



CAPSTONE PROJECT FOUNDATIONS OF DATA SCIENCE

CRAIG CALDER

MARCH 2016



OVERVIEW

1. Problem statement
2. Data sources
3. Data challenges and preparation
4. Exploratory analysis
5. Top models
6. Summary

PROBLEM STATEMENT

Create a model that a lending organization could use to evaluate a potential customer's likelihood of being a good candidate for a medium term (36 or 60 month) loan based on a modest number of inputs such as annual income, loan amount, term, revolving line utilization rate, etc.

DATA SOURCES

Data: <https://www.lendingclub.com/info/download-data.action/loan.csv>

Contains 163,987 observations with 15 variables. The variable we want to predict is `bad_loans`. The dataset is split between these observations:

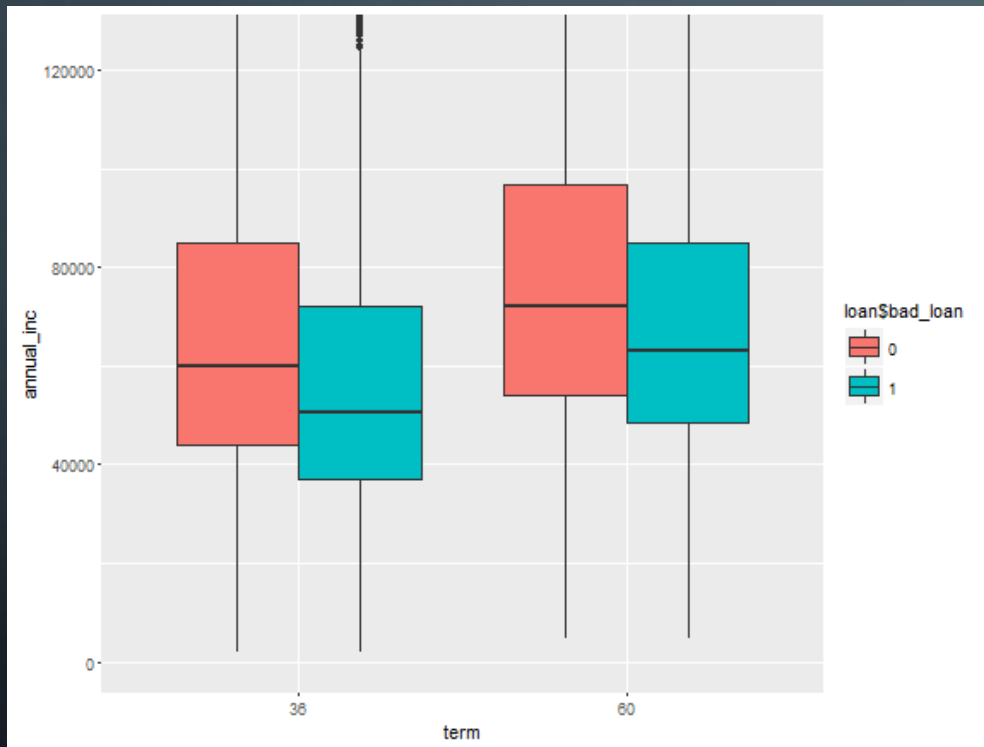
- Good Loans (`bad_loan = 0`): 133,971 observations (82%)
- Bad Loans (`bad_loan = 1`): 30,016 observations (18%)

