Credit Risk Analytical Report

Patrick Knott

3 Nov 2017

Introduction

Following the recent discovery that the risk models used previously were too ad-hoc to be suitable for a bank environment with strict regulatory requirements, the task of completely rebuilding the loan default model has arisen. New management has ordered for a 'ground-up' rebuild of the credit risk models using appropriate statistical tools, methods and performance measures. Specifically, with regards to the new credit risk model, management requires answers to their several primary concerns:

1. How does this model perform compared to the previous? How can it be expected to perform on new loans?
2. What are the important variables in this model and how do they compare to variables the bank has found to be traditionally important in its own modelling?
3. What assurances and justifications can be made about the statistical rigor of the model and modelling methodology?

To formally address management's inquiries, statistical techniques can be applied whereby a binomial general linear model is fit to this binary data to provide the base for further inference. Successfully estimating a model will allow conclusions to be formed in relation to the merit of the model compared to the benchmark, and reveal the importance of loan characteristics as determinants of default.

Before statistics can take over and provide answers to the above questions, there is an important step of pre-processing and checking the quality of the underlying data. The next three sections provide a first insight into the patterns inside the loan data, and an insight on the trustworthiness of the data itself.

Section One: Data Viewing

A credit scoring model is a tool that is typically used in the decision-making process of accepting or rejecting a loan. It is the direct result of a statistical model which, based on information about the borrower (e.g. annual income, number of previous loans, etc.), allows banks to distinguish between "good" and "bad" loans and give an estimate of the probability of default. A database export of historical lending between 2007 and 2011 has been provided to act at the foundation of this new model. The goal is to identify which characteristics of loans/loan holders hold a relationship with default, and use this information to build a model to apply for future predictions.

The dependent variable in this circumstance is identified as a binary variable – where 0 indicates that specific loan holder did not default, while 1 indicates they did. Looking into the scope of the given data, there are 38 478 observations, and 15.15% of the loans over this five year period ended in default. This percentage of default is adequate for the modelling process to take place – as any less percentage of defaults may not provide enough data to accurately identify which characteristics influence default.

```
dim(dataset)
```

```
## [1] 38478    41
```

```
sum(dataset$repay_fail)/length(dataset$repay_fail)
```

```
## [1] 0.1514632
```

```
head(dataset)
```

```
##   X     id member_id loan_amnt funded_amnt funded_amnt_inv      term
## 3 3 545583    703644      2500        2500            2500 36 months
## 4 4 532101    687836      5000        5000            5000 36 months
## 5 5 877788   1092507      7000        7000            7000 36 months
## 6 6 875406   1089981      2000        2000            2000 36 months
## 7 7 506439    652909      3600        3600            3600 36 months
## 8 8 981465   1204637      8000        8000            8000 36 months
##   int_rate installment grade sub_grade emp_length home_ownership
## 3    13.98       85.42     C        C3    4 years           RENT
## 4    15.95      175.67     D        D4    4 years           RENT
## 5     9.91      225.58     B        B1  10+ years       MORTGAGE
## 6     5.42       60.32     A        A1  10+ years           RENT
## 7    10.25      116.59     B        B2  10+ years       MORTGAGE
## 8     6.03      243.49     A        A1       <NA>       MORTGAGE
##   annual_inc verification_status issue_d
## 3      20004        Not Verified  1/7/10
## 4      59000        Not Verified  1/6/10
## 5      53796        Not Verified  1/9/11
## 6      30000        Not Verified  1/9/11
## 7     675048        Not Verified  1/4/10
## 8      77736            Verified 1/10/11
##                                         loan_status            purpose
## 3 Does not meet the credit policy. Status:Fully Paid              other
## 4                                        Charged Off debt_consolidation
## 5                                         Fully Paid              other
## 6                                         Fully Paid debt_consolidation
## 7 Does not meet the credit policy. Status:Fully Paid              other
## 8                                         Fully Paid              other
##   zip_code addr_state   dti delinq_2yrs earliest_cr_line inq_last_6mths
## 3    487xx         MI 19.86           0           1/8/05              5
## 4    115xx         NY 19.57           0           1/4/94              1
## 5    751xx         TX 10.80           3           1/3/98              3
## 6    112xx         NY  3.60           0           1/1/75              0
## 7    352xx         AL  1.55           0           1/4/98              4
```

```
## 8    853xx          AZ  6.07           0           1/7/96              0
##   mths_since_last_delinq open_acc pub_rec revol_bal revol_util total_acc
## 3                     NA        7       0       981     21.30%        10
## 4                     59        7       0     18773     99.90%        15
## 5                      3        7       0      3269     47.20%        20
## 6                     72        7       0         0         0%        15
## 7                     25        8       0         0         0%        25
## 8                     NA       12       0      4182     13.60%        49
##   total_pymnt total_pymnt_inv total_rec_prncp total_rec_int recoveries
## 3    3075.292         3075.29         2500.00        575.29          0
## 4    2948.760         2948.76         1909.02        873.81        151
## 5    8082.392         8082.39         7000.00       1082.39          0
## 6    2161.663         2161.66         2000.00        161.66          0
## 7    4206.031         4206.03         3600.00        606.03          0
## 8    8724.972         8724.97         8000.00        724.97          0
##   last_pymnt_d last_pymnt_amnt next_pymnt_d last_credit_pull_d repay_fail
## 3       Jul-13           90.85       Aug-13             Jun-16          0
## 4       Nov-11          175.67         <NA>             Mar-12          1
## 5       Mar-14         1550.27         <NA>             Mar-14          0
## 6       Feb-14           53.12         <NA>             Jun-16          0
## 7       May-13          146.75       Jun-13             Jun-16          0
## 8       Apr-14         1423.66         <NA>             Apr-14          0
##   time_since_first_loan
## 3              4.917808
## 4             16.178082
## 5             13.512329
## 6             36.690411
## 7             12.008219
## 8             15.260274
```

Accompanying the dependent variable are 37 characteristics of the loans. These independent variables are a mix of loan characteristics, personal characteristics of the applicant, personal financial status of the applicant, and recordings of loan specific quantities. For a variety of reasons, not all of these predictors will be eligible for the binomial model, and can be culled accordingly. The following section will conduct multiple methods to reduce this set of 37 characteristics into a small group of variables that can be applied to the forthcoming statistical processes.

Section Two: Manually Culling Variables

The aim here is to build a model that predicts the probability of loan default based on information known at the time of application. Looking into the variables given, many are ones that are revealed throughout the duration of the loan, such as number of recoveries or last payment amount, so they could not be considered for this model. As a result of this reasoning, 16 independent variables were immediately removed from the potential model candidates.

- Funded amount
- Funded amount Investors
- Issue Date
- Loan Status
- Months since last Delinquency
- Revolving Balance
- Revolving Line Utilisation Rate
- Total Payment
- Total Payment Investors
- Total principal received
- Total interest received
- Recoveries
- Last Payment Date
- Last Payment Amount
- Next Payment Date
- Last Credit Pull Date

It is understood that sometimes this information can be collated from previous loans in a customer's name and used to predict their new information, however these variables were specifically related to the loan that was recorded as defaulted or not – so they could not be included here.

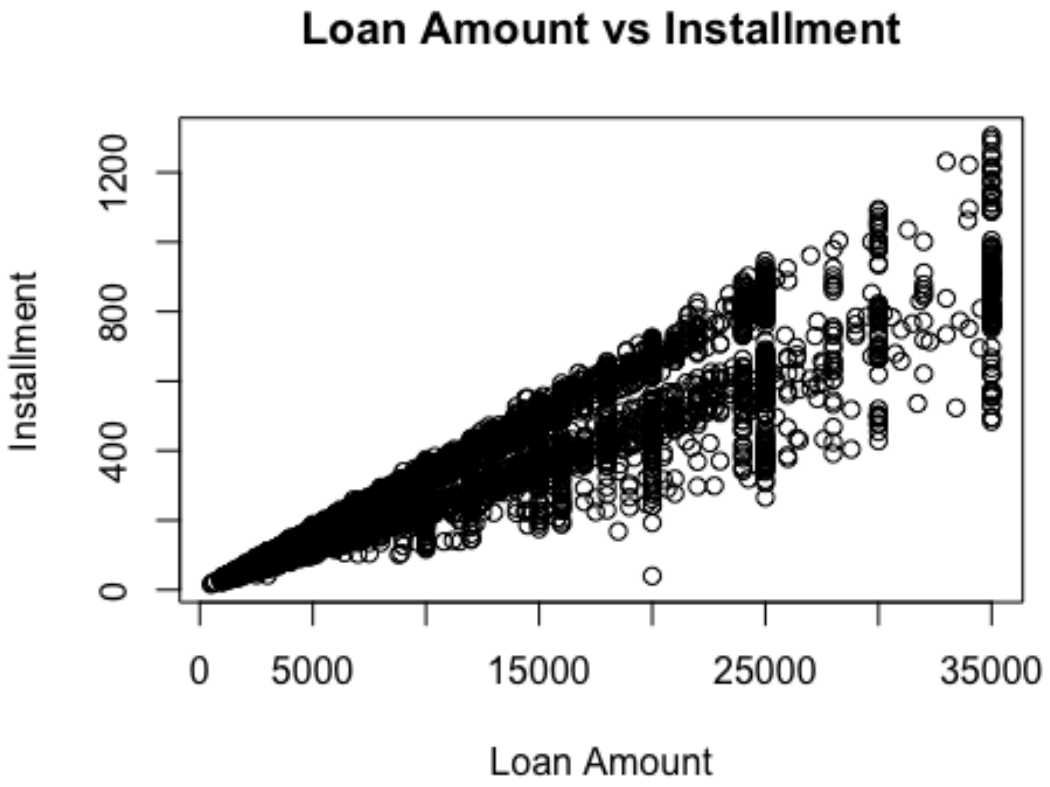
With further verbal reasoning and industry advice, grade, sub-grade and verification status were all removed as they relate to a previous score system for applicants that have since been replaced, thus they are no longer relevant. Similarly, interest rate was also disregarded as it is more of an economic variable rather than a personal characteristic, so again, it doesn't suit the model's scope.

Using the two dates provided in the dataset, a new variable referring to the time in months of this new loan since the first loan in the applicant's credit history was introduced. This

was created using the difference between the ‘Issue date’ and ‘Earliest credit line’. It is possible that such a measurement may reveal insight into a borrower’s history and perhaps be a predictor of default – so it is added to the list of variables to be statistically considered.

###Correlation Matrix With 15 potential predictor variables remaining, a correlation matrix was constructed including any characteristic measured in a continuous way – such as the dollar value of loan amount. When two or more variables are highly correlated, it means they follow the same trend and have the same effect on the dependent variable – so usually only one is kept to describe that phenomenon. The data was investigated for any correlations that exceeded 0.85 – where 1 describes perfect correlation. As is shown in the correlation matrix and plot, loan amount and instalment were found to be highly correlated and display a very similar trend. Specifically, as loan amount increases, instalment is also found to increase at an almost identical linear manner. From both intuition and professional advice, instalment was removed. It was therefore concluded that loan amount would appropriately cover any insight into defaulting probabilities that would have been provided by instalment.

	loan_amnt	installment	annual_inc	dti	delinq_2yrs	inq_last_6mths	pub_rec	total_acc	time_since_first_loan
loan_amnt	1	0.93	0.28	0.06	-0.03	-0.03	-0.05	0.26	0.21
installment	0.93	1	0.28	0.05	-0.02	-0.01	-0.05	0.23	0.18
annual_inc	0.28	0.28	1	-0.11	0.02	0.03	-0.02	0.24	0.19
dti	0.06	0.05	-0.11	1	-0.04	0.02	0	0.24	0.06
delinq_2yrs	-0.03	-0.02	0.02	-0.04	1	0.02	0.01	0.07	0.06
inq_last_6mths	-0.03	-0.01	0.03	0.02	0.02	1	0.06	0.09	-0.04
pub_rec	-0.05	-0.05	-0.02	0	0.01	0.06	1	-0.01	0.05
total_acc	0.26	0.23	0.24	0.24	0.07	0.09	-0.01	1	0.39
time_since_first_loan	0.21	0.18	0.19	0.06	0.06	-0.04	0.05	0.39	1



Section Three: Checking the Data

Addressing Missing Information

Before the modelling could begin, the data was checked for the absence of information. Of the remaining variables being considered for the binomial regression, two held missing points. For annual income, one information point was missing – therefore that observation was simply removed. For employment length, however, there were over 900 observations with this information missing, so it was not appropriate to remove these before investigating whether the presence of "NA" held an effect of its own. A new level/category of employment length called NA was introduced, which used an indicator variable to identify these observations – 1 being the observation of annual income was not present, and 0 being the information was present. A univariate binomial generalised linear model

was fit containing default as the binary dependent variable and employment length as the sole dependent with the purpose of investigating whether NA showed up as statistically significant or not. As shown, NA was extremely statistically significant, therefore there was strong evidence to suggest these observations should not be removed from the data, but rather an additional application category of NA should be added to the options. In terms of application categories for a new loan, the options associated with employment length less than one year, one year, …, 10+ years, with the new addition of NA. (Perhaps this may indicate an applicant is unemployed).

```
missing.count
```

```
##   repay_fail loan_amnt term installment emp_length home_ownership
## 1          0         0    0           0        993              0
##   annual_inc purpose dti delinq_2yrs inq_last_6mths pub_rec total_acc
## 1          1       0   0           0              0       0         0
##   time_since_first_loan
## 1                     0
```

Addressing Outliers

In assessing the distribution of the data amongst the predictor variables, it was noticed that most of the annual incomes were grouped together in a similar range - apart from some extreme outliers. It is possible that these outliers could hold the ability to skew the effects of the income variable, which would cause the need for a change in data form. To investigate this, a univariate model was fit containing default as the binary dependent variable and annual income in its continuous form as the dependent.

