LOAN PAYMENT

MAT-8406 Regression Methods

Submitted to Dr. Zhang

***-Team GLM*-Adhiraj Roka, Haichuan Du, Quentin Williams**

TABLE OF CONTENTS

INTRODUCTION

Loan is the lending of money by one or more entities to other individuals, organizations etc. where the recipient (i.e., the borrower) incurs a debt and is usually liable to pay the principal amount along with interest. In our project, using the loan dataset from *Kaggle*, we are *primarily* trying to model how many days an individual has to pay back a loan because it can give valuable information on how quickly or how long it would take a person to finish paying off a loan. We will be using a general linear model to achieve this goal as such information can be used to better utilize terms for quicker liquidity turn around.

Secondarily, we are trying to see if we can model whether a loan would be paid by the due date because this information can be used to predict which individuals will default on loans. If we are able to model the likelihood of a person paying off their loan, companies or other entities can use this information to predict which individuals will default on loans, allowing them to eventually minimize risk. We will be using a type of generalized linear model: logistic regression to find our model here.

DATA PREPARATION

TABLE 2.1: DESCRIPTION OF DATA

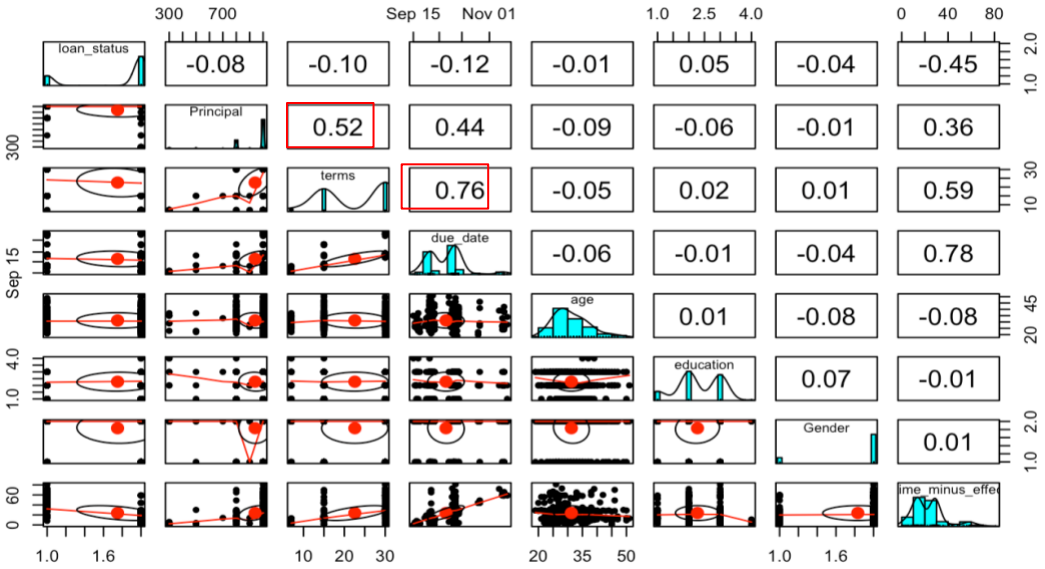
Variables	**Description**
Loan_ID	unique loan number assigned to each loan customers
Loan_status	loan is *paid off*, in *collection*, new customer yet to payoff, or paid off after the collection efforts (*Paid off/ Collection Paid off/ Collection)*
Principal	Basic principal loan amount at the origination
Terms	weekly (7 days), biweekly, and monthly payoff schedule (in days)
Effective_date	When the loan got originated and took effects
Due_Date	When the loan needs to be paid off

Paidoff_time	The actual time a customer pays off the loan
Pastdue_days	How many days a loan has been past due
Age, education, gender	A customer's basic demographic information

For our first model, our response variables will be **Paid off time - Effective Date** (*from here on out we will call it the loan length for interpretations*). We will subset only **collection paid off** and **paid off** as we're modeling how many days an individual has to pay back a loan otherwise the number of days is indefinite. The education levels were fixed: "*Bechalors* "was fixed to Bachelors. Furthermore, *Pastdue_days* NA's were there for those who already paid before due date and were converted to 0.

