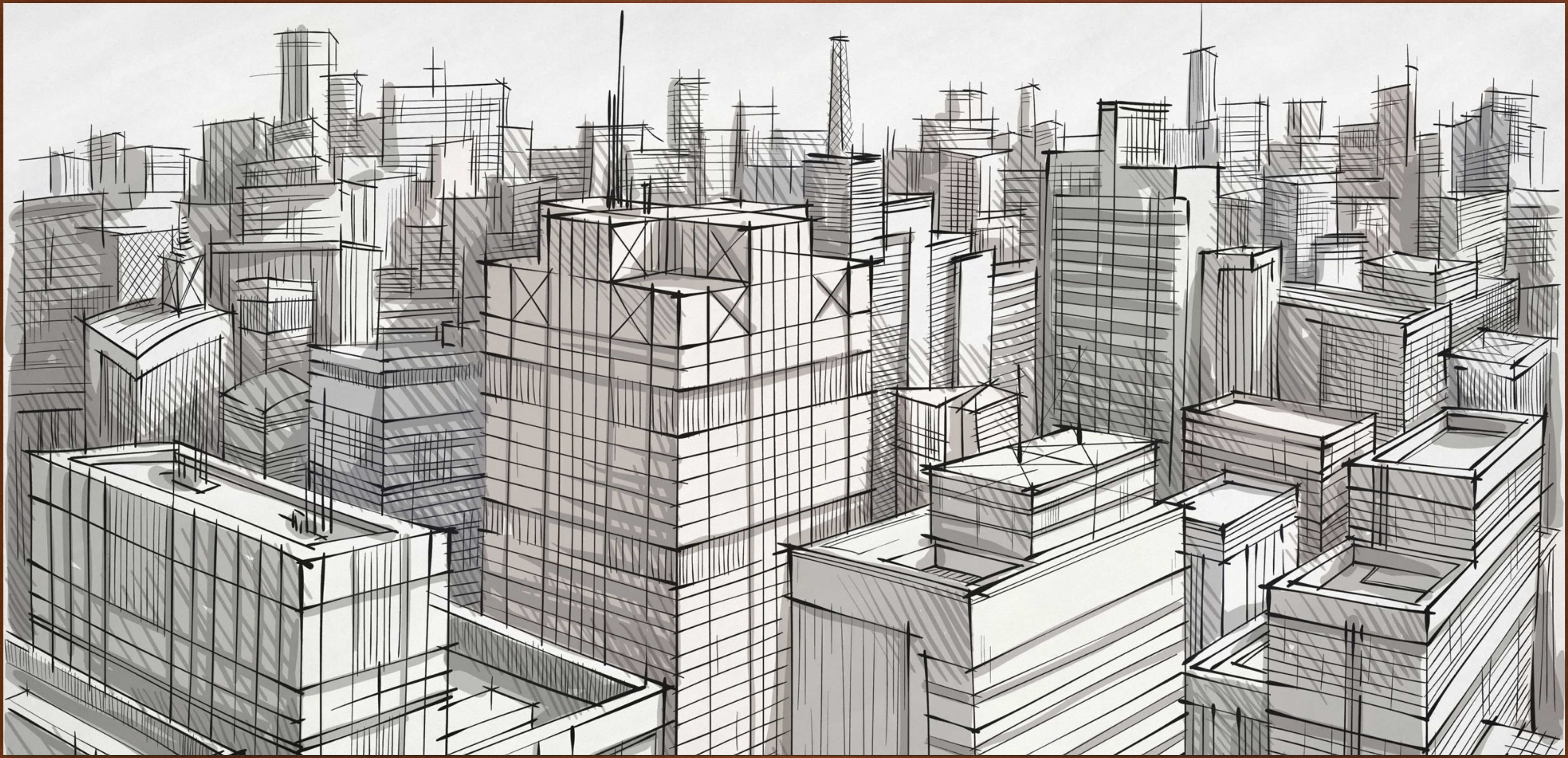
CAPSTONE PROJECT:

GIVE ME SOME CREDIT

FOUNDATIONS OF DATA SCIENCE WORKSHOP, AUTUMN 2015

DAVID CROOK



GIVE ME SOME CREDIT

KAGGLE.COM

- Credit Scoring modeling competition, originally conducted Autumn 2011. Entries from 925 teams.
- Data contains over 250,000 borrower observations (~100,000 of them for scoring competitor entry), with variables like `age`, `DebtRatio` and `NumberRealEstateLoansOrLines`
- A well-defined ***binary classification*** problem: Predict a probability on a binary dependent variable.
    - Specifically, whether or not a given borrower will experience "serious financial distress" within two years. 6.6% of 150,000 in training dataset did.