DATA PREPARATION

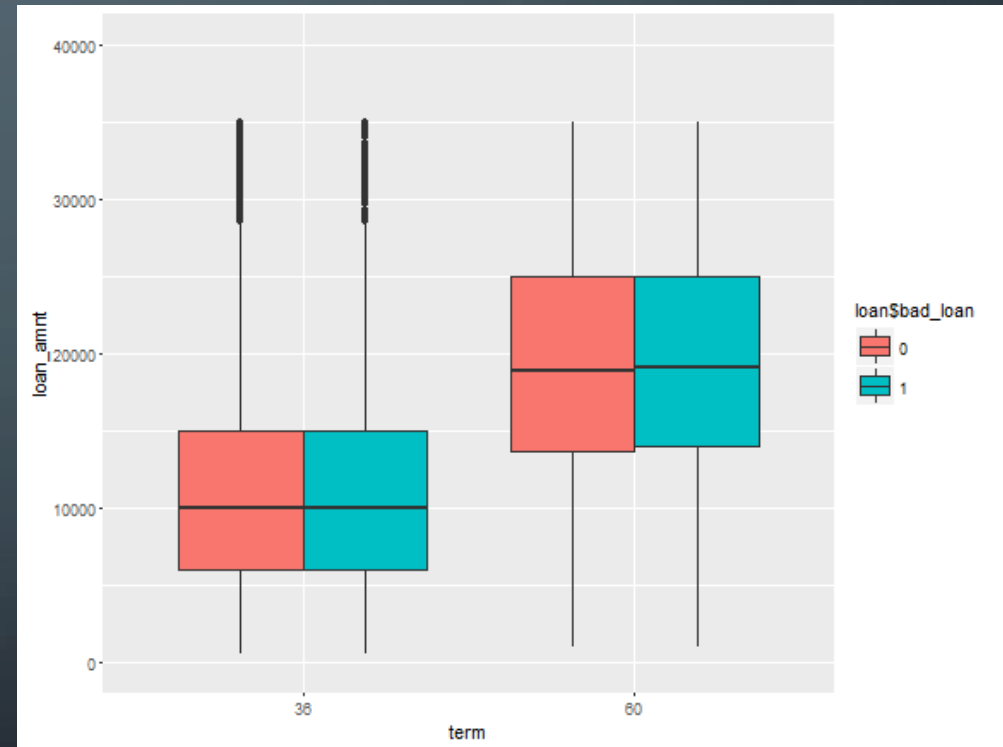
- Wrangling
 - Imputed Missing Values using MICE package
 - Created Dummy Variables from Categorical Variables
 - Identified and Excluded Outliers (Annual Income > \$1M)
 - Balance and Split Training/Test Data using SMOTE

EXPLORATORY DATA ANALYSIS

Unsurprisingly mean annual income is distinctly lower for a bad loan state...