MODEL FORMULATION

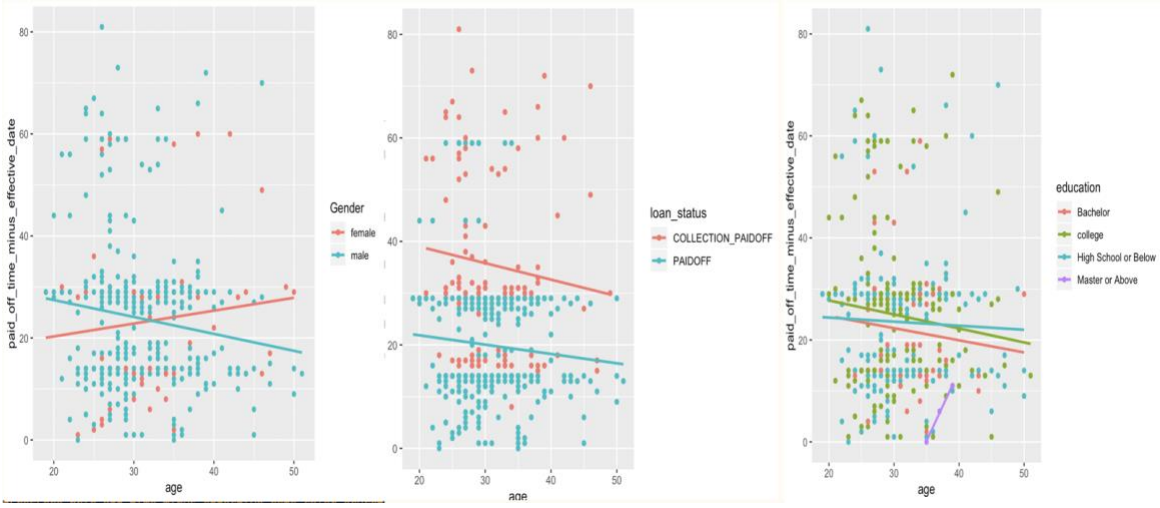
The Pastdue_days will be removed from the model because it provides essentially the same information as our response variable [Paid off time - Effective Date**]**. The former is zero when the loan is paid off before due date, whereas in the same situation the response variable still records how many days since the effective date until paid off; as a result, it is more informative.



Looking at the pairs panel plot from R above, we can see that terms and due date have a moderate correlation with our response variable. There could also potentially be some multicollinearity issue between principal and terms (r=0.52), and terms and due date (r= 0.76). After checking for VIFs, we found out that all of the values were less than 5 and concluded that there was no serious issue for multicollinearity.

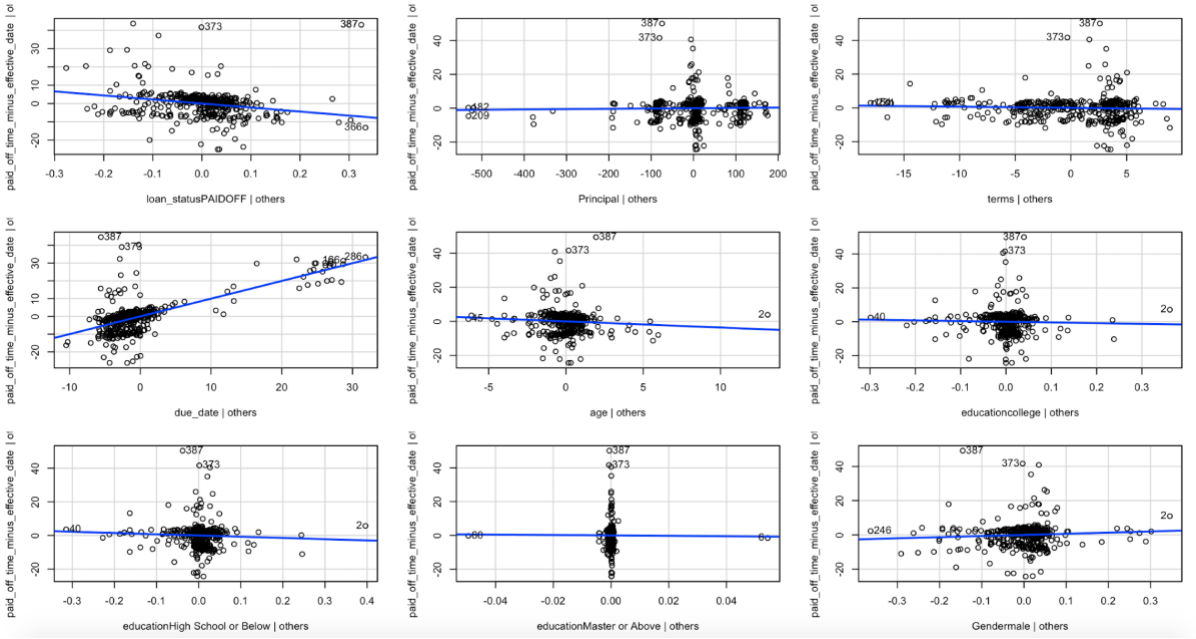
To initialize our model, we potentially tested for interaction and higher order terms. After testing for possible interaction terms, we found out that there could be potential interaction between age and gender, age and loan status and age and education by looking at the graphs (*see figure 2.3*).