Continuous Annual Income

```
summary(fit.cont.annual_inc)$coefficient[,4]
```

```
##  (Intercept)   annual_inc
## 0.000000e+00 3.444344e-19
```

```
summary(fit.cont.annual_inc)$aic
```

```
## [1] 32629.34
```

Discretized Annual Income

```
summary(fit.fac.annual_inc)$coefficient[,4]
```

```
##                       (Intercept)     fac.annual_inc(3e+04,3.7e+04]
##                      0.000000e+00                      4.294214e-01
##  fac.annual_inc(3.7e+04,4.44e+04]    fac.annual_inc(4.44e+04,5e+04]
##                      4.616798e-03                      7.192085e-05
##    fac.annual_inc(5e+04,5.86e+04] fac.annual_inc(5.86e+04,6.53e+04]
##                      6.859291e-07                      8.223736e-05
## fac.annual_inc(6.53e+04,7.53e+04]    fac.annual_inc(7.53e+04,9e+04]
##                      7.473333e-11                      2.067183e-14
```

```
##    fac.annual_inc(9e+04,1.16e+05]    fac.annual_inc(1.16e+05,6e+06]
##                      7.772423e-22                      8.479773e-18
```

```
summary(fit.fac.annual_inc)$aic
```

```
## [1] 32559.98
```

Following this, annual income was grouped into 10 equal bins. As is shown, the AIC for the categorical depiction of Annual Income is smaller than that of the continuous.While identifying numerical values into specific groups can sometimes lead to a loss of information, it is proven to enhance the statistics in th. is situation by removing the effects of these extreme outliers – therefore annual income will be treated in this categorical manner for the remaining investigation.

###Variables Remaining for Model Fitting After completing this manual culling of the data through intuition and statistical analysis, 14 independent variables remain as potential candidates for the model fitting process. The variables are a good mix of loan characteristics, personal characteristics of the applicant and personal financial status of the applicant – however it is the statistical process used in the following section that will decide which of these can be used as predictors of loan default.

Loan Characteristics:

- Loan amount
- Term
- Purpose

Personal Characteristics

- Employment length
- Home ownership status
- Annual income
- Public Record

Personal Financial Status

- Debt to income ratio
- Delinquencies in the last 2 years
- Earliest credit line
- Inquires in the last 6 months
- Number of open accounts
- Total accounts
- Time since the first loan

#Section Four: Model Fitting ###Logistic Regression Assuming past behaviour is a predictor of future behaviour, the aim is to create a statistical model to be able to predict

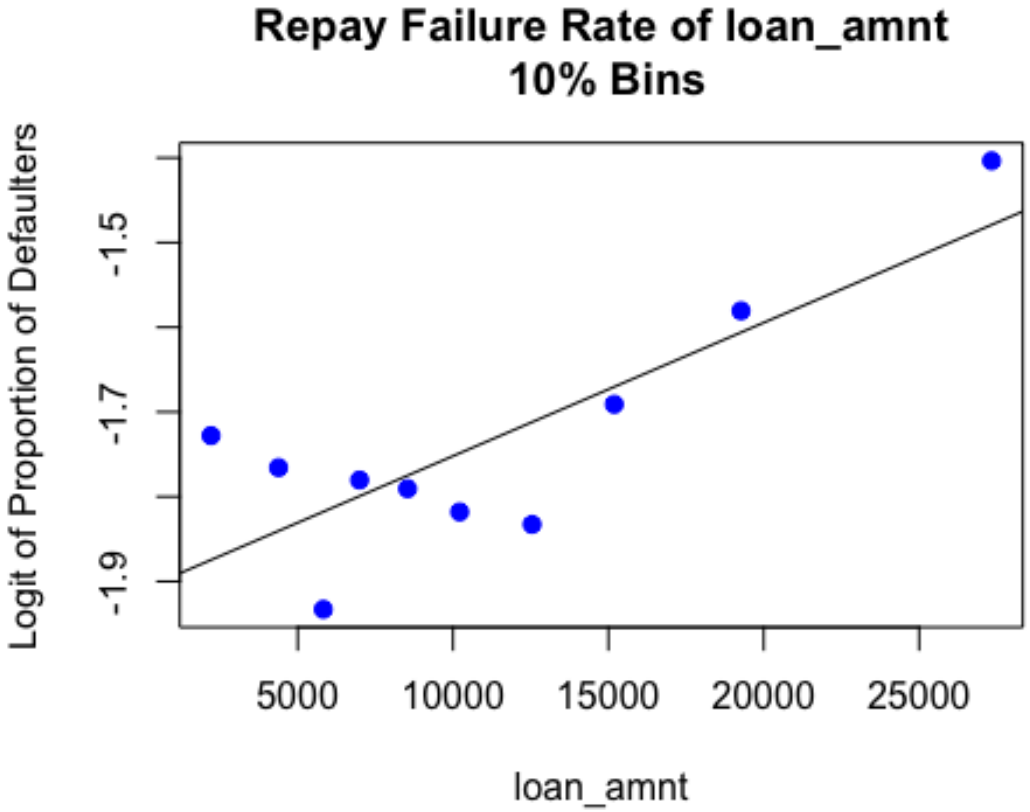
the probability that a new debtor will not repay the debt-holder. One of the most common and successful ways to create a model from binary data via a binomial generalised linear model – specifically with the logistic function. The logistic regression is suited to the credit risk model because many of the independent variables are categorical, the result is required to be a probability/percentage (which is not possible for the linear regression model) and the variability of the dependent variable is not constant (whereas it is assumed constant in the linear regression). This logistic function is the log of odds ratio and provides a ratio of the probability of default occurring against it not.

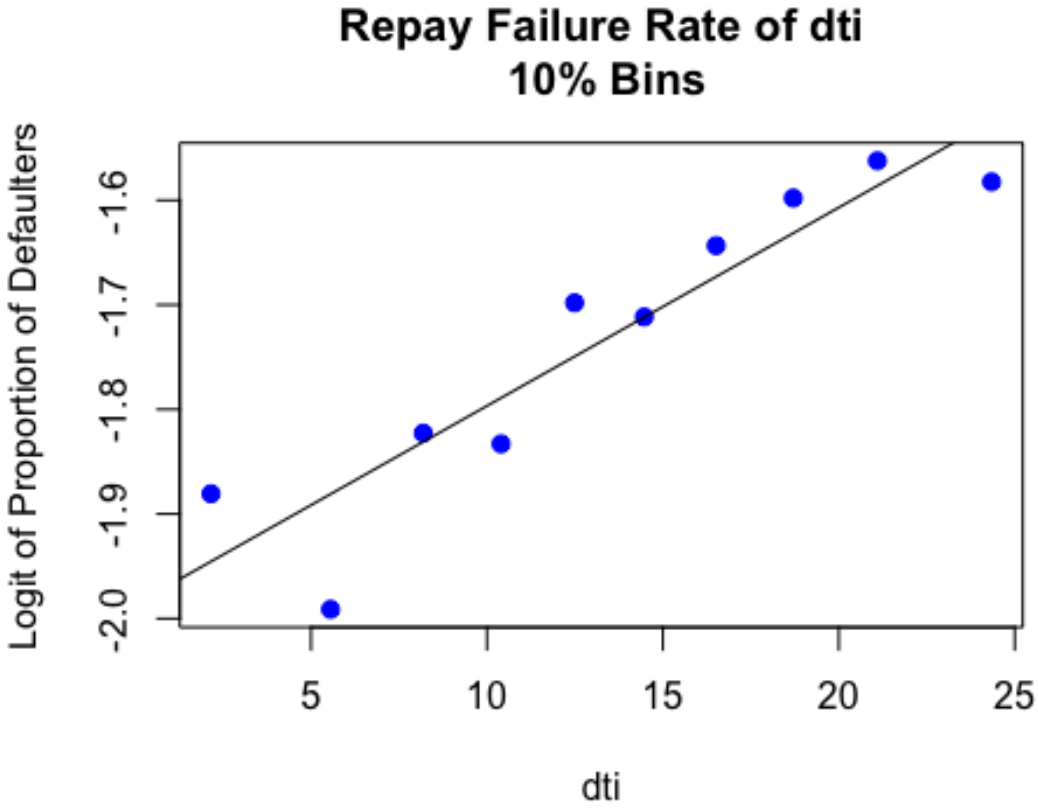
The logistic model as shown takes as input the client characteristics and outputs the probability of default.

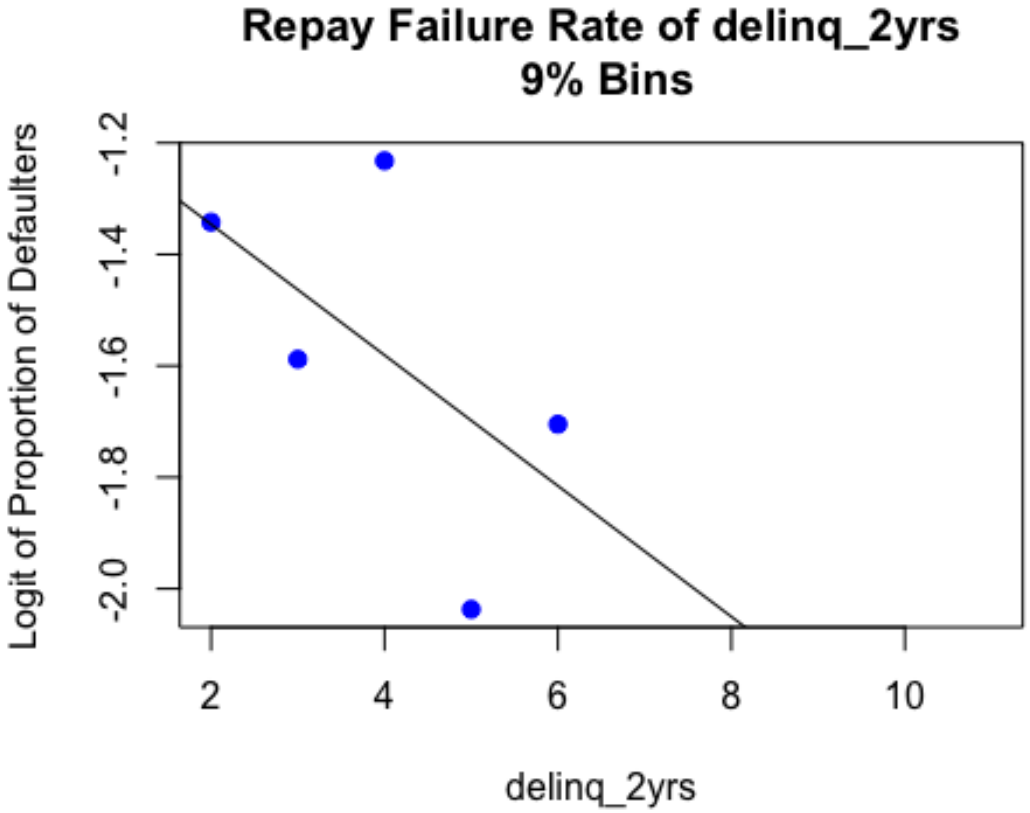
$$p = \exp(\beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n) / (1 + exp(\beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n))$$