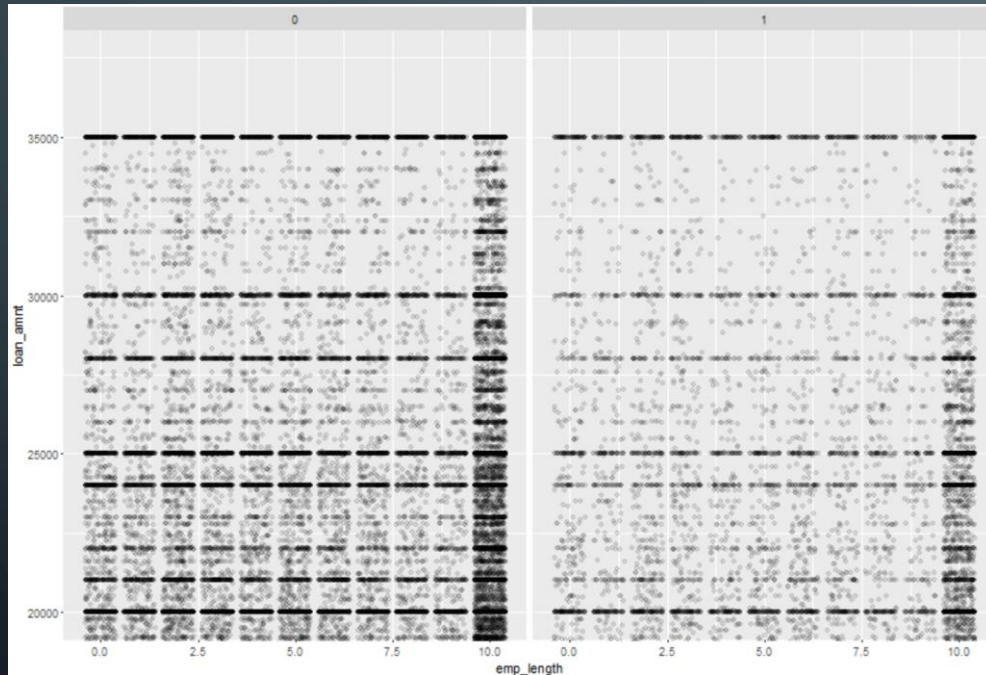


and while loan amounts increase for longer terms, but the means are comparable

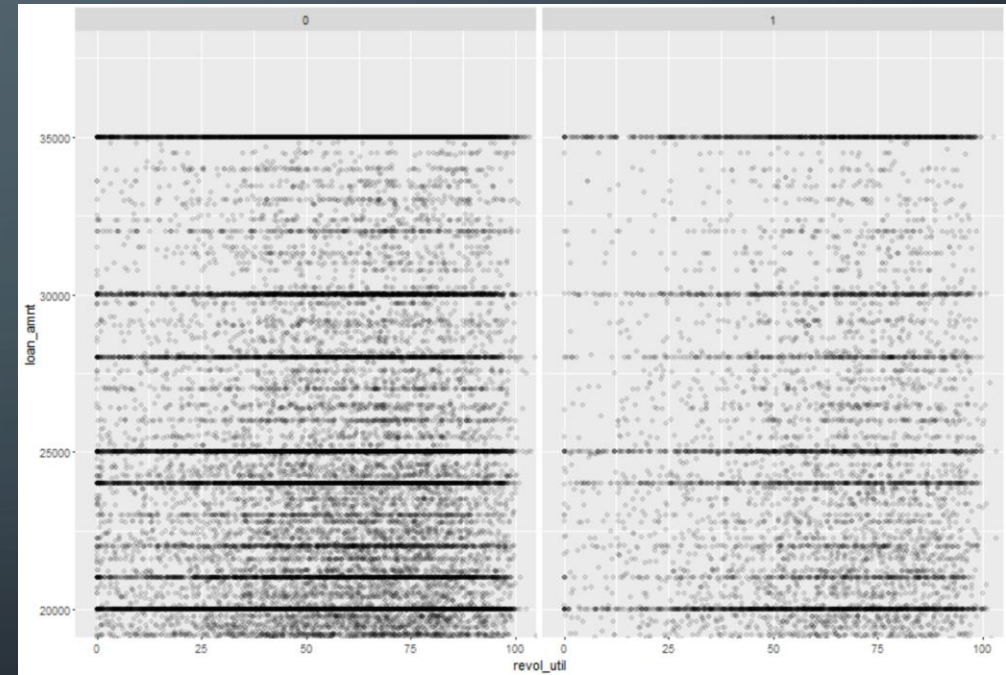


EXPLORATORY DATA ANALYSIS CONT...

Most loans are for similar amounts across all employment lengths. Comparing Good Loans (0) to Bad Loans (1), employment length doesn't appear to be a factor in whether or not a loan will be bad.



Revolving Line Utilization does appear to be a predictive variable of if a loan will be bad as one can see bad loans (1) increases as revol_util increases.



MODEL: LOGISTIC REGRESSION

The table on the right displays the logistic regression models output. From this we see employment length, annual income, total accounts, credit length, living in a low or medium average bankruptcy state and having a mortgage are all inversely related to bad loans.

Below we see the effect of different threshold values on the confusion matrices. It seems $t=.55$ provides balances the desire maximize true positives and true negatives while managing false positives and false negatives.

Threshold	TP	FP	TN	FN
$t > .70$	106,393	6,505	998	4,995
$t > .65$	101,629	5,867	1,636	9,759
$t > .60$	94,395	5,036	2,467	16,993
$t > .55$	83,695	4,056	3,447	27,693
$t > .50$	69,625	2,919	4,584	41,763
$t > .45$	53,333	1,911	5,592	58,055