Figure 2.3 Interaction plots



We also looked at partial regression plots (*see figure 2.4*) to see if our model would need any higher order terms; however, this was not the case. There weren't any serious issues but there is indication for potential outliers or influential observations.

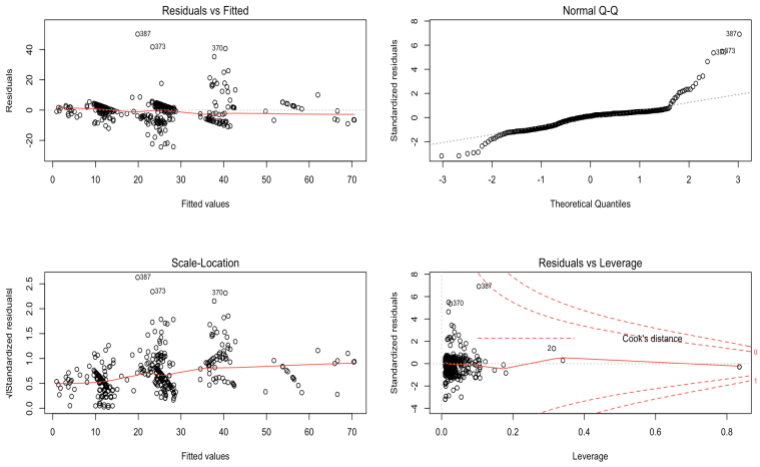
Figure 2.4 Partial Regression plots



From the initialized model, we can see that loan_status, due date and interaction of loan_status and age were pretty significant in the model. Also, the model has overall significance. Moreover, *from output 2.5* below, after checking for model diagnostics we can see that the constant variance assumption, linearity seems reasonable. There are short *bands* in the graph of residual vs. fitted which indicates replication in our data. However, checking for normality, q-q plot shows that normality assumption has been violated. Hence, we decided to use a *Box-Cox transformation* on our response as its helps with not only solving non-normality issues but also stabilizes the variance. Since, the lambda value was 0.58, we used a *sqrt* transformation on our response.