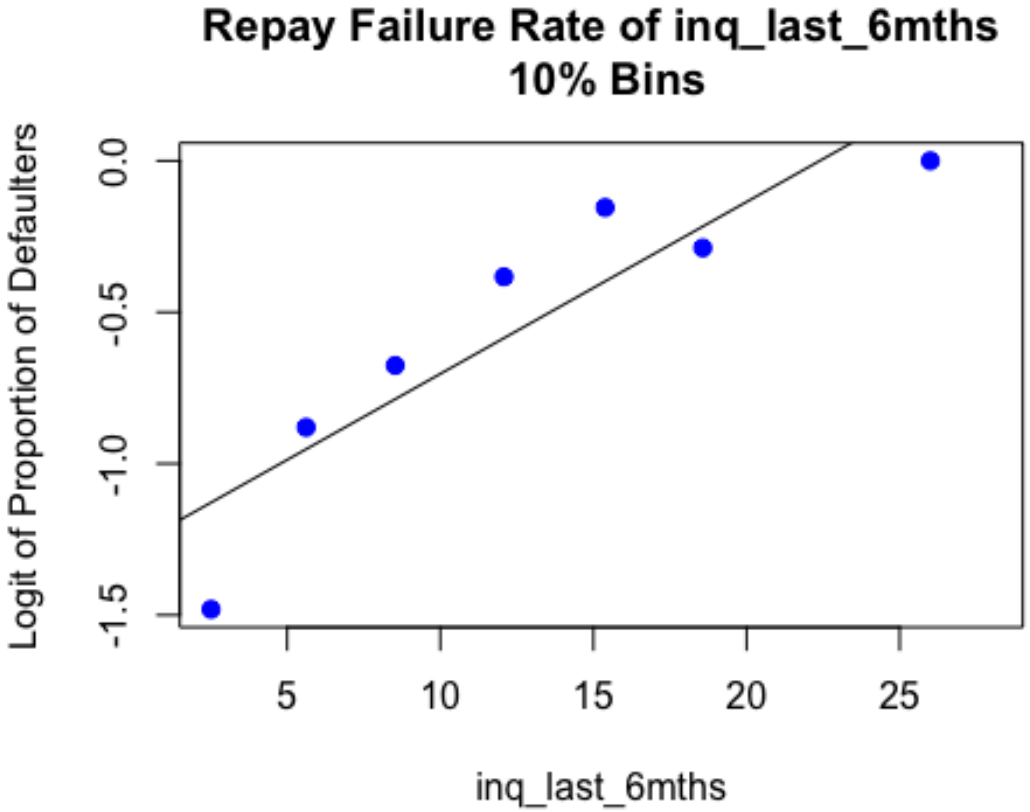
Where:
P= probability of default
Xi = the explanatory variable i
Bi = the regression coefficient of the explanatory variable i
N = the number of explanatory variables

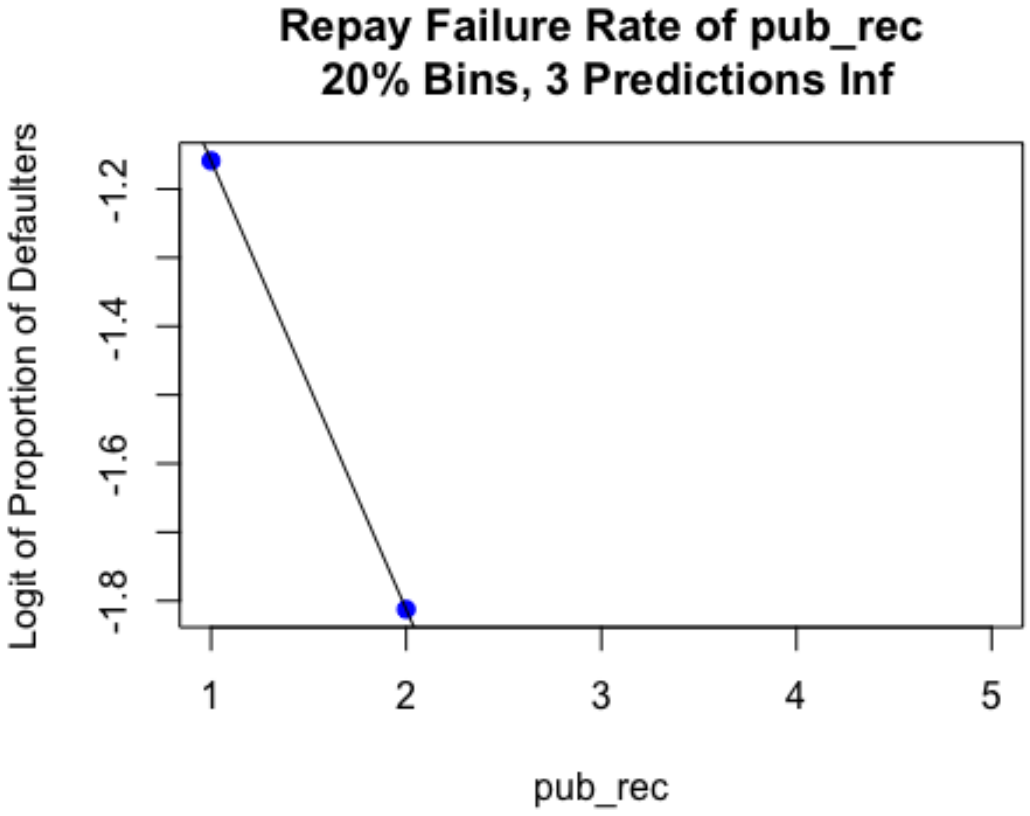
###Preliminary Plots For the logistic model to be appropriate in this content and with the given data, plots of each of the independent variables against the logit of the dependent variable need to be assessed. The plots are analysed for their visual form, where an approximate linear increase or decrease suggests the variable is suited to the logistic regression.

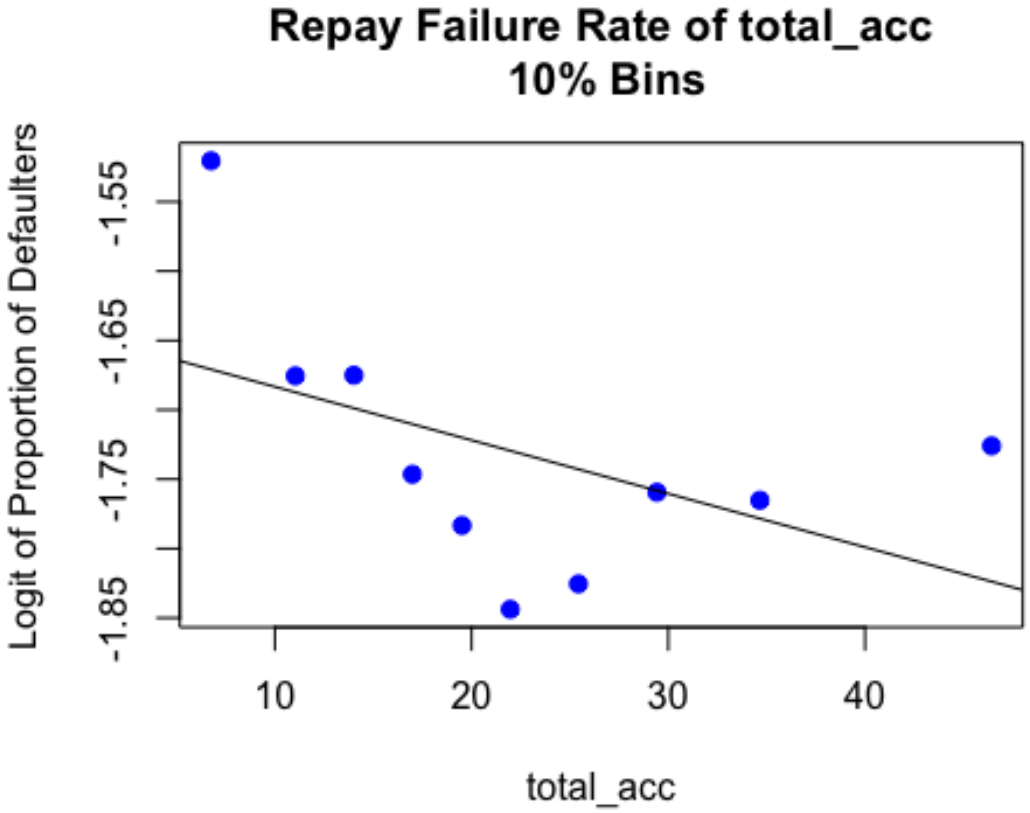

Repay Failure Rate of loan_amnt
10% Bins
Logit of Proportion of Defaulters
-1.9
-1.7
-1.5
5000
10000
15000
20000
25000
loan_amnt

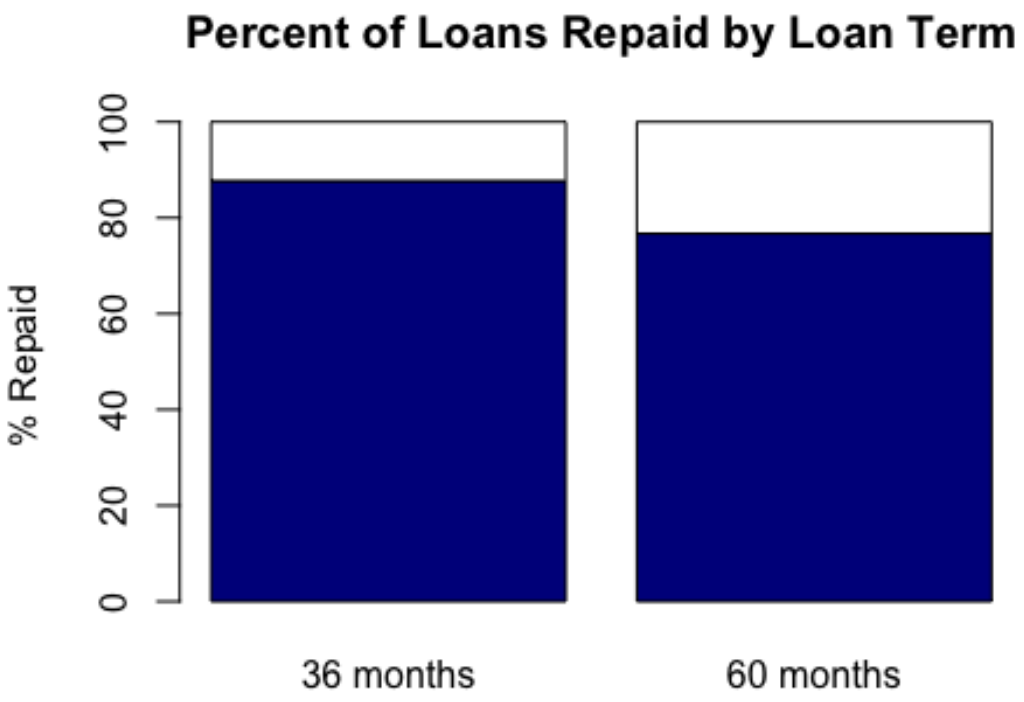

Repay Failure Rate of dti
10% Bins
Logit of Proportion of Defaulters
-2.0
-1.9
-1.8
-1.7
-1.6
5
10
15
20
25
dti

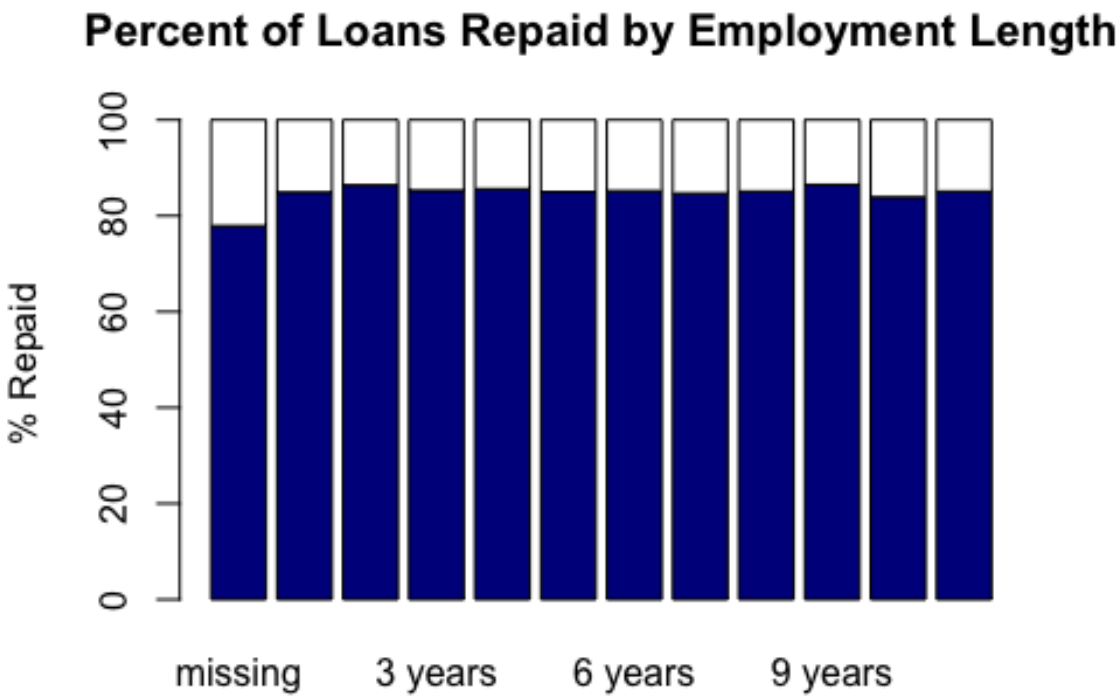

Repay Failure Rate of delinq_2yrs
9% Bins
Logit of Proportion of Defaulters
-1.2
-1.4
-1.6
-1.8
-2.0
2
4
6
8
10
delinq_2yrs