THE APPROACH

ORDER OF OPERATIONS

- Define problem (was defined by competition) and Obtain data
- Explore and clean data*
- Build models
- Evaluate + validate models
- Iterate

Refer to paper and code for full details.

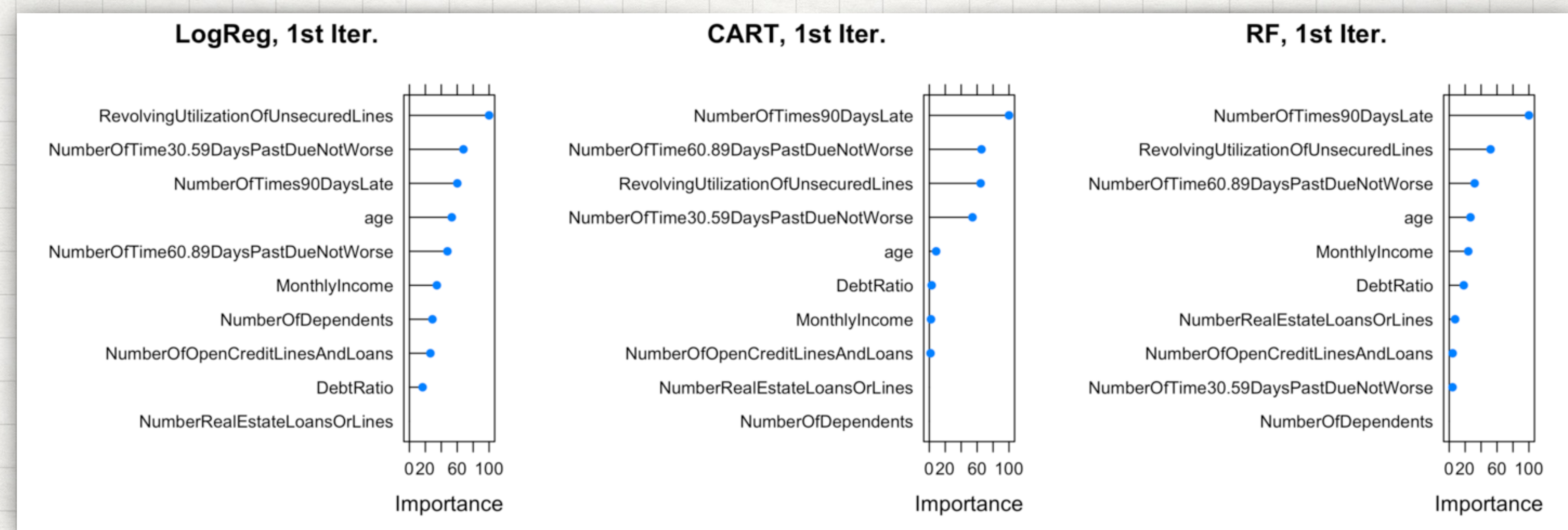
** Some Feature engineering was done in this step*

