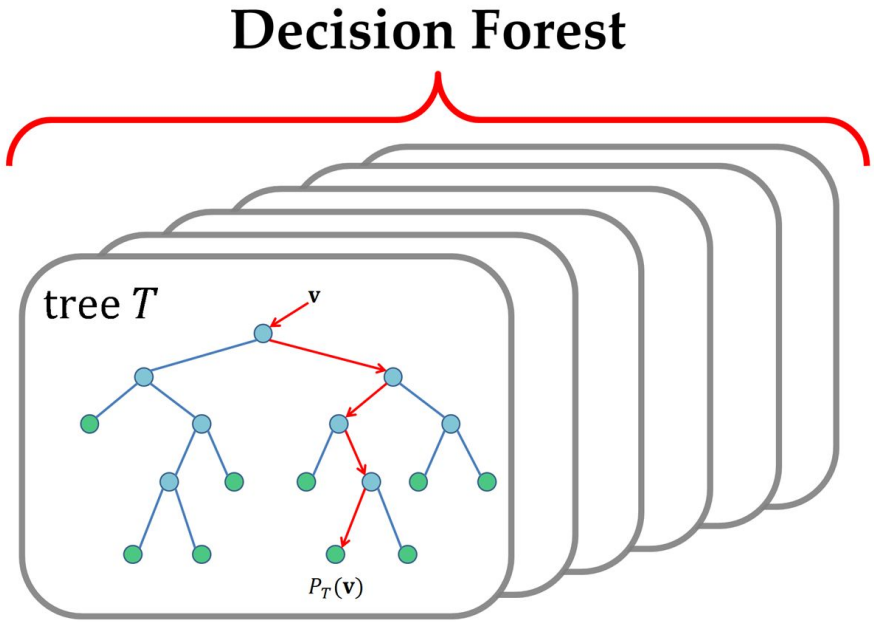
- RF Model: 0.852
- LR Model: 0.806

Model Comparison

Logistic Regression	Random Forest
Deals well with high variance in explanatory and noise variables (reference 8)	Better false positive rate when importance of noise variables is large (reference 8)
Higher true negative and lower false positive rate	Higher true positive and lower false negative rate
High interpretability	Medium interpretability
Requires greater amount of data preprocessing	Requires less data preprocessing

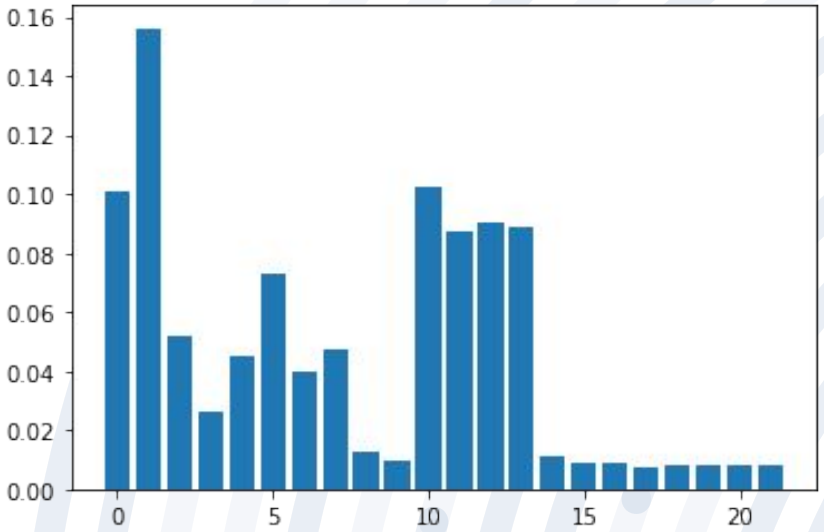
Chosen Model: RF

- Noise
- Maximizing value
  - Loss reduction
- How does it work?
  - Aggregate of decision trees



Feature importance

- High: debt, utilization and past due amount
- Low: prior XYZ customer, state, accounts opened in last 36 months



Model implementation

- Decision making: complete or hybrid
- XYZ Prior customers
  - Brand loyalty vs. Importance
- Explaining rejections
  - Transparency vs. Security

Thank you,
Are there any questions?