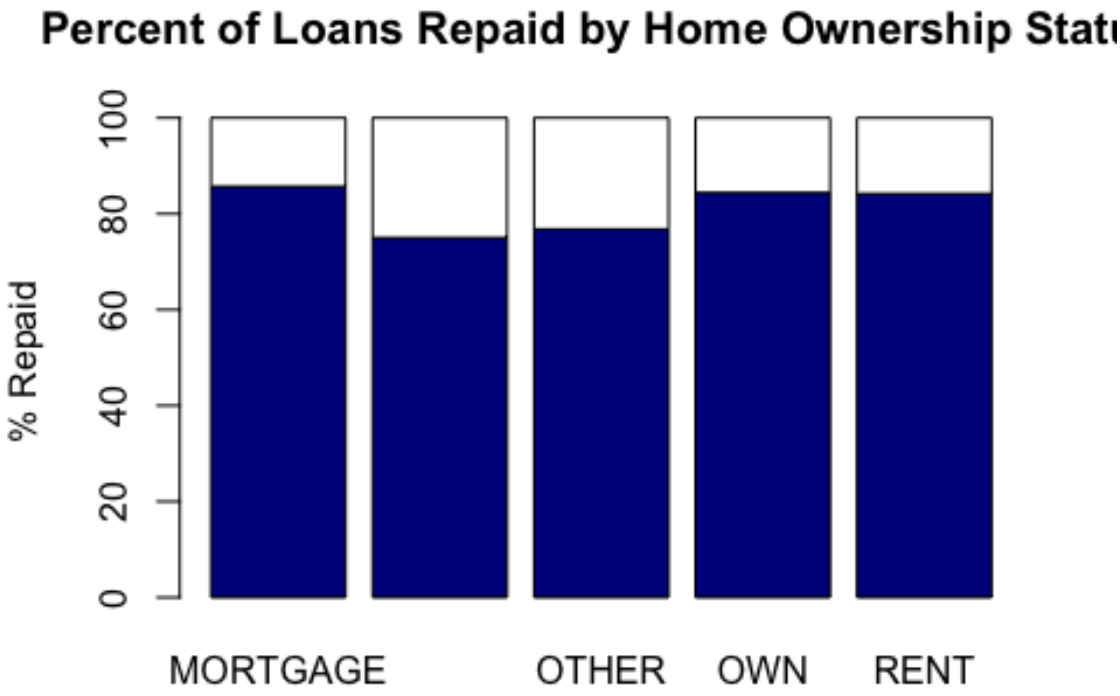

Repay Failure Rate of inq_last_6mths
10% Bins
Logit of Proportion of Defaulters
0.0
-0.5
-1.0
-1.5
5
10
15
20
25
inq_last_6mths

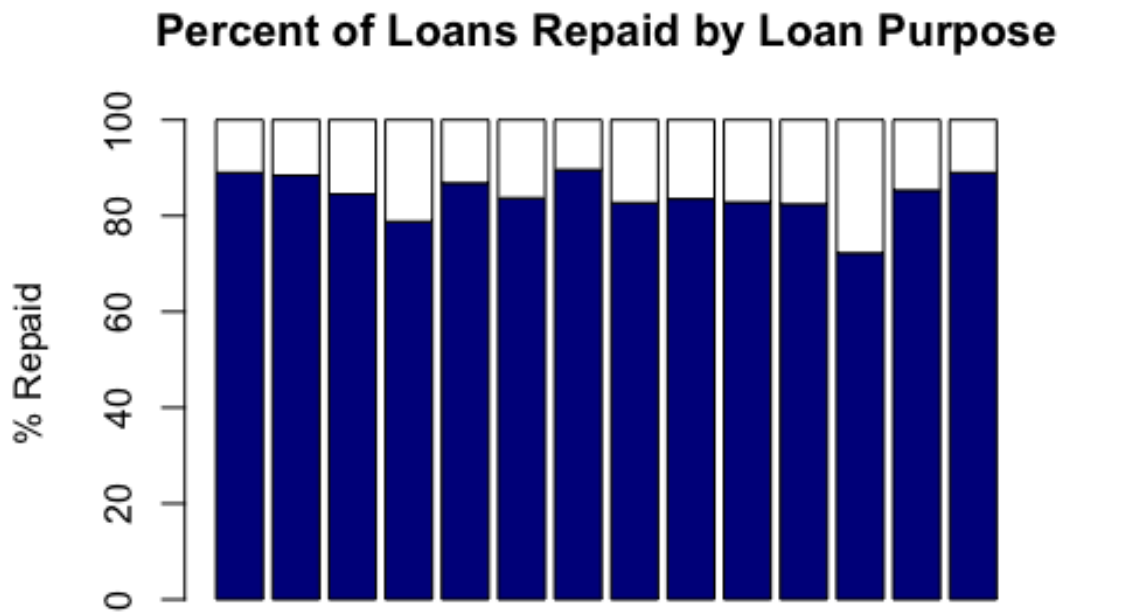

Repay Failure Rate of pub_rec
20% Bins, 3 Predictions Inf
Logit of Proportion of Defaulters
-1.2
-1.4
-1.6
-1.8
1
2
3
4
5
pub_rec


Repay Failure Rate of total_acc
10% Bins
Logit of Proportion of Defaulters
-1.55
-1.65
-1.75
-1.85
10
20
30
40
total_acc

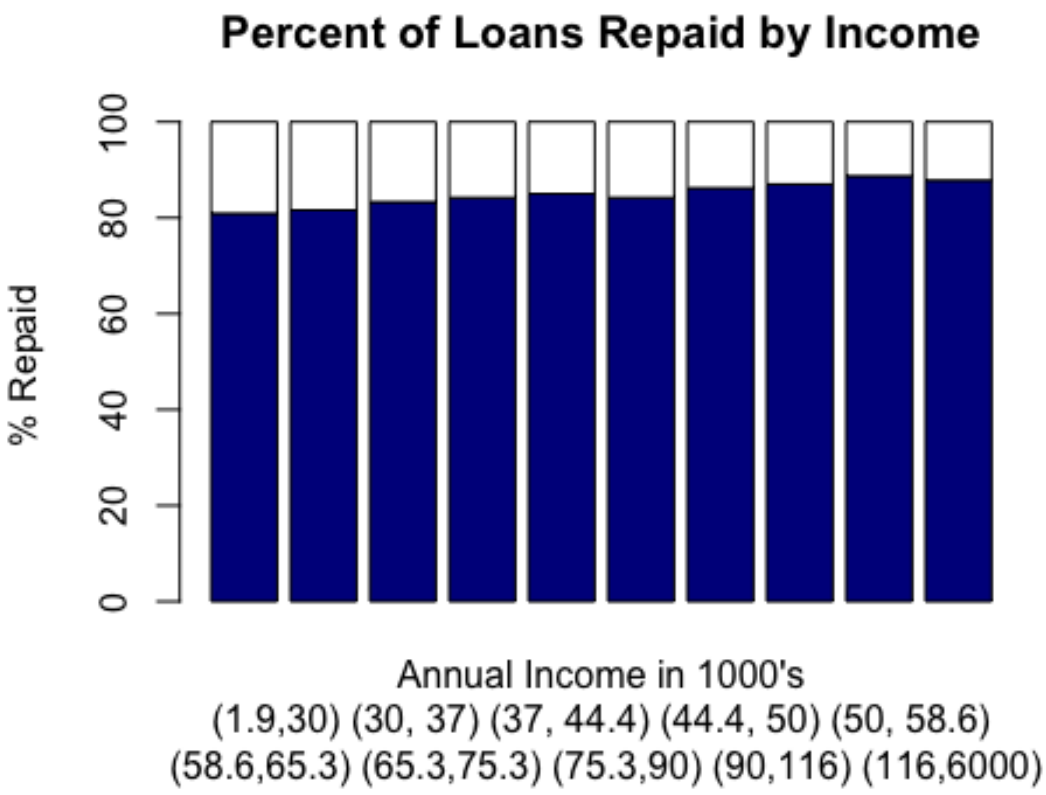

Percent of Loans Repaid by Loan Term
% Repaid
0
20
40
60
80
100
36 months
60 months


Percent of Loans Repaid by Employment Length
% Repaid
0
20
40
60
80
100
missing
3 years
6 years
9 years


Percent of Loans Repaid by Home Ownership Statu
% Repaid
0
20
40
60
80
100
MORTGAGE
OTHER
OWN
RENT



r, credit_card, debt_consolidation, educational, home_improvement,
_purchase, medical, moving, other, renewable_energy, small_busine



As is shown in the plots above, most reasonably suggest the assumption is met with their linear increases. For laon amount and total accounts clear curvature is present. Statistically speaking, there are probably ways to improve these variables through transformations. It is noted, however, that management has strongly advised against this methodology for the time being. It has been requested that the credit risk model remain as simple as possible for a variety of expected users, and the introduction of transformations is known to significantly change and complicate the interpretation of the effect of predictor variables – therefore, transformations will not be applied.

If the inadequate plots are being overlooked, the possession of wrong functional form in certain variables needs to be identified. Since some variables clearly do not fit the logistic regression requirements in their natural state, the process of model fitting may suggest that a certain loan characteristic is not a predictor of default simply because of the incorrect functional form. This phenomenon will not negatively affect the model, it may just label certain predictors as irrelevant when in fact if they were correctly represented, they may help explain the variation in loan defaults. This is a direct result of attempting to achieve management's request of achieving maximum statistical power while keeping the model as simple as possible for a variety of expected users.

Splitting up the Data

With the statistical assumptions addressed, the data needed to be split into two sections: 75% making up a training set and 25% as the test set. The point of this split is to save true but unseen data until the end of the modelling process to assess the predictive power of the model. The create partition function in R was used to generate a random split following these specifications. To keep consistency across the analysis, each data set contained 15% of defaults.

```
train.index <-
  createDataPartition(dataset$repay_fail,  p = .75, list = FALSE)
train <- dataset[train.index,]
train.full <-train # for putting predictors back in after they've been
removed from train subset
test <- dataset[-train.index,]
sum(train$repay_fail) / length(train$repay_fail)

## [1] 0.1521242

sum(test$repay_fail) / length(test$repay_fail)

## [1] 0.1493918
```

###Modelling Process and Statistical Measures Within the training set, a generalised linear binomial model with the logit link function was fit that included all 14 predictor variables. With the intention of removing variables that were not related to default or were not enhancing the model, an automated loop was created to remove and add variables one at a time – producing a platform to investigate the statistics associated with a variety of predictor combinations. To guide the model choice, the following statistical measures were adopted:

- the Gini coefficient. It is the risk-reward trade-off between a model's ability to correctly identify defaulters, and correctly identify the safe customers. It gives an output between 0 and 1, where 1 describes perfect model fit and 0 being completely inefficient – so optimizing the model with respect to Gini would give the risk management department an indication of the strength of the model and help assign their risk appetite.

- the difference of deviance test, which conducts a goodness of fit measure for the full model against possible reduced ones. This statistically rigorous tests the hypothesis that the reduced model is more appropriate than the full model – where the rejection of the null would indicate that statistical power is lost if that tested variable is removed. This component of the model building process has been applied to add a second level of support to what the Gini indicates.