THE APPROACH - BUILD MODELS

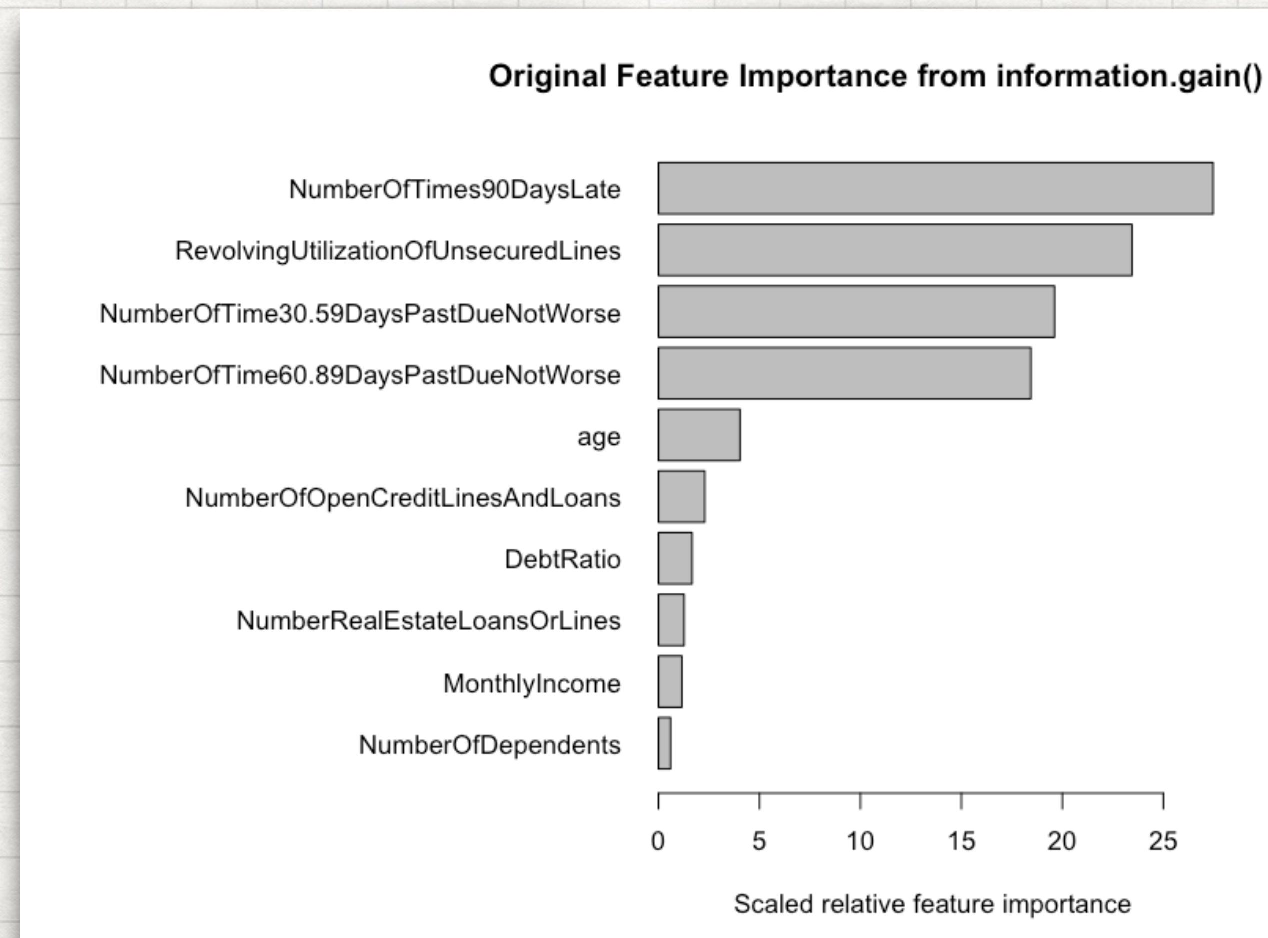
BUILD ML MODELS TRAINED WITH BORROWER DATA

- Three model types built: *Logistic Regression, Classification Tree,* and *Random Forest*
- Classification Trees can be more understandable than other types
- Logistic Regressions are binary classifiers
- Random Forest is an ensemble method using classification trees.
  - RF algorithm reduces chance of over-fitting compared to normal classification tree.