Variable	Coefficient	P-Value	Significance
Loan_amount	1.84E-05	1.35E-55	***
emp_length	-0.00182616	0.485009021	
annual_inc	-5.39E-06	5.49E-104	***
dti	0.035582836	1.32E-177	***
delinq_2yrs	0.056976745	3.57E-05	***
revol_util	0.011514081	2.61E-195	***
total_acc	-0.00902606	4.00E-28	***
longest_credit_length	-0.00746845	3.89E-08	***
bankrptc_state_low (dummy)	-0.09342886	4.57E-05	***
bankrptc_state_med (dummy)	-0.04470727	0.032735737	*
bankrptc_state_medhigh (dummy)	0.059659539	0.003381705	**
bankrptc_state_high (dummy)	0.096607377	0.000106993	***
homeown_other (dummy)	0.060802301	0.072699274	.
homeown_mort (dummy)	-0.26221461	1.30E-43	***
homeown_rent (dummy)	0.244811429	5.09E-38	***
term (factor)		1.64E-297	***
purpose (factor)		3.41E-16	***
vstatus_verified (factor)		1.54E-67	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

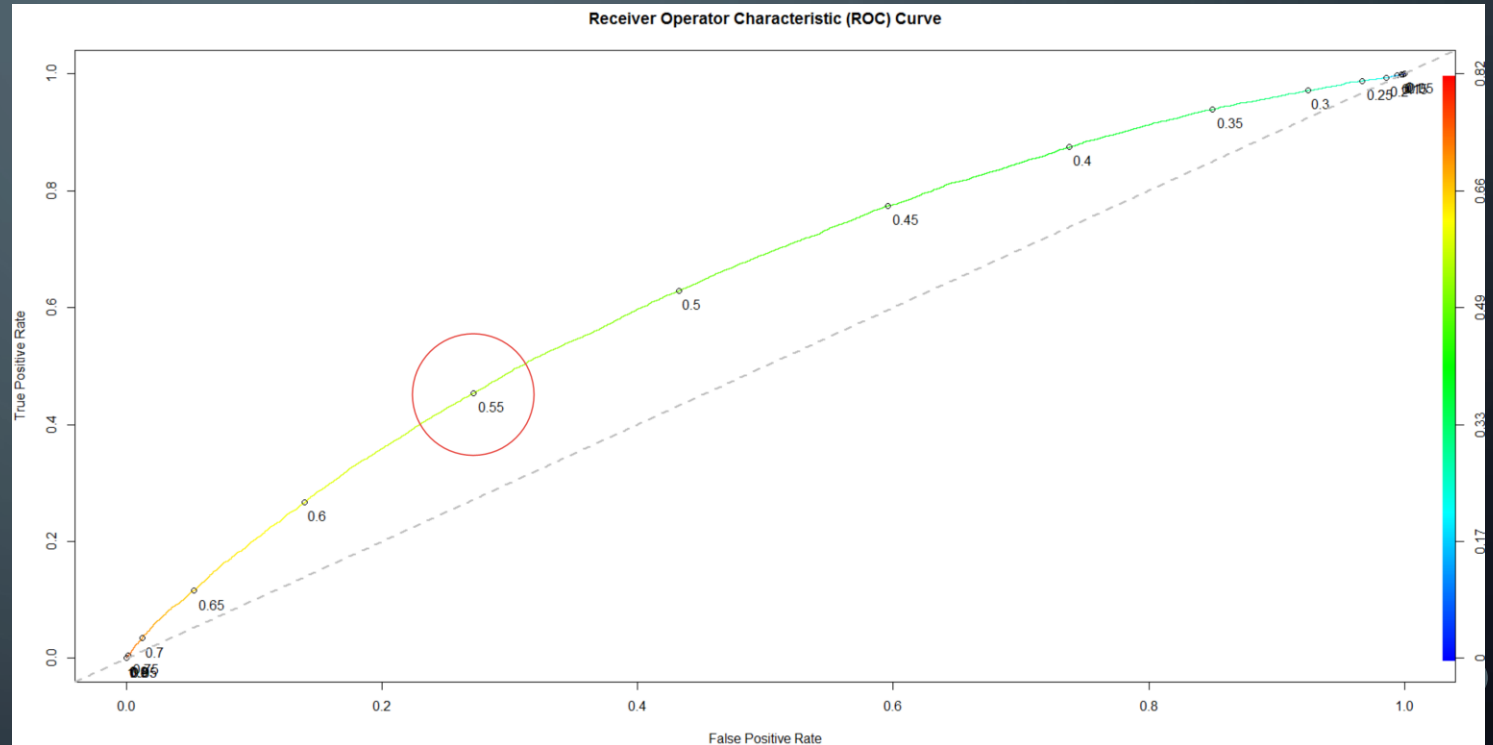
MODEL VALIDATION - ROCR

Area Under the Curve
(AUC) = 0.6618582.