- Aikike's Information Criterion was also used for data exploration, but other than suggesting that income as a factor was a better predictor than as a covariate, AIC always suggested the model with fewer predictors was better

The output of this automated loop recorded the Gini coefficients and difference of deviance test results in separate excel spreadsheets with the variable names. Similar functions were created for difference of deviance and AIC.

```
    # Gini removal function and call----

Gini.remove.predictor <-
  function(train,
           Gini.model.summary,
           Gini.model.summary.counter,
           Gini.model.pvalues.list) {
    Gini.train.model.check.df <-
      as.data.frame(matrix(nrow = dim(train)[2] - 1, ncol = 4))
    colnames(Gini.train.model.check.df) <-
      c("Dropped Predictor", "Gini", "AIC", "Column Index")

    # store the pvalues of each model so we can pick out the p-values of the
best model
    temp.pvalues.list <- list()
    for (i in 2:dim(train)[2]) {
      fit <-
        glm(paste("repay_fail~", (paste(
          names(train)[-c(1, i)], collapse = "+"
        ))),
        data = train,
        na.action = na.exclude)
      prob = predict(fit, type = c("response"))
      g <- roc(repay_fail ~ prob, data = train)
      Gini.train.model.check.df[i - 1, 1] <- names(train)[i]
      # calculate Gini
      Gini.train.model.check.df[i - 1, 2] <- 2 * g$auc - 1
      # AIC
      Gini.train.model.check.df[i - 1, 3] <- fit$aic
      # so we know the position in the list of columns of the predictor to
drop
      Gini.train.model.check.df[i - 1, 4] <- i
      temp.pvalues.list[[i]] <- summary(fit)$coefficients[, 4]
    }

    Gini.model.summary.counter <- Gini.model.summary.counter + 1

    # so largest Gini is at the top
    Gini.train.model.check.df.ascending <-
      Gini.train.model.check.df[order(Gini.train.model.check.df[, 2],
decreasing = TRUE),]

    ############################ copy results to Gini.model.summary
    Gini.biggest.index <- Gini.train.model.check.df.ascending[1, 4]
```

```
    # copy models predictors
    Gini.model.summary[Gini.model.summary.counter, 1] <-
      paste(names(train)[-c(1, Gini.biggest.index)], collapse = "+")
    # copy models Gini
    Gini.model.summary[Gini.model.summary.counter, 2] <-
      Gini.train.model.check.df.ascending[1, 2]
    # copy AIC
    Gini.model.summary[Gini.model.summary.counter, 3] <-
      Gini.train.model.check.df.ascending[1, 3]
    # copy that this is a predictor REDUCTION to summary
    Gini.model.summary[Gini.model.summary.counter, 4] <-
      "Reduction"
    # what predictor is being assessed
    Gini.model.summary[Gini.model.summary.counter, 5] <-
      Gini.train.model.check.df.ascending[1, 1]
    # add blank lines to the console window for easier navigation
    temp.list <-
      list(
        Gini.model.summary,
        Gini.model.summary.counter,
        Gini.biggest.index,
        Gini.train.model.check.df.ascending[1, 1],
        temp.pvalues.list[[Gini.biggest.index]]
      )
    return(temp.list)
  }
call the predictor removal function
temp <-
  Gini.remove.predictor(train, Gini.model.summary,
Gini.model.summary.counter)

# update variables outside the function environment
Gini.model.summary <- temp[[1]]
Gini.model.summary.counter <- temp[[2]]
biggest.Gini.index <- temp[[3]]
Gini.model.pvalues.list[[Gini.model.summary.counter]] <- temp[[5]]

# copy the removed predictors name to the removed.predictor vector so it can
be used later
Gini.removed.predictors[length(Gini.removed.predictors) + 1] <-
  names(train[, biggest.Gini.index])
# drop the predictor
train <- train[, -c(biggest.Gini.index)]
# check names of remaining predictors
names(train)[-1]
```

	Model	Gini	Reducing or adding a predictor?	Predictor being assessed
1	Full model	0.349	Nil	Nil
2	loan_amnt+term+home_ownership+purpose+dti+delinq_2yrs+inq_last_6mths+pub_rec+total_acc+fac.annual_inc	0.349	Reduction	time_since_first_loan
3	loan_amnt+term+purpose+dti+delinq_2yrs+inq_last_6mths+pub_rec+total_acc+fac.annual_inc	0.348	Reduction	home_ownership
4	loan_amnt+term+purpose+dti+delinq_2yrs+inq_last_6mths+pub_rec+fac.annual_inc	0.346	Reduction	total_acc
5	loan_amnt+term+purpose+dti+inq_last_6mths+pub_rec+fac.annual_inc	0.344	Reduction	delinq_2yrs
6	loan_amnt+term+purpose+inq_last_6mths+pub_rec+fac.annual_inc	0.341	Reduction	dti
7	loan_amnt+term+purpose+inq_last_6mths+pub_rec+fac.annual_inc + dti	0.344	Addition	dti
8	term+purpose+inq_last_6mths+pub_rec+fac.annual_inc	0.338	Reduction	loan_amnt
9	term+purpose+inq_last_6mths+fac.annual_inc	0.333	Reduction	pub_rec
10	term+purpose+inq_last_6mths	0.305	Reduction	fac.annual_inc
11	term+inq_last_6mths	0.266	Reduction	purpose
12	term	0.168	Reduction	inq_last_6mths
13	term + inq_last_6mths	0.266	Addition	inq_last_6mths

Gini based predictor selection

Gini Value and Difference of Deviance Output

As is evident on the excel spreadsheet output, there are 13 models, each with different combinations of predictor variables and their associated Gini values. To begin, the full 14 variable model is included and contains a Gini of 0.354. Interestingly, the Gini values do not significantly differ. After removing the chosen 10 variables from the automated loop, the Gini dropped ever so slightly to 0.339 for the model with just 4 terms.

At this stage, the results of the difference of deviance test were important to identify if some clarification about which model was statistically superior to others was provided. The reason more clarification was needed instead of just choosing the full model with the "best" Gini is because the difference between the best and worst of the first seven models is only 0.015 – therefore essentially the same level of accuracy is received from the full and complicated model as the much simpler model. Due to the requirement from management about creating the simplest model as possible, the aim is to achieve something of statistical merit with minimal components.

	Model	p-value	Predictor being assessed
2	loan_amnt+term+home_ownership+purpose+dti+delinq_2yrs+inq_last_6mths+pub_rec+total_acc+fac.annual_inc	0.879	time_since_first_loan
3	loan_amnt+term+purpose+dti+delinq_2yrs+inq_last_6mths+pub_rec+total_acc+fac.annual_inc	0.809	home_ownership
4	loan_amnt+term+purpose+dti+delinq_2yrs+inq_last_6mths+pub_rec+fac.annual_inc	0.181	total_acc
5	loan_amnt+term+purpose+dti+inq_last_6mths+pub_rec+fac.annual_inc	0.167	delinq_2yrs
6	loan_amnt+term+purpose+inq_last_6mths+pub_rec+fac.annual_inc	0.154	dti
7	term+purpose+inq_last_6mths+pub_rec+fac.annual_inc	0.050	loan_amnt
8	term+purpose+inq_last_6mths+fac.annual_inc	0.013	pub_rec
9	term+purpose+inq_last_6mths	0.007	fac.annual_inc
10	term+inq_last_6mths	0.004	purpose
11	term	0.000	inq_last_6mths

As is shown in the difference of deviance spreadsheet output, the results were very much in support of the four-term model. With a p-value on the border of significance of 0.115 for a five-term model, then a highly significant 0.0431 for the four-term model compared to each of the three-term models, there was no evidence to suggest that statistical power was improved with a smaller model. With these statistics taken into account, the four-term model was chosen. The following section includes further tests to assess the accuracy of this model and investigates the potential of occurrences such as under fitting.

###Variables in the Four-Term Model The model that is currently considered optimal includes:

- Loan term (categorical)
- Loan purpose (categorical)
- Number of inquiries the applicant has made in the last 6 months (integers)

- Annual income (categorical)

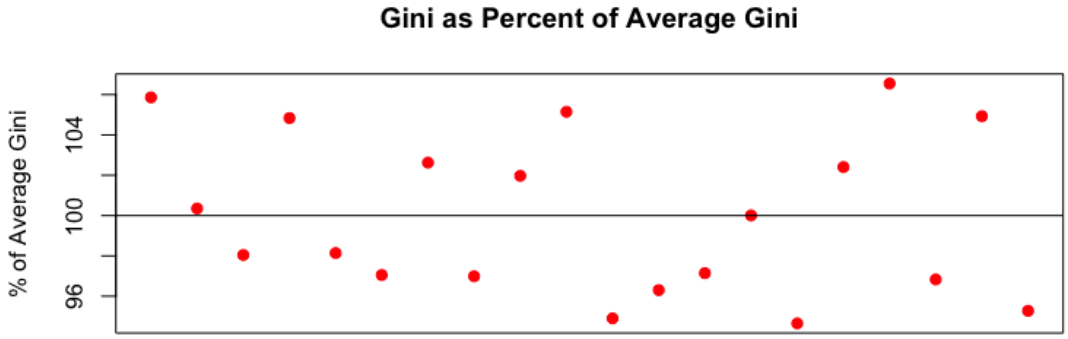
```
##
## Call:
## glm(formula = repay_fail ~ term + purpose + inq_last_6mths +
##     fac.annual_inc, family = "binomial", data = dataset)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7418  -0.5975  -0.4900  -0.3937   2.5138
##
## Coefficients:
##                                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)                      -2.328224   0.092241 -25.241  < 2e-16
## term60 months                     0.862395   0.030898  27.911  < 2e-16
## purposecredit_card                0.298687   0.096061   3.109 0.001875
## purposedebt_consolidation         0.520033   0.087199   5.964 2.47e-09
## purposeeducatiol                  0.867765   0.154172   5.629 1.82e-08
## purposehome_improvement           0.341654   0.101743   3.358 0.000785
## purposehouse                      0.409673   0.164927   2.484 0.012993
## purposemajor_purchase             0.076595   0.111767   0.685 0.493147
## purposemedical                    0.674035   0.134021   5.029 4.92e-07
## purposemoving                     0.614681   0.143807   4.274 1.92e-05
## purposeother                      0.642773   0.095072   6.761 1.37e-11
## purposerenewable_energy           0.780084   0.292942   2.663 0.007746
## purposesmall_business             1.248980   0.100606  12.415  < 2e-16
## purposevacation                   0.522978   0.173502   3.014 0.002576
## purposewedding                    0.155209   0.136927   1.134 0.256996
## inq_last_6mths                    0.175761   0.008489  20.705  < 2e-16
## fac.annual_inc(3e+04,3.7e+04]    -0.085669   0.061268  -1.398 0.162038
## fac.annual_inc(3.7e+04,4.44e+04] -0.235054   0.058866  -3.993 6.52e-05
## fac.annual_inc(4.44e+04,5e+04]   -0.319442   0.059690  -5.352 8.71e-08
## fac.annual_inc(5e+04,5.86e+04]   -0.379741   0.060391  -6.288 3.21e-10
## fac.annual_inc(5.86e+04,6.53e+04] -0.345283   0.059792  -5.775 7.71e-09
## fac.annual_inc(6.53e+04,7.53e+04] -0.516096   0.061811  -8.350  < 2e-16
## fac.annual_inc(7.53e+04,9e+04]   -0.600087   0.061906  -9.694  < 2e-16
## fac.annual_inc(9e+04,1.16e+05]   -0.788046   0.066431 -11.863  < 2e-16
## fac.annual_inc(1.16e+05,6e+06]   -0.711142   0.064206 -11.076  < 2e-16
##
## (Intercept)                      ***
## term60 months                    ***
## purposecredit_card               **
## purposedebt_consolidation        ***
## purposeeducatiol                 ***
## purposehome_improvement          ***
## purposehouse                     *
## purposemajor_purchase
## purposemedical                   ***
## purposemoving                    ***
## purposeother                     ***
## purposerenewable_energy          **
```

```
## purposesmall_business              ***
## purposevacation                    **
## purposewedding
## inq_last_6mths                     ***
## fac.annual_inc(3e+04,3.7e+04]
## fac.annual_inc(3.7e+04,4.44e+04]   ***
## fac.annual_inc(4.44e+04,5e+04]     ***
## fac.annual_inc(5e+04,5.86e+04]     ***
## fac.annual_inc(5.86e+04,6.53e+04]  ***
## fac.annual_inc(6.53e+04,7.53e+04]  ***
## fac.annual_inc(7.53e+04,9e+04]     ***
## fac.annual_inc(9e+04,1.16e+05]     ***
## fac.annual_inc(1.16e+05,6e+06]     ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 32721  on 38475  degrees of freedom
## Residual deviance: 31046  on 38451  degrees of freedom
##   (1 observation deleted due to missingness)
## AIC: 31096
##
## Number of Fisher Scoring iterations: 5
```