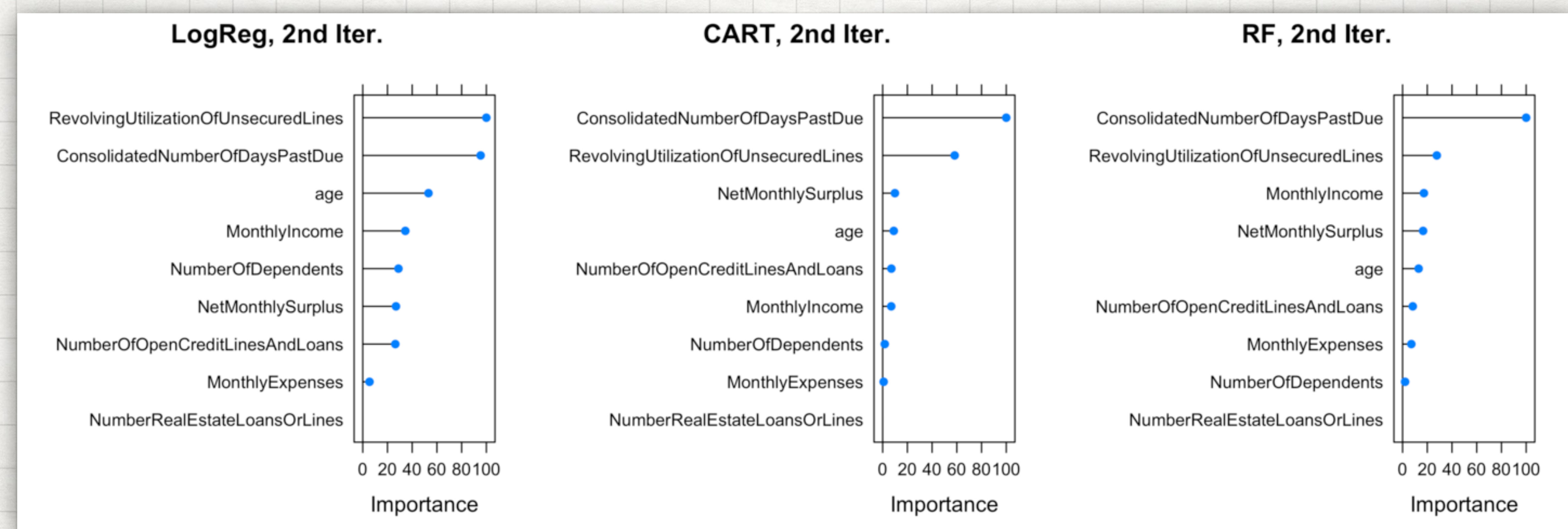
VARIABLE IMPORTANCE - OBS IN MODEL VS. PREDICTED