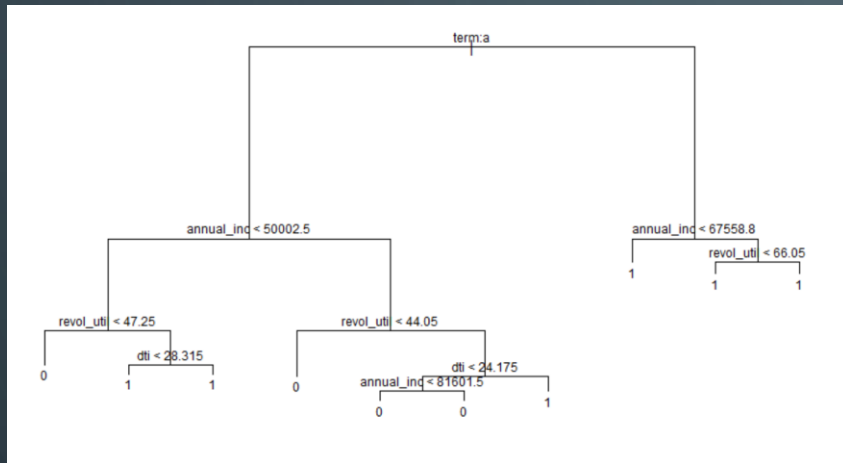
The ROC curve shows the tradeoff
between sensitivity and specificity.

The table below highlights the exact
values at select thresholds, where a
threshold of .55 balances the True
Positive Rate well against False Positive
Rate.

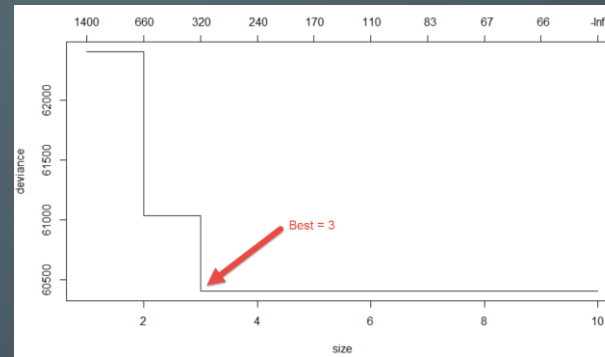
Threshold	Sensitivity	Specificity
$t > .70$	0.9552	0.1330
$t > .65$	0.9124	0.2180
$t > .60$	0.8474	0.3288
$t > .55$	0.7514	0.4594
$t > .50$	0.6251	0.6110
$t > .45$	0.4788	0.7453



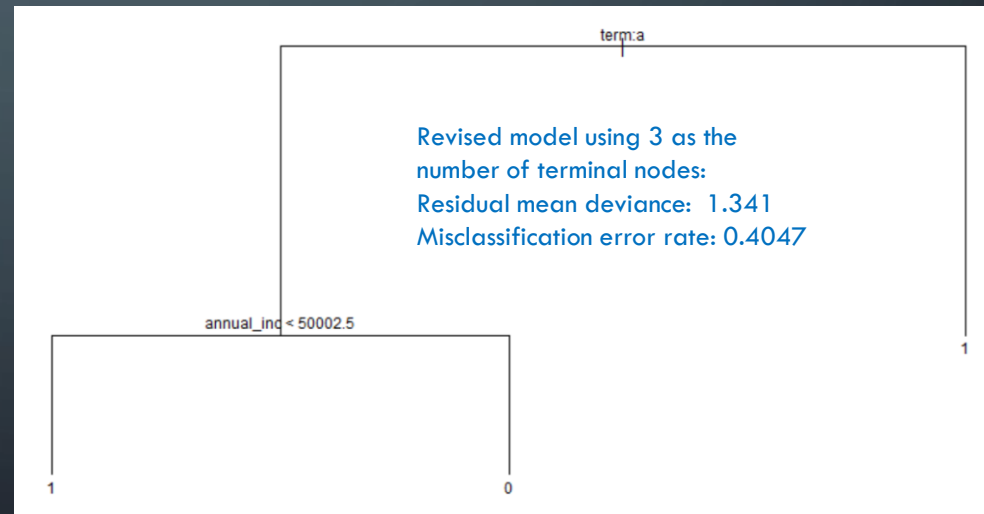
MODEL: CLASSIFICATION TREE



The classification model run with all variables.
Number of terminal nodes: 10
Residual mean deviance: 1.318
Misclassification error rate: 0.3888