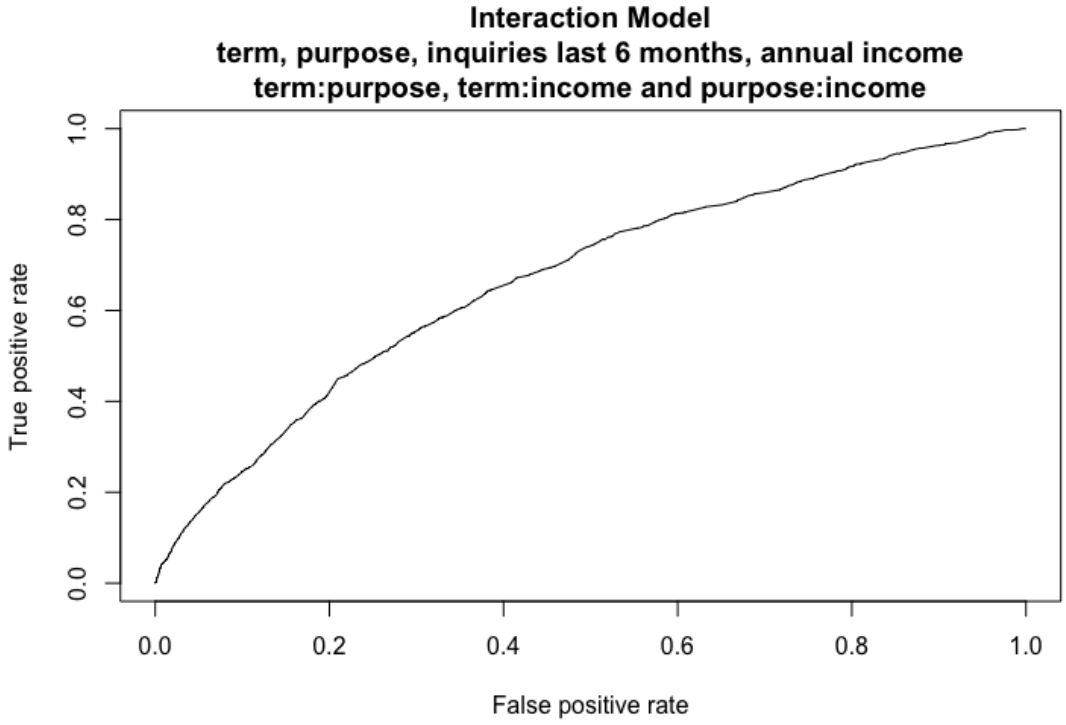
Industry experts may wonder why seemingly important variables such as loan amount or debt to income ratio are not included in this model – however this can simply be addressed by reiterating that the analysis conducted concluded that keeping these variables does not provide a stronger foresight into an applicants' likelihood of loan default – so there is no benefit in using them, and there is potential for a big downside of overfitting the model which would result in large variation between individuals with very similar application characteristics.

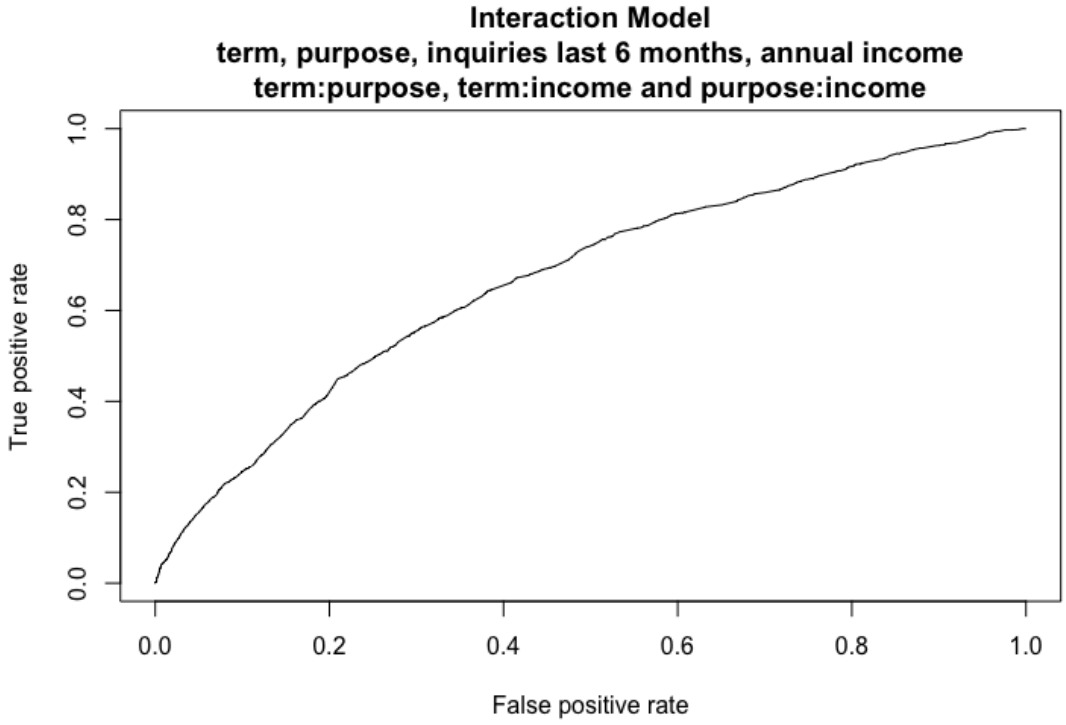
#Addressing Management's Inquiries: Justifications about the Statistical Rigor of the Model
###Investigating Under-fitting With a "simple" model chosen, the potential for under-fitting remains prominent – which would imply that key features of the underlying ideal relationships could be missed. To assess this need for repeatability, the data was further split into training and testing subsets twenty times. The four term model was fit for each of these subsets and their corresponding Gini values were calculated. The plot shown below displays the red horizontal line as the average Gini for the 20 versions of the four term model (denoted as 100%), and each of the dots display the Gini values relative to the 100% average. As the points vary from approximately 93 to 105%, it can be concluded that the

consistency suggests the model choice is not overly simplistic or biased.



#Assessing Interactions Although variables are not being transformed to correct for incorrect functional forms, interactions between certain variables can be assessed. To keep the model to a reasonable scope, only two-way interactions are investigated. The two-way interactions were first assessed on their own, to simply provide an indication as to whether there was any reason to believe default was influenced by combinations of predictors. As is evident, many had significant Gini scores in a model consisting only of the interactions. Therefore, interactions were further investigated by adding term-to-purpose, term-to-income and purpose-to-income them to the model. In assessing the quality of the fit with these interactions against the basic model, the new Gini value was calculated. As is shown, the Gini was in fact lowered by around 0.005 when the interactions were introduced. As a result, keeping in mind the need to find an appropriate medium between statistical complexity and interpretability, the simple model was kept and interactions were disregarded.





###4-Fold Cross Validation Adding a third metric to the Gini and Difference of deviance measures, 4-fold cross validation was applied to the entire training data set. The training data (75% of the entire dataset) was split into four subsets of equal size and each maintaining a 15% default rate. Three subsets were combined to estimate the model parameters, then used to predict the remaining subset – with the process being completed

a total of four times. The new Gini scores were calculated for each of the four stages of the process and the average was calculated. The results confirmed the earlier conclusions. Starting with a larger model and removing variables one-by-one down to the four previously defined had very little reduction in Gini, but reducing from four to three there was a big drop off. In trying to find the simplest model for a given predictive power, the four-term model is again proven superior.

```
##
V1
## 1
<NA>
## 2
<NA>
## 3
<NA>
## 4
<NA>
## 5
loan_amnt+term+emp_length+purpose+delinq_2yrs+inq_last_6mths+pub_rec+fac.annu
al_inc
## 6
loan_amnt+term+emp_length+purpose+inq_last_6mths+pub_rec+fac.annual_inc
## 7
<NA>
## 8
loan_amnt+term+purpose+inq_last_6mths+pub_rec+fac.annual_inc
## 9
term+purpose+inq_last_6mths+pub_rec+fac.annual_inc
## 10
term+purpose+inq_last_6mths+fac.annual_inc
## 11
term+inq_last_6mths+fac.annual_inc
## 12
term+purpose+inq_last_6mths+fac.annual_inc+term:purpose+term:fac.annual_inc
## 13
<NA>
##          V6
## 1        NA
## 2        NA
## 3        NA
## 4        NA
## 5  0.338400
## 6  0.335300
## 7        NA
## 8  0.334775
## 9  0.331750
## 10 0.325250
## 11 0.297525
## 12 0.328800
## 13       NA
```

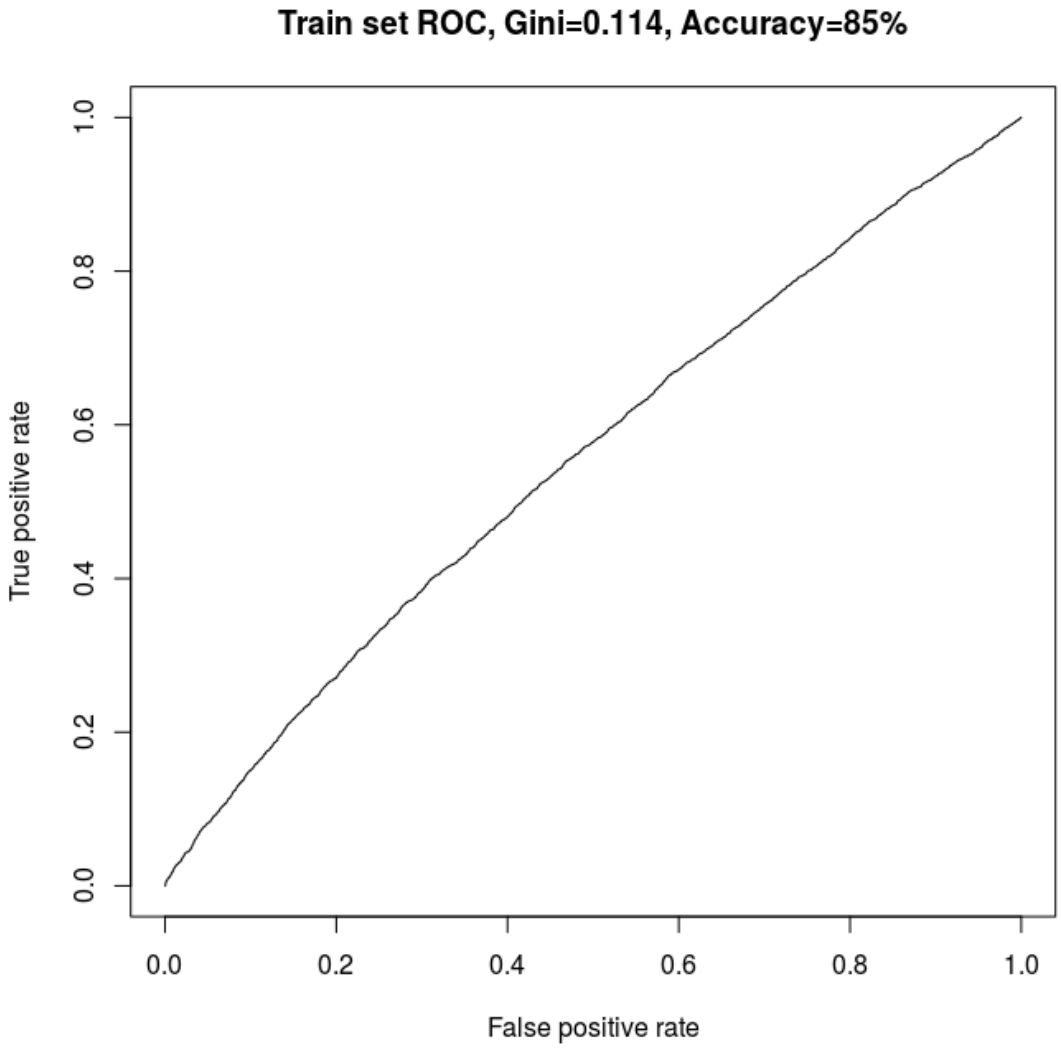
###Test Set Predictions With the four term model confidently chose, test set validation could be undertaken. The data was split into 10 training and testing folds, again calculating a Gini value to quantify the accuracy. Similar to the training data results, the Gini scores had a tipping point below four inputs, reinforcing that the four-term model performs slightly less well than larger models, and much better than smaller ones.

```
##
Model
## 1
loan_amnt+term+emp_length+purpose+dti+inq_last_6mths+pub_rec+fac.annual_inc
## 2
term+emp_length+purpose+dti+inq_last_6mths+pub_rec+fac.annual_inc
## 3
term+purpose+dti+inq_last_6mths+pub_rec+fac.annual_inc
## 4
term+purpose+inq_last_6mths+pub_rec+fac.annual_inc
## 5
term+purpose+inq_last_6mths+fac.annual_inc
## 6
term+inq_last_6mths+fac.annual_inc
## 7
term+purpose+inq_last_6mths+fac.annual_inc+term:purpose+term:fac.annual_inc
## 8
<NA>
##   Average Gini
## 1    0.3055555
## 2    0.3032919
## 3    0.3026011
## 4    0.3007538
## 5    0.2955581
## 6    0.2748252
## 7    0.2957024
## 8           NA
```