Output 2.5 Summary/ Diagnostics output from *intitalized* model

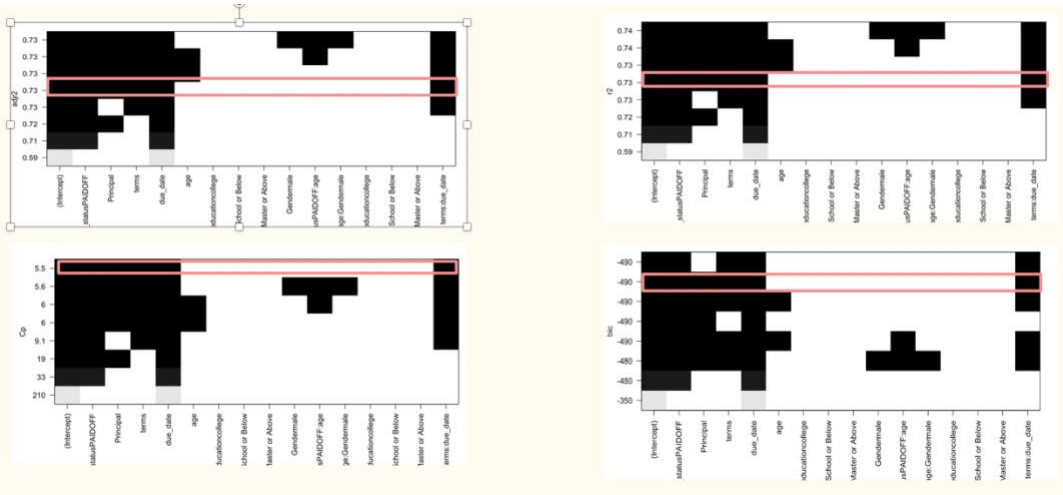
```
Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                   -1.699e+04  9.275e+02 -18.322  < 2e-16 ***
loan_statusPAIDOFF            -2.201e+01  4.705e+00  -4.679 4.01e-06 ***
Principal                      1.937e-03  4.006e-03   0.483   0.6290
terms                         -7.027e-02  7.920e-02  -0.887   0.3755
due_date                       9.976e-01  5.444e-02  18.326  < 2e-16 ***
age                           -3.626e-01  2.559e-01  -1.417   0.1573
educationcollege              -4.171e+00  7.524e+00  -0.554   0.5796
educationHigh School or Below -7.374e+00  7.343e+00  -1.004   0.3159
educationMaster or Above      -1.272e+01  1.024e+02  -0.124   0.9012
Gendermale                     6.680e+00  5.146e+00   1.298   0.1950
loan_statusPAIDOFF:age         3.063e-01  1.481e-01   2.069   0.0393 *
age:Gendermale                -1.897e-01  1.577e-01  -1.203   0.2298
age:educationcollege           1.253e-01  2.319e-01   0.540   0.5893
age:educationHigh School or Below 2.398e-01  2.241e-01   1.070   0.2853
age:educationMaster or Above   4.005e-01  2.760e+00   0.145   0.8847
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.667 on 385 degrees of freedom
Multiple R-squared:  0.7436,    Adjusted R-squared:  0.7343
F-statistic: 79.75 on 14 and 385 DF,  p-value: < 2.2e-16
```



VARIABLE SELECTION

Using manual backward elimination, we removed each variable form the model that had the highest p-values and were not significant (p-values >0.05) while maintaining model hierarchy. We ended up removing Age: education, education, Age: Gender, Gender, Age: Loan status, Age in that order until all the variables were significant in the model. We also used *regsubsets* () function in R to find out the best model from all different combination using measuring statistics like Cp, Bic, R^2 and R^2_{adj}. (*see figure 3.1*)

Figure 3.1 Regsubsets Plot



From the figure above, the model with predictors loan_status, principal, terms, due date and terms: due date has the lowest Cp, Bic and one of the highest R_2 and R_{2adj} values. We can also see similar results from table 3.2 where the same model has comparatively lower PRESS, AIC and MSE statistical values. Hence, we decided this to be our final model.

Table 3.2 Table statistics for different model

L-loan status, P-Principal, T-terms, D-Due Date, A-age

Models	MSE	R^2	R^2adj	PRESS	AIC	BIC
L,P,T,D,A,L:A,T:D	0.629	0.736	0.732	258.17	-177.68	-143.96
L,P,T,D,A,T:D	0.630	0.735	0.731	256.86	-177.6	-148.1
L,P,T,D,T:D	0.630	0.734	0.730	255.94	-178.14	-152.85
L, T, D,T:D	0.64	0.730	0.727	258.29	-174.42	-153.34
L, T, D	0.66	0.718	0.717	268.23	-159.91	-143.05

MODEL DIAGNOSTICS

As previously stated, there is no serious violations with constant variance and linearity assumptions for our model. There is also no reason to believe non-independence in our model. However, non-normality still exists. From *Shapiro-Francia* normality test, we found enough evidence to conclude that our model error term was not normal. We decided to check for model lack of fit to see if our model was not adequate enough. Conversely, our p-value was high enough to conclude that our model had reasonable/adequate fit.