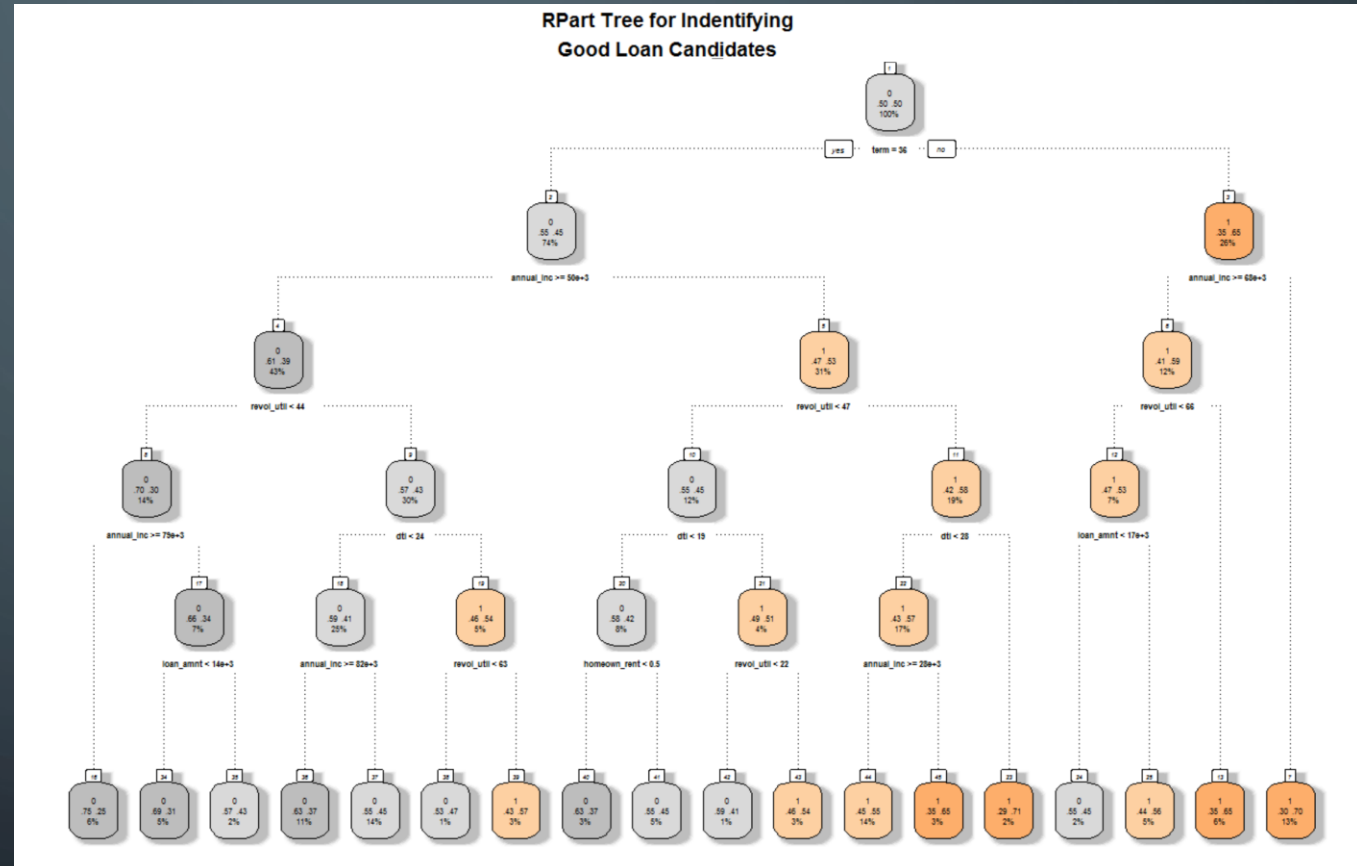
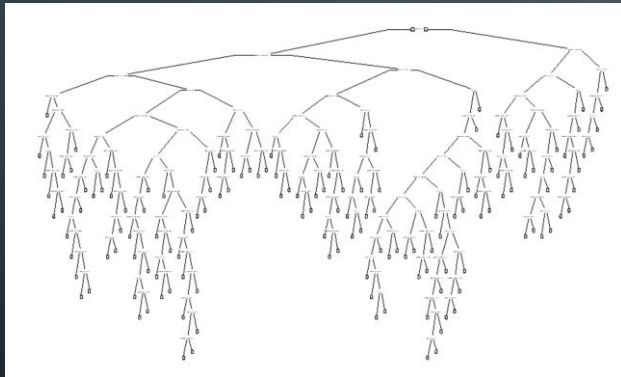
Do K-fold cross-validation to find the best number of nodes for the tree that minimizes overfitting.