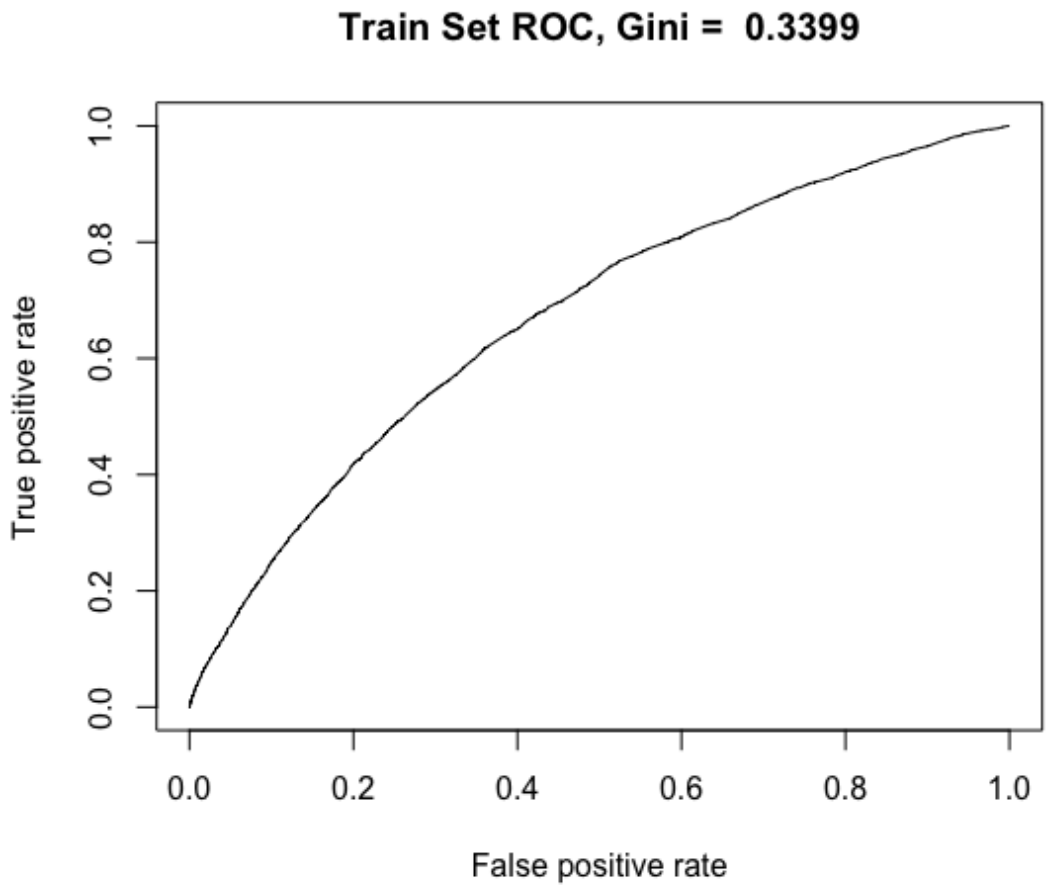
#Addressing Management's Inquiries: Comparison to the Benchmark ###ROC Curves To understand the Gini value visually, the receiver operating characteristic, or ROC curve, can be considered. This is a way to summarise the ability of a binary classifier, where a Gini of zero, would sit on the one-to-one diagonal, and a perfectly predictive model with a Gini of 1 would sit curved in the top left corner of the graph, like a right-angled triangle. This curve helps answer the second managerial query of 'how does this model perform compared to the previous one, and how can it be expected to perform in the future?'.

The Gini values and ROC curves based on the training data for the previous model are shown on the left, and the new model on the right. In the old model, it can be seen that the "curve" is close to a straight line, with a Gini of about 0.11. Conversely, the new one is visibly curving with a Gini of 0.33. In terms of the test set information, the old model is again at 0.11 and new at substantially better at 0.32 and more curved. Therefore, it is evident that the previous model has been significantly improved on by increasing the correct prediction of bad loans, and in reducing the incorrect predictions about good loans.

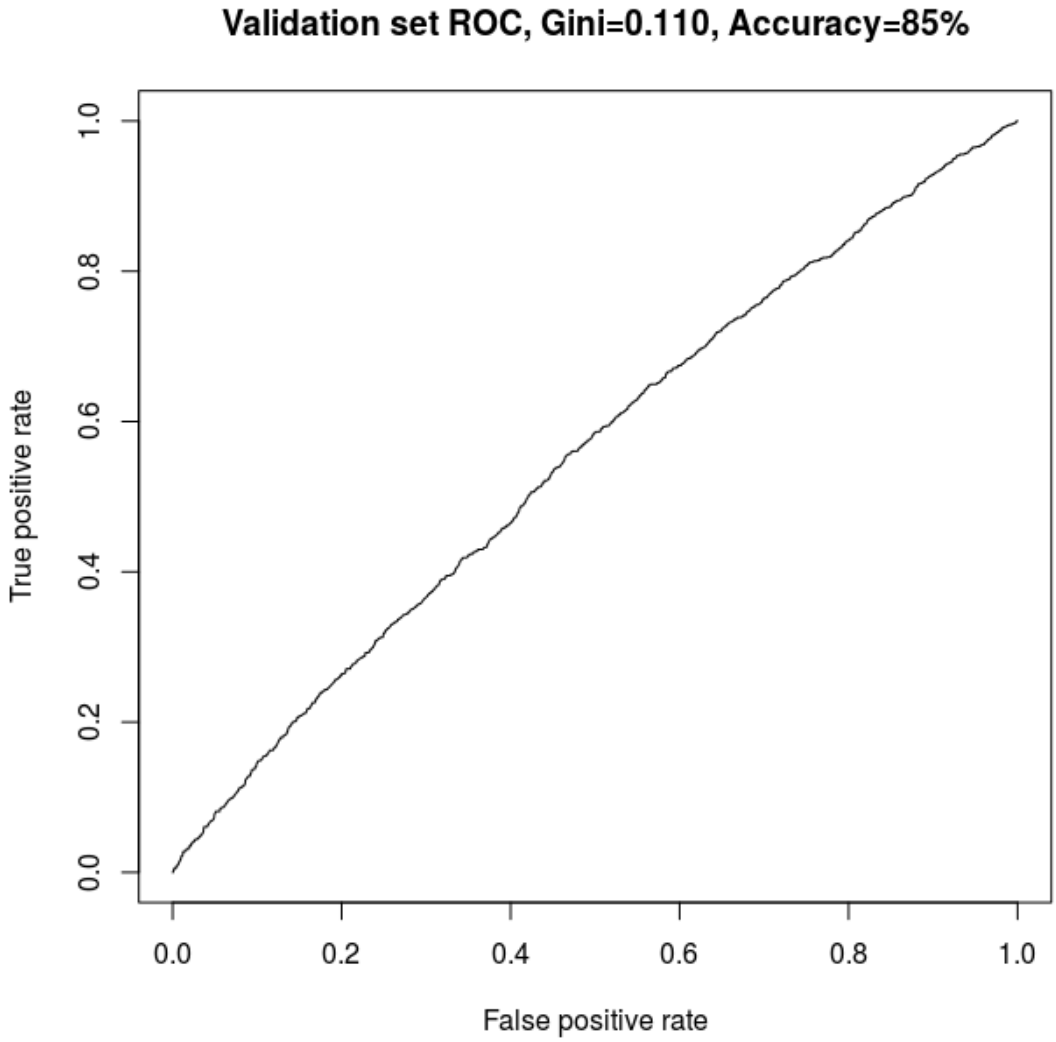
statistical values indicate our model is more successful than the previous in assessing potential bank customers, so it is a much sturdier platform on which to build a personal loan business.