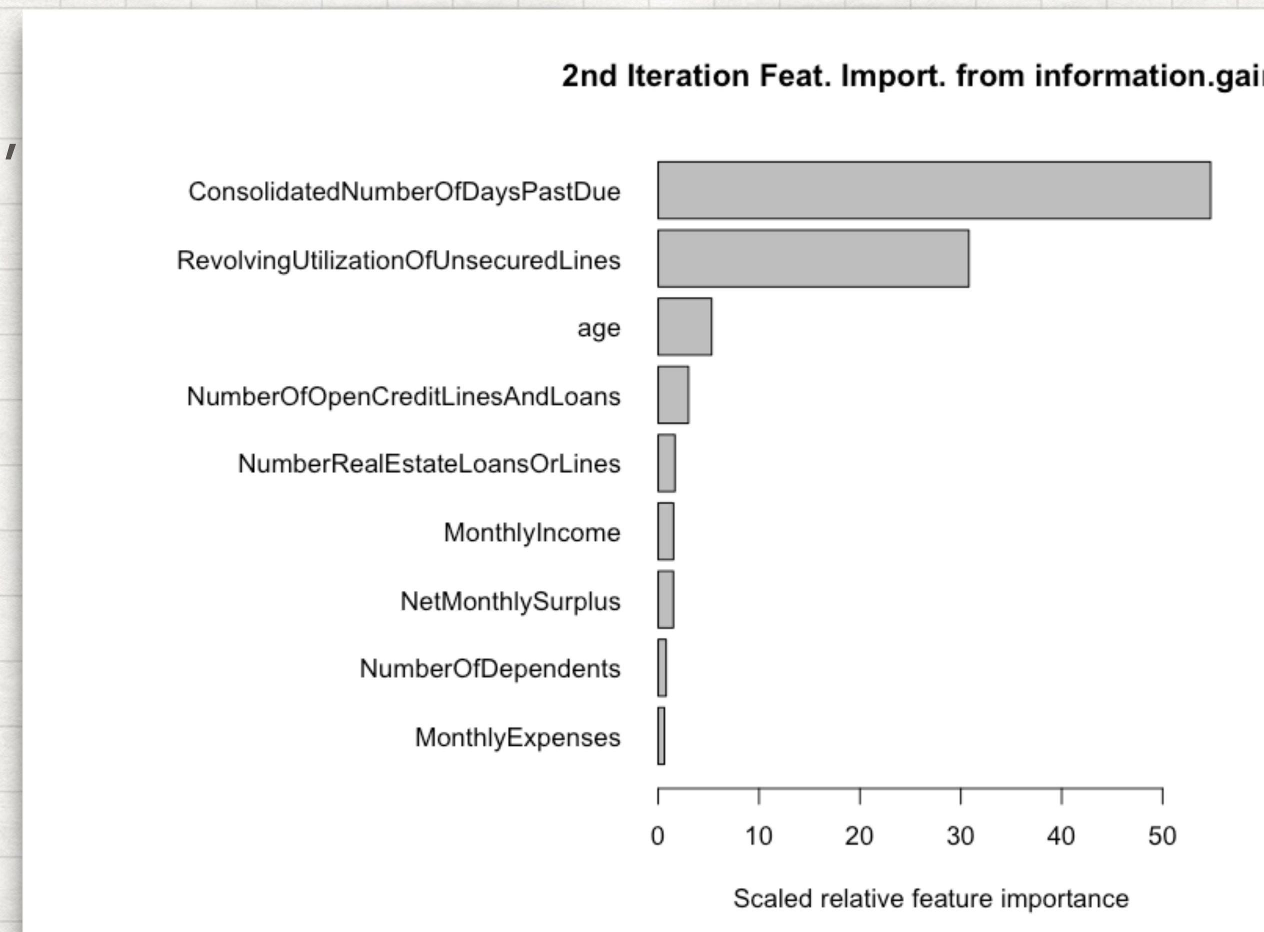
In First Iteration, built models with original variables.