Revised model using 3 as the
number of terminal nodes:
Residual mean deviance: 1.341
Misclassification error rate: 0.4047

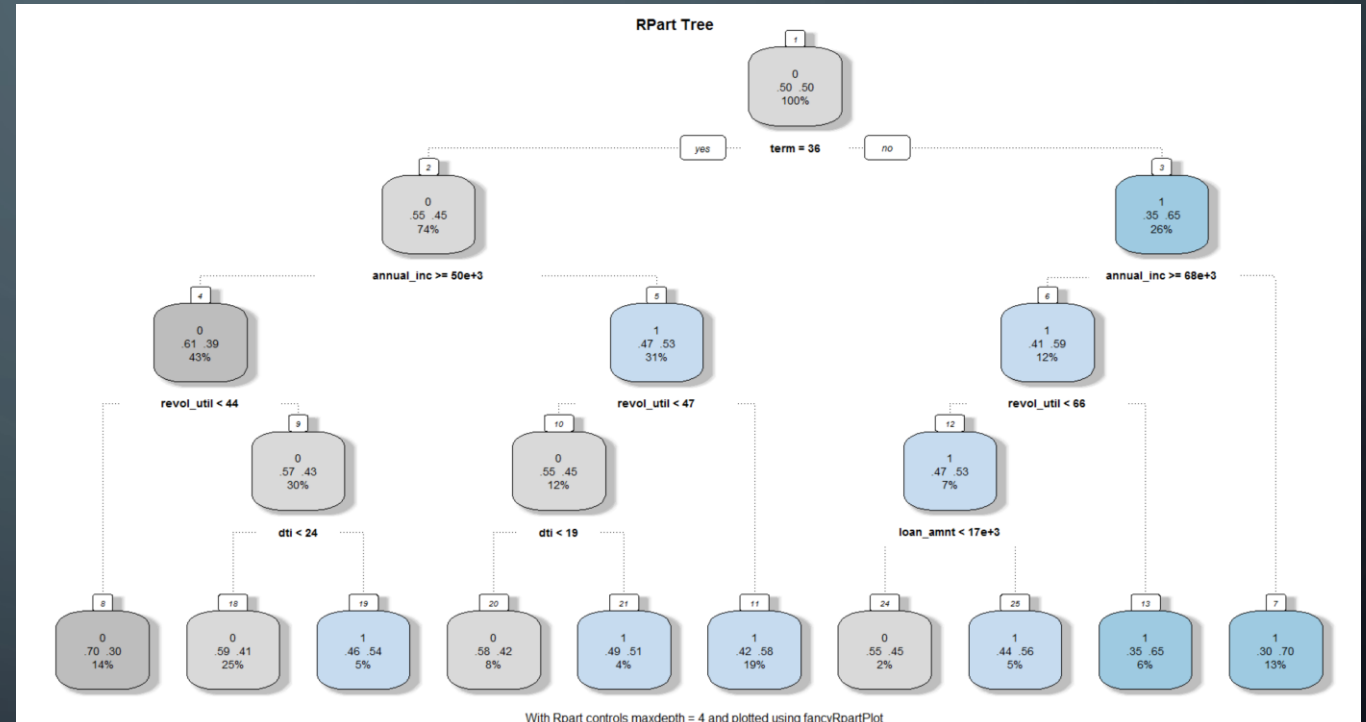
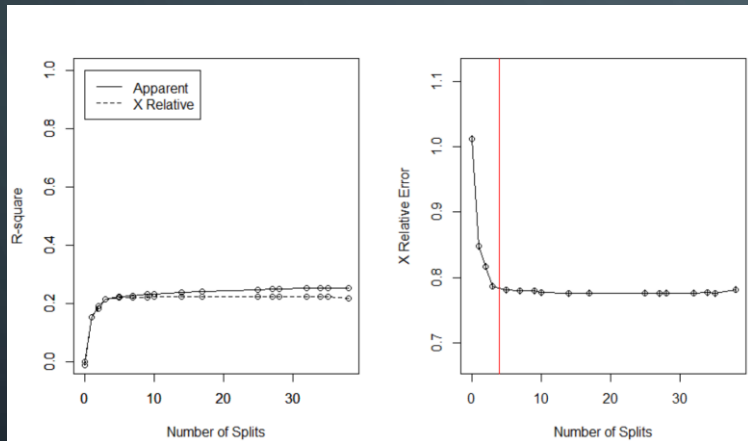
MODEL: RTREE