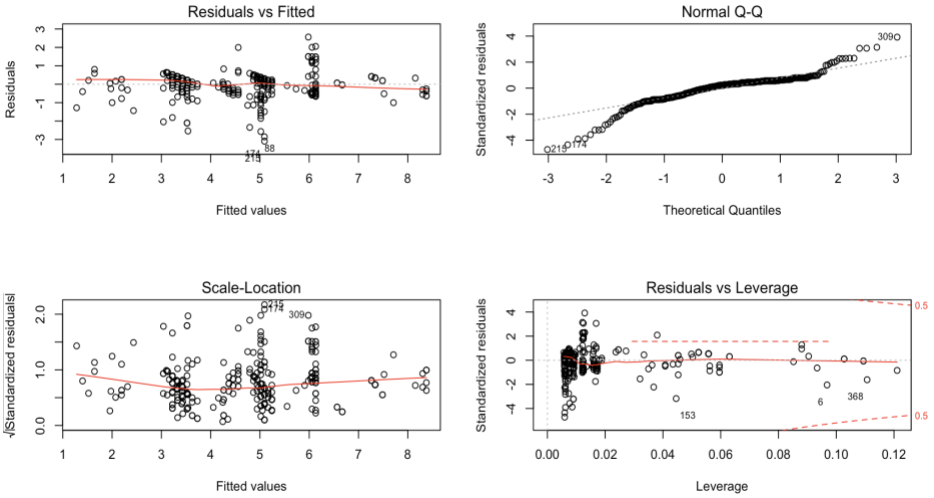
Further checking for residuals diagnostics, using measures like *Dfbetas*, *DfFits*, *leverage* (see Table 4.1) we found a cluster of observations that were influential. Removing these observations individually starting from observation 387, ID: xqd20160487, we can see that normality assumption is still violated from q-q plot (see figure 4.2). Taking a closer look at some of these observations, some of them are loans that are paid the same or the next day of taking out the loan. These could potentially be individuals who are solely trying to improve their credit score. Also, there are some individuals in the data who do not pay their loans for several days past due

date. However, there is not enough convincing evidence to subset these observations out of our analysis, and since our dataset is already small enough, doing so is not recommended.

Table 4.1 Residual Analysis (influential/outliers shown) [obs. no matches to *loan.new* dataset in R code]