VARIABLE IMPORTANCE - OBS IN MODEL VS. PREDICTED



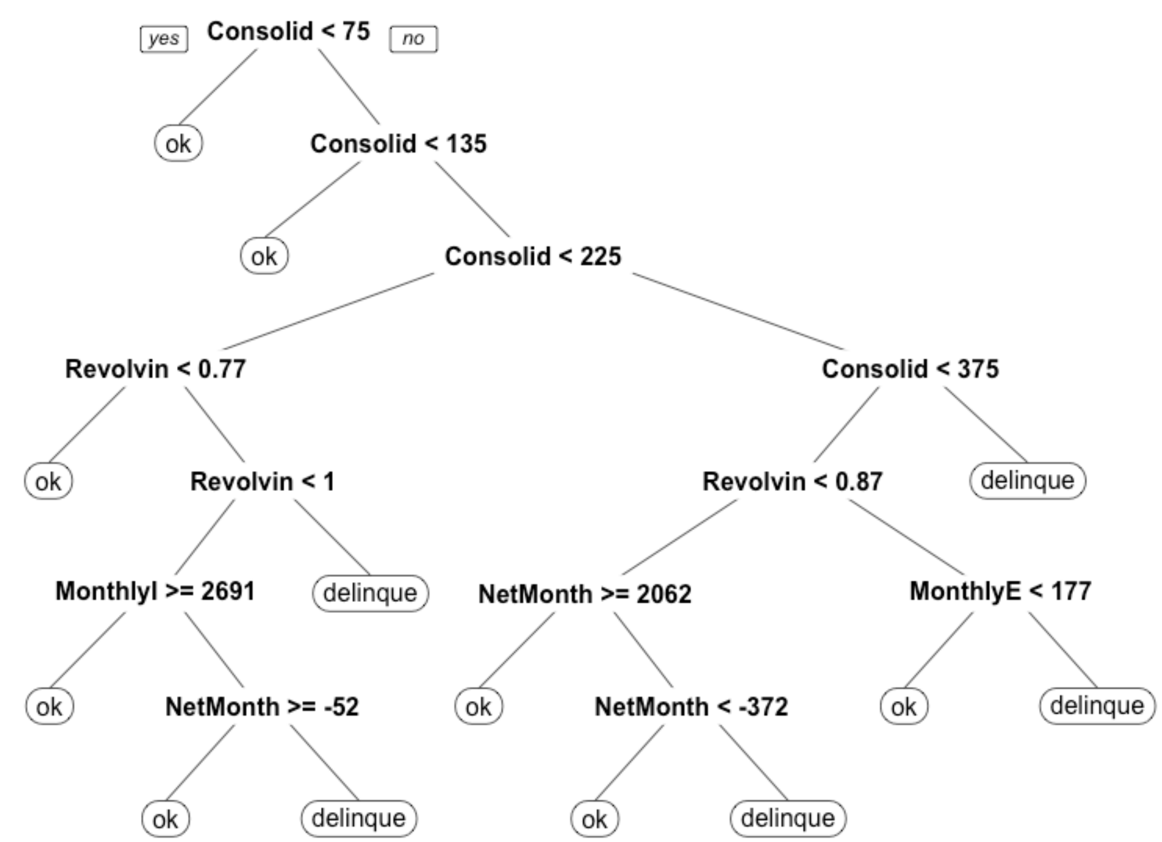
In Second Iteration, built models with constructed features.



EXAMPLE DECISION TREE

FIVE PREDICTORS USED, BUILT ON FULL TRAINING DATASET

- ConsolidatedNumberOfDays PastDue - critical feature
- RevolvingUtilizationOfUnsecuredL ines - also very important
- NetMonthlySurplus
- MonthlyExpenses
- MonthlyIncome



"delinque" -> "Delinquent"