The Rtree model, shown below using default values, creates an overfitted tree. But by increasing the number of observations assigned to a minimum split, or by controlling the maximum depth, we can reduce the model down to a meaningful number of nodes.





MODEL: RTREE VALIDATION

Analyzing for the optimal depth that minimizes overfitting. The analysis below suggests that either three or four splits are the optimal. In trying both, four resulted in a more compelling tree and is shown on the right.



Variable definitions:
revol_util = Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit.
dti = A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, divided by the borrower's self-reported monthly income.

SUMMARY

Model	Technique	Variable.Imp	AUC
1	Logistic	term, annual_inc, loan_amnt, dti, delinq_2yrs, revol_util, total_acc, longest_credit_length, bankrpc_state_low, bankrpc_state_high, homeown_mort, homeown_rent, purpose, vstatus_verified	66.186% 
2	Decision Tree (Tree)	term, annual_inc	63.492%
3	Decision Tree (RTree)	term, annual_inc, revol_util, dti, loan_amnt	65.718% 

APPENDIX: ORIGINAL VARIABLE DEFINITIONS

#	Variable Name	Description
1	loan_amnt	The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value. Loan amount ranges from 500 to 35,000.
2	term	The number of months to pay the loan. Values are either 36 or 60 months (character type)
3	int_rate	Ranging from 5.42 the best candidate (e.g. lowest risk) to 26.06 for high risk customers. (dependent variable)
4	emp_length	Employment length in years.
5	home_ownership	The home ownership status provided by the borrower during registration. Our values are: RENT, OWN, MORTGAGE, OTHER (character type).
6	annual_inc	The self-reported annual income provided by the borrower during registration.
7	purpose	A category provided by the borrower for the loan request.
8	addr_state	The state provided by the borrower in the loan application
9	dti	A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower's self-reported monthly income.
10	delinq_2yrs	The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years
11	revol_util	Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit.
12	total_acc	Total number of accounts
13	bad_loan	Numeric: 1 = Bad Loan, 0 = Good loan. Captures if the consumer was either a good or bad loan.
14	longest_credit_length	In months.
15	verification_status	Based on the state of the loan application

APPENDIX: ADDED VARIABLES DEFINITIONS

New Variables	Description
bankrpc_state_high	States, grouped into four categories, based on bankruptcy filings by State: http://www.valuepenguin.com/bankruptcy-filings-state
bankrpc_state_medhigh	
bankrpc_state_med	
bankrpc_state_low	
vstatus_verified	Loan status Verified = 1, Not Verified = 0
homeown_other	Collapsed the homeownership value into categories other, mortgage, or rent.
homeown_mort	
homeown_rent	