						DfBetas					
obs.no.	|SRES|	LEVERAGE	COOK'S D	|DfFits|	(Intercept)	loan_statusPAIDOFF	Principal	terms	due_date	terms:due_date
1	2.89843058	0.00753861	0.0104	0.2526	0.0098	0.09143	0.002658	0.050129	0.009702	0.050005
6	1.6678903	0.08872481	0.0449	0.5204	0.201055	0.007221	0.389485	0.207255	0.200518	0.207311
46	1.25145737	0.03054939	0.0082	0.2222	0.16648	0.014046	0.022071	0.172007	0.166407	0.171962
64	5.13626732	0.00609174	0.0253	0.4021	0.00037	0.166575	0.008427	0.032635	0.000511	0.032471
81	2.26002407	0.00609174	0.0052	0.1769	0.000163	0.073295	0.003708	0.01436	0.000225	0.014288
82	0.96598903	0.04368679	0.0071	0.2065	0.14816	0.012613	0.114513	0.1516	0.148203	0.151541
88	3.18461177	0.00609174	0.0101	0.2493	0.000229	0.10328	0.005225	0.020235	0.000317	0.020133
136	5.13626732	0.00609174	0.0253	0.4021	0.00037	0.166575	0.008427	0.032635	0.000511	0.032471
148	1.06705057	0.08045695	0.0166	0.3156	0.075287	0.008978	0.258269	0.084183	0.074943	0.084206
153	2.64172936	0.04138167	0.0495	0.5489	0.129582	0.058031	0.510927	0.132806	0.130138	0.132729
174	3.56195922	0.00585335	0.0121	0.2733	0.004353	0.116584	0.006703	0.009655	0.004443	0.009554
193	2.66475732	0.00877739	0.0103	0.2508	0.00136	0.06906	0.152674	0.010351	0.001238	0.010436
209	2.15843056	0.08045695	0.0673	0.6385	0.15229	0.018161	0.522426	0.170286	0.151595	0.170332
215	3.87545884	0.00585335	0.0142	0.2974	0.004736	0.126845	0.007293	0.010505	0.004834	0.010395
245	1.81829404	0.03547117	0.0201	0.3487	0.21917	0.028498	0.203299	0.231106	0.219248	0.230981
259	0.87447925	0.03547117	0.0047	0.1677	0.105406	0.013706	0.097774	0.111147	0.105444	0.111087
279	3.24053413	0.00830989	0.0143	0.2966	0.127991	0.084625	0.147551	0.12419	0.128178	0.124238
293	4.07213287	0.00574242	0.0154	0.3095	0.014343	0.135743	0.009628	0.018627	0.014428	0.018712
309	3.17579604	0.01257306	0.0209	0.3584	0.025451	0.270183	0.002034	0.010903	0.025434	0.010788
337	2.45122003	0.01214226	0.0122	0.2718	0.022468	0.207749	0.000976	0.000544	0.02245	0.000623
357	2.51489648	0.01183341	0.0125	0.2752	0.025949	0.212348	0.000393	0.009746	0.025925	0.009818
370	3.58223486	0.01183341	0.0249	0.392	0.036962	0.302469	0.000559	0.013882	0.036927	0.013985
373	4.61361925	0.01522092	0.0521	0.5736	0.048708	0.429975	0.202476	0.071847	0.049127	0.071881
381	2.36895914	0.01635034	0.0154	0.3054	0.020376	0.216804	0.131396	0.00642	0.020401	0.006345
387	5.03392696	0.01522092	0.0615	0.6258	0.053145	0.469146	0.220922	0.078392	0.053603	0.078429
cutoff	1.966	0.025	0.87	0.2236	0.1	0.1	0.1	0.1	0.1	0.1

Figure 4.2 Model Diagnostics



Hence, we decided to go with **robust regression** as it protects against non-normal residuals by down weighting the outliers, not by excluding them, and bisquare-weighting was used for this.

```
Coefficients:
                  Value       Std. Error  t value
(Intercept)   -17456.2004    849.3239    -20.5531
loan_statusPAIDOFF   -4.4349      0.2767    -16.0269
Principal             0.0024      0.0012      2.0342
terms                10.7714     31.6410      0.3404
due_date              1.0235      0.0498     20.5674
terms:due_date       -0.0006      0.0019     -0.3412

Residual standard error: 2.537 on 394 degrees of freedom
```

Output 4.3 Robust Regression

$$\widehat{(Paid_{Off_{time}} - Effective_{date})} = -17456.2004 - 4.4349 * I(loan\ status = Paidoff) + 0.0024 * Principal + 10.77 * terms + 1.0235 * due_{date} - 0.0006 * terms{:}\ due\_date$$

LOGISTIC REGRESSION

DATA PREPARATION

The purpose of this model is to determine whether a loan can be paid by the due date by using a logistic regression model and come up with an optimum model to maximize the correct classification of potential and non-defaulters. In the logistic model, **Loan_status_binary** will be a binary response variable, where 1 represents *paid off* (60% of data), and 0 otherwise. However, not all of the predictors are intuitively useful for our logistic models, such as the loan ID and past_due_days. Thus, we removed such fields from our model. We also removed paidoff_time for which 25% of the data were missed. Finally, our predictors include Principal, terms, effective_data, due_date, age, education, and Gender.