EVALUATE THE MODELS

COMPARE THE RELATIVE PREDICTIVE PERFORMANCE

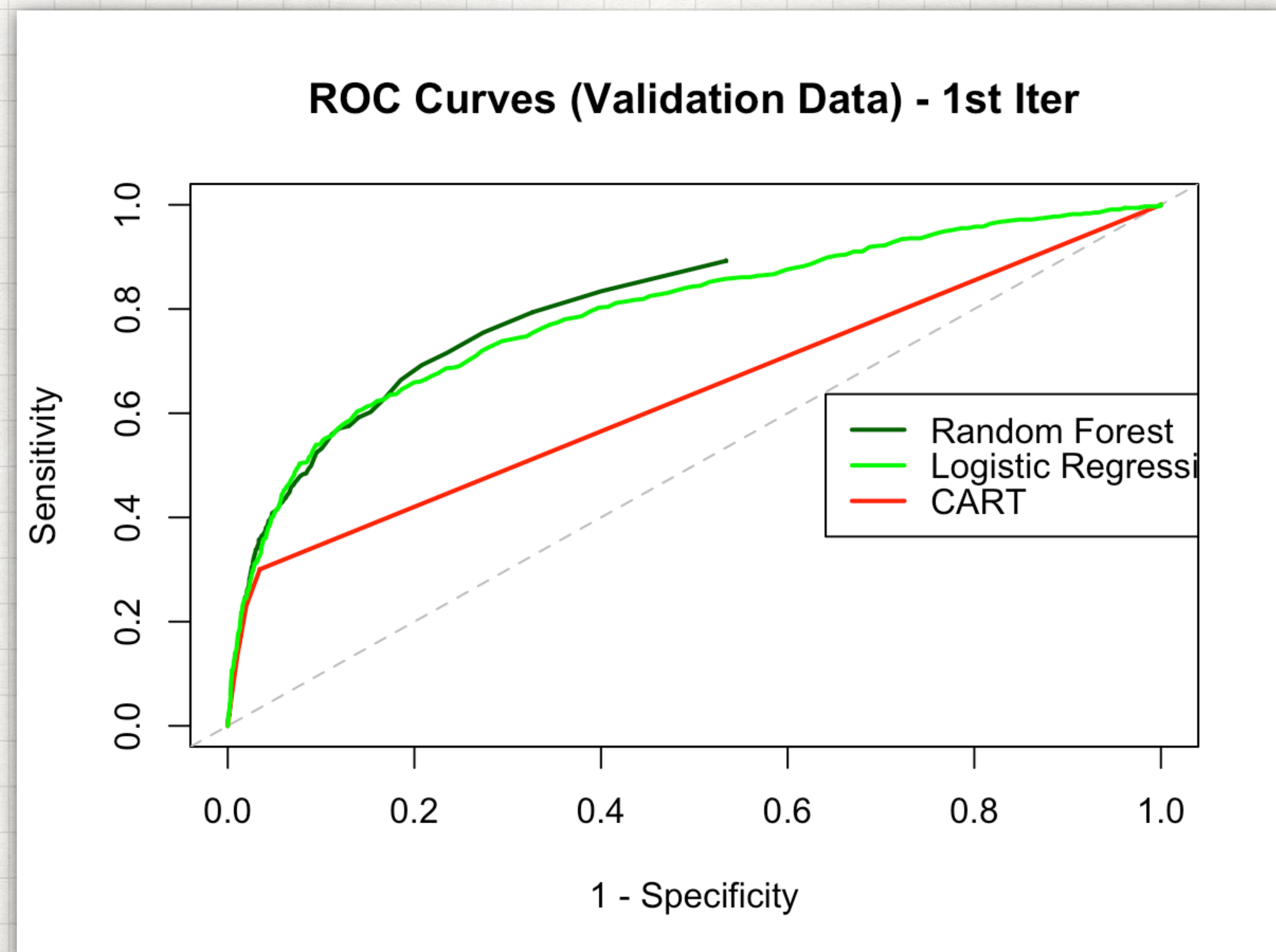
- Mainly ROC Curves were used to evaluate models

- ROC curve visually demonstrates how much better model does at predicting the dependent binary variable compared to choosing an observation at random from the sample data.

  - The area under the curve (AUC) on an ROC chart provides a quantitative measure of how well the model does over random predictions. Baseline performance is shown on ROC charts as a diagonal line from (0,0) to (1,1) and represents an AUC of 0.5-- which is the chance that a model will, at random, predict correctly the dependent variable.