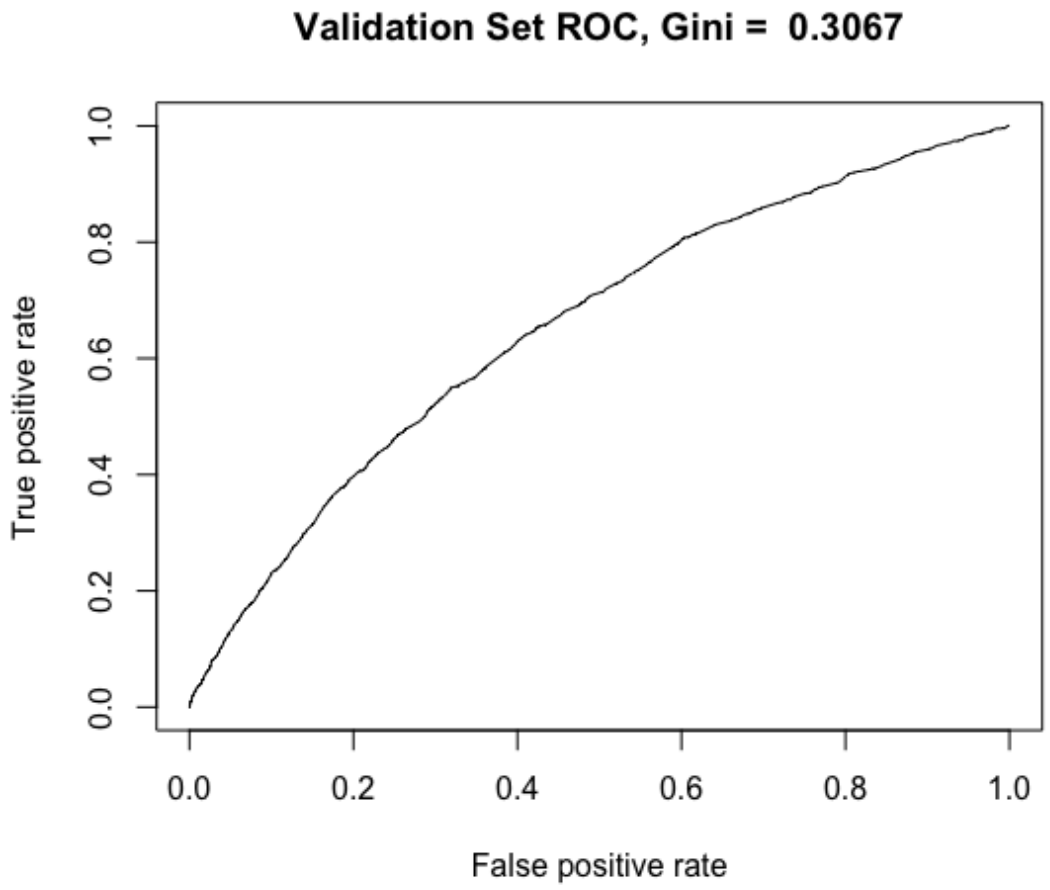
Old Training Set Roc Plot



New Training Set Roc Plot



Old Validation Set Roc Plot



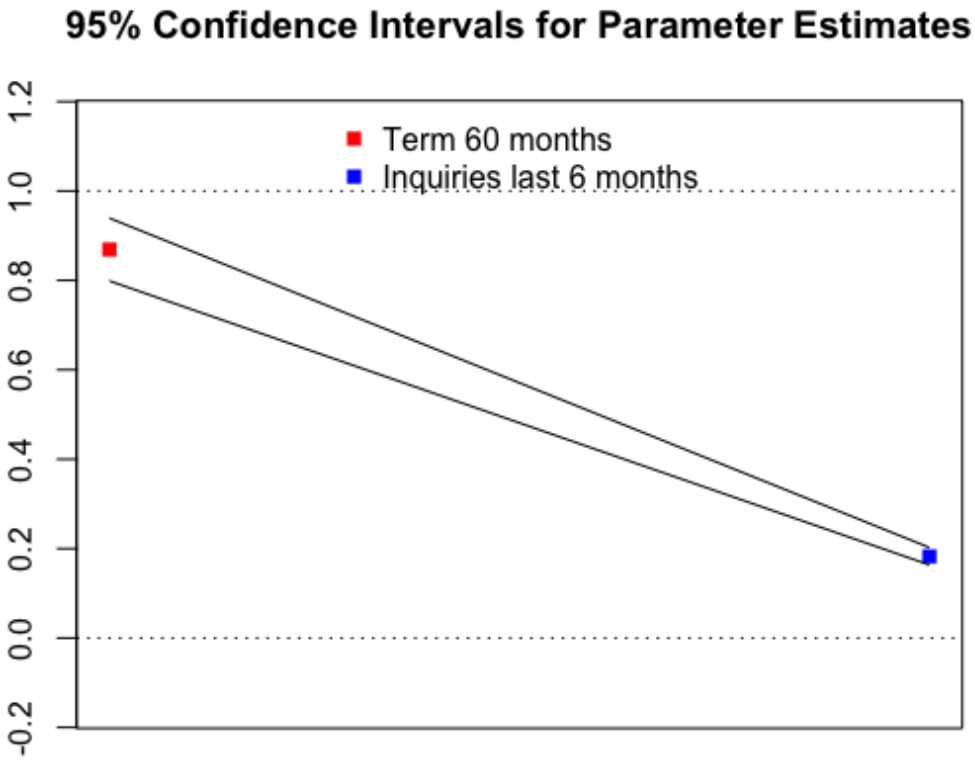
New Validation Set Roc Plot

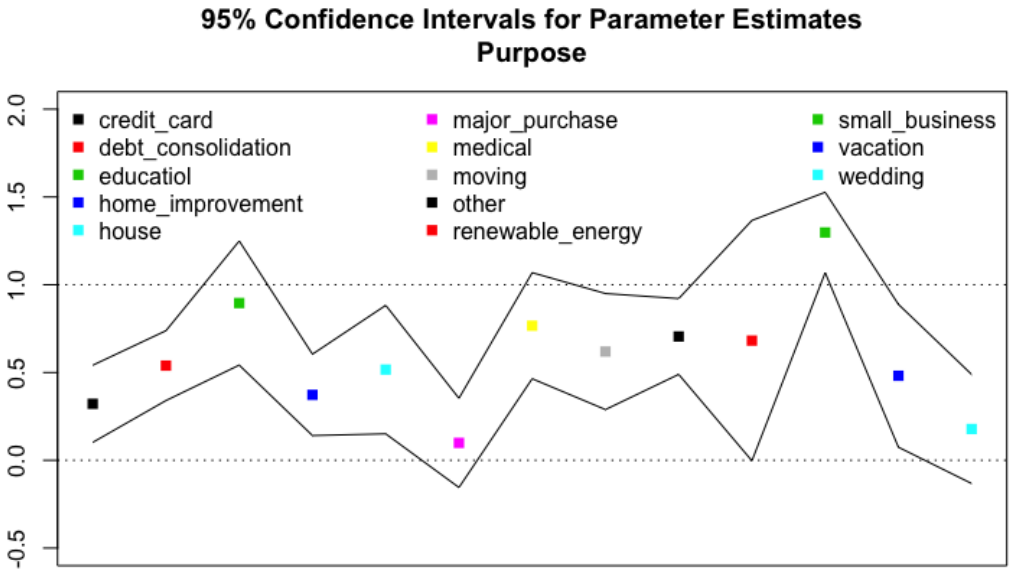
It is noted here that a confusion matrix displaying the true positive and true negative rates has not been calculated. Creating such a matrix requires a threshold be specified indicating which probability level acts as the cut off between classifying a customer's loan as "good" or "bad". Undertaking this process requires expert judgement and is an internal requirement based on what risk appetite the bank is willing to accept at a certain point in time. As a result, management requested the model's be compared solely on their Gini values and ROC plots, which would provide substantial information to identify which model is superior.

#Addressing Management's Inquiries: Important Variables and their Effects ###Effect Estimates With the model justified, the specifics of the important variables can now be addressed. The horizontal dotted lines on the plot below show the values of zero towards the bottom, and 1 above. On the left is the effect of a longer loan term, which is positive, and on the right is the effect of each extra recent inquiry by the applicant, which is also positive, so they each increase the likelihood of a default. The solid black lines show the upper and lower bounds of a 95% confidence interval for the parameter estimates. That they sit closely, relative to the absolute value, to each estimate indicates that the effect of these variables is sharply calculated.

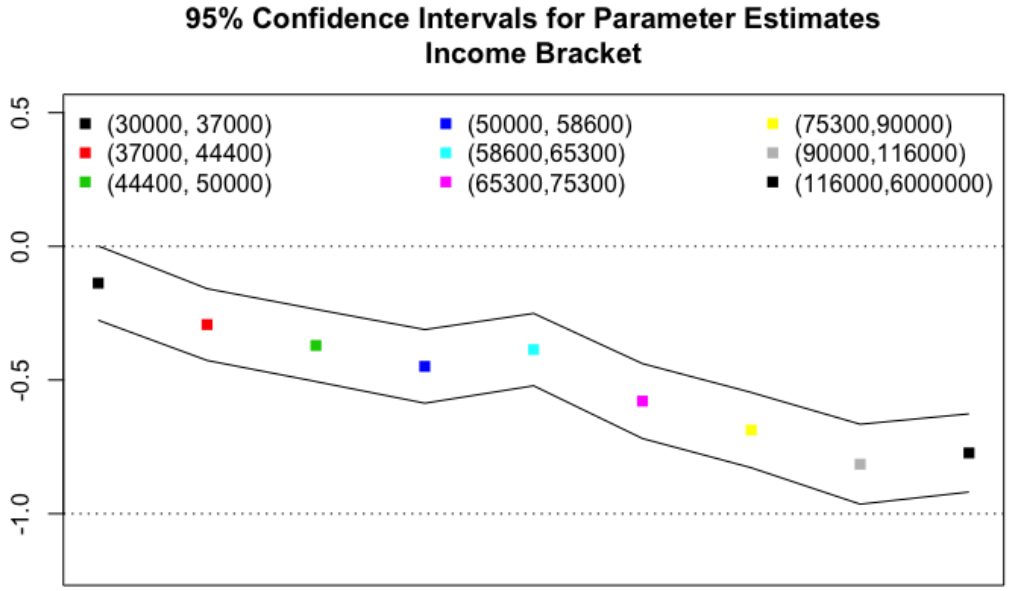
This plot shows the effect of a loans purpose on default likelihood, again with zero and one as the horizontal dashes. Other than car, wedding and 'major purchase', each type of purpose increases the probability of a bad loan. The confidence intervals on these estimated effects are much larger than before. In saying this however, removing purpose from the model substantially decreased the Gini score – so while the confidence intervals allow for large variation in estimates, there is no reason to question the inclusion of purpose in the model. The Cis could be improved by either the gathering of more data, or perhaps by grouping some of the 14 purposes into like categories (under industry expert's discretion).

The final plot displays annual income, with the smallest grouping of income level on the left, to largest on the bottom right. This time, the horizontal dashes are for -1 below and zero above, so as expected, increasing income decreases the chance of loan default.





Base level: Car



Base level: (1900,30000)

```
##       Predictor             Lower Bound Estimate  Upper Bound
##  [1,] "Intercept"           "-2.51"     "-2.33"   "-2.15"
##  [2,] "Term 60 months"      "0.802"     "0.862"   "0.923"
##  [3,] "credit_card"         "0.11"      "0.299"   "0.487"
##  [4,] "debt_consolidation" "0.349"     "0.52"    "0.691"
##  [5,] "educatiol"           "0.566"     "0.868"   "1.17"
```

```
##  [6,] "home_improvement"   "0.142"     "0.342"    "0.541"
##  [7,] "house"              "0.0864"    "0.41"     "0.733"
##  [8,] "major_purchase"     "-0.142"    "0.0766"   "0.296"
##  [9,] "medical"            "0.411"     "0.674"    "0.937"
## [10,] "moving"             "0.333"     "0.615"    "0.897"
## [11,] "other"              "0.456"     "0.643"    "0.829"
## [12,] "renewable_energy"   "0.206"     "0.78"     "1.35"
## [13,] "small_business"     "1.05"      "1.25"     "1.45"
## [14,] "vacation"           "0.183"     "0.523"    "0.863"
## [15,] "wedding"            "-0.113"    "0.155"    "0.424"
## [16,] "inq_last_6mths"     "0.159"     "0.176"    "0.192"
## [17,] "30000, 37000"       "-0.206"    "-0.0857"  "0.0344"
## [18,] "37000, 44400"       "-0.35"     "-0.235"   "-0.12"
## [19,] "44400, 50000"       "-0.436"    "-0.319"   "-0.202"
## [20,] "50000, 58600"       "-0.498"    "-0.38"    "-0.261"
## [21,] "58600,65300"        "-0.462"    "-0.345"   "-0.228"
## [22,] "65300,75300"        "-0.637"    "-0.516"   "-0.395"
## [23,] "75300,90000"        "-0.721"    "-0.6"     "-0.479"
## [24,] "90000,116000"       "-0.918"    "-0.788"   "-0.658"
## [25,] "116000,6000000"     "-0.837"    "-0.711"   "-0.585"
```

###Summary of effect estimates

Variable	Effect on Probability of Default
Loan term	Longer loan = increase
Number of inquiries in last six months	More inquiries = increase
Loan Purpose	Car/wedding/major purchase = neutral All others = increase
Annual income	Larger income = decrease

###Summary of answers to management's questions

```
1.How does this model perform compared to the one you used previously? How
can it be expected to perform on new loans?
```

As was shown in the Gini values and ROC plots, the previous model has been substantially improved upon. With a previous Gini of approximately 0.11 achieved in both the training and test set measures, the old model appeared to have a small amount of predictive power and large amounts of variation in defaults were left unexplained. In the new, four-term model, the Gini was significantly increased to approximately 0.32, and the ROC plot reflected a much more prominent curve. This occurrence proves the new model explains more variation in defaults, and the combination of the loan characteristics included are used to provide estimated of "good" and "bad" loans that are largely more accurate than the given benchmark.

2. What are the important variables in this model and how do they compare to variables the bank has found to be traditionally important in its own modelling?

The model chosen contained four characteristics of the given loan, including:

- Loan Term
  - 36 months, 60 months
- Loan purpose
  - Credit card, debt consolidation, major purchase, small business, wedding, medical, car, home improvement, vacation, educational, house, other
- Annual income
  - (1900,30000), (30000, 37000), (37000, 44400), (44400, 50000), (50000, 58600), (58600,65300), (65300,75300), (75300,90000), (90000,116000), (116000,6000000)
- Number of inquiries the applicant has made in the last six months

While there are seemingly important characteristics such as loan amount or debt to income ratio absent from this model, it is reiterated that the statistical methodology applied in this context formally concluded that the inclusion of any more terms did not increase the predictive ability of the model or explain any more variation in defaults – therefore no benefit was gained by including them.

3. What assurances and justifications can you make about the statistical rigor of your model and modelling methodology?

The justifications and assurances made about the statistical rigor of the chosen model arise from the processes of investigating Gini values, difference of deviance tests, ROC curves, repeatability test, cross validation techniques and predicting the unseen test set data. Throughout these procedures, the four-term model was consistently proven to be statistically superior to other models containing different combinations of the loan characteristics. The model chosen was found to be statistically simple, yet was not under fit, and the predictive power was extremely good in relation to other models and the previous benchmark.

#Recommendations, strengths and weaknesses

In evaluating the methods used to create a new credit risk model for personal loans provided in a banking environment, there are numerous strengths and weaknesses. In terms of strengths, generalised linear models were appropriately applied, and the logistic regression was well suited to the loan default situation. The model chosen provides an output between zero and one that denotes the probability that a new applicant will default based on their input data relevant to the model characteristics. A good medium was found between creating a model that had strong statistical merit, but also remained simple and interpretable for a variety of expected users within a banking environment. The output also gives a bank the opportunity to specify their risk appetite and thus redefine the threshold they wish to use to classify the difference between “good” and “bad” loans at any point in time.

Potential weaknesses of this process include the loss of information and accuracy as a result of not applying transformations to variables that did not fit the functional form of the logistic regression. This main downfall perhaps lead to variation in defaults remaining unexplained that could have been captured if the variables were represented correctly.

As a result, the following recommendations for further work on this credit risk model are made:

1. Investigate transformations and identify if their presence enhances the accuracy and predictive power of the models. Quantify this success through the Gini values, ROC curves and cross validation.
2. Investigate interactions further to assess if loan characteristics interact in their effect on default. This may become more effective once variables have undergone transformations.
3. Gain industry advice on how to better represent characteristics of the loan on new applications – for example, how to better group "purpose" to make less than the current large selection of 14 categories
4. Investigate other loan characteristics that have been overlooked in this situation – potentially including age of the applicant and more information surrounding their past banking habits

As with any mathematical representation of a real-life scenario, it is important to identify that the human factor associated with loans holds a certain level of uncertainty. As such, it is encouraged that the model be used as a guide, and the possibility of unpredictable actions may occur for any new applicant.