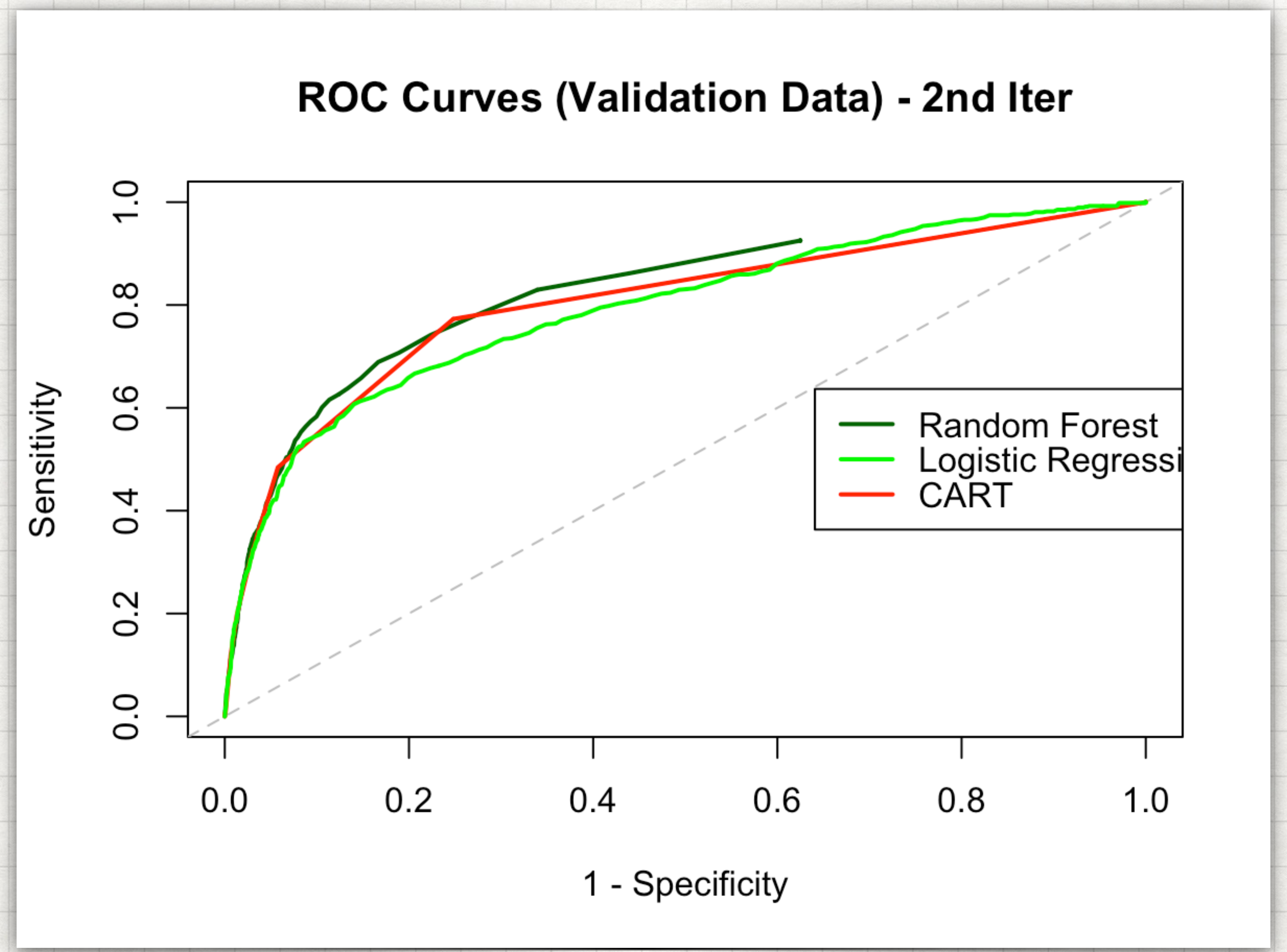
FIRST ITERATION

BUILT WITH ORIGINAL INDEPENDENT VARIABLES IN DATASET



SECOND ITERATION - FEATURE CHANGES

SOME CONSTRUCTED FEATURES USED IN MODEL BUILD



WHICH MODEL TYPES PERFORMED BEST?

OVER THE TWO ITERATIONS, THE ROC CURVES SHOW IT

- The **CART model** performed much better in second iteration with the engineered features included.

- The **logistic regression model** actually degraded slightly in second iteration

- The **Random Forest model** is clearly the best performing model among the three. It improved its lead in the second iteration.

KAGGLE SUBMISSION

CART AND RF AND RF2

- Submitted a few entries on kaggle competition website
- As expected the RF (Random Forest) models performed the best. However, using the entire sample 150,000 training dataset was time consuming
- Below is how my best RF-based submission fared, better than only ~1/5 of teams that were in the competition, with an AUC of **0.832493**. But not bad for a novice like myself I think! My data science journey has started.

712	↑6	dataEngine	0.832583	1	Thu, 01 Dec 2011 03:44:52	
-		**DCrook**	**0.832493**	-	**Tue, 10 Nov 2015 00:51:28**	**Post-Deadline**
713	↑10	PZ	0.832043	8	Sun, 27 Nov 2011 02:44:46 (-24.4h)	

Post-Deadline Entry
If you would have submitted this entry during the competition, you would have been around here on the leaderboard.

KEY LEARNINGS

MAIN LESSONS I LEARNED

- Data cleanup is important and time consuming
- Feature engineering can help make better use of strangely defined data variables
- Obtaining better performing machine learning models can be computationally expensive
- Kaggle competitions really are competitive. They seem to be a good way to hone machine learning chops and learn about other state-of-the-art techniques.