MODEL FORMULATION (PURPOSEFUL MODEL SELECTION)

1. Fit “simple” logistic regression models for each of the predictors separately. Eliminate any predictor values with large p-values (say >0.2). Principal, terms, effective_date, due_date and Gender have p-value less than 0.2, so we put them into the model together.

```
Call:
glm(formula = loan_status_binary ~ Gender + Principal + terms +
    effective_date + due_date, family = binomial, data = newDataSet)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8903  -1.1645   0.5941   0.9847   2.0578

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    -1.113e+04  1.795e+03  -6.201 5.60e-10 ***
Gendermale     -4.785e-01  2.821e-01  -1.696   0.0899 .
Principal      -6.211e-04  1.055e-03  -0.589   0.5561
terms           2.851e-03  1.925e-02   0.148   0.8822
effective_date  6.810e-01  1.071e-01   6.359 2.04e-10 ***
due_date       -2.805e-02  1.295e-02  -2.166   0.0303 *
---                                                    .
```

2. Conduct backward stepwise selection with remaining predictors, usually using a more stringent cut-off, such as p-value<0.1 Our resulting predictors included Gender, effective_date and due_date.

```
Call:
glm(formula = loan_status_binary ~ effective_date + due_date +
    Gender, family = binomial, data = newDataSet)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8663  -1.1676   0.6166   1.0063   2.0547

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    -1.113e+04  1.793e+03  -6.211 5.27e-10 ***
effective_date  6.822e-01  1.063e-01   6.418 1.38e-10 ***
due_date       -2.932e-02  8.771e-03  -3.343 0.000828 ***
Gendermale     -4.792e-01  2.821e-01  -1.699 0.089378 .
---
```

3. Consider adding in any variables that were not included in the model after Step 1 or Step 2. A predictor can be added in even if p-value>0.1 if it changes the estimated β coefficients by at least, say, 10%. Then, starting with a model with just Gender, effective_date and due_date, we tried adding back in each of the other predictors but changed nothing significantly.

4. Attempt adding plausible interactions among variables in the model, usually using somewhat stricter standards such as p-value<0.05 (can consider non-linear predictor terms, like quadratic effects, in this step as well). After considering all interaction terms did not produce anything extra.

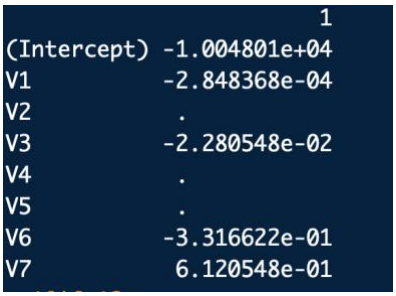
Thus, our final model included Gender, effective_date and due_date as predictors for Loan_status_binary. We also tried automatic backward selection and interestingly it kept everything but Gender. This makes sense because gender has a larger p-value. However, since the model with Gender, effective_date and due_date has the smallest AIC, our final model included Gender, effective_date and due_date as predictors for Loan_status_binary.

MODEL DIAGNOSTICS

Traditionally, accuracy has been the most commonly used performance measure in binary classification problems. For this model AUC score is 0.7, and accuracy is 0.63. Our binary classification model predicted the outcome with 63% accuracy which is considered not that bad. Sensitivity is 0.72, which refers to the part of a loan that the model predicts is positive and not in default, and specificity 0.5 measures negative samples that actually default on the loan. Since we are given an imbalanced dataset, accuracy tends to emphasize majority class, which may mislead the true performance of the model. Therefore, in order to balance both sensitivity and specificity, we used $G\_Mean = \sqrt{sensitivity * specificity}$ to avoid bias towards the majority class. In the final model, the G_mean is about 0.6, which is still not considered as good as we think.

Since there could be potentially multicollinearity issues between principal, terms and due date, we decided to use penalized logistic regression. We have accuracy 0.618 for lasso regression, which was slightly smaller than that of the selection model. Finally, we have G_mean 0.67 for lasso regression, and all predictors from the selection model are kept in the lasso regression model. This proves to a certain extent that these variables have significant contribution to our model. Since we are given an imbalanced dataset, we decided to take lasso regression as our final model, which included Principal, Gender, effective_date and due_date as predictors for *loan_status_binary*. However, since we are using our data to predict the outcomes of our data, this accuracy level in our model is still somewhat misleading and will tend to be somewhat optimistic. To avoid this, with a sufficient amount of data, we can partition the data into a training set and a validation set and also considering sampling techniques like up or down-sampling to deal with imbalance data.

Output: Lasso regression coefficients for principal, due date, gender, effective date respectively



	1
(Intercept)	-1.004801e+04
V1	-2.848368e-04
V2	.
V3	-2.280548e-02
V4	.
V5	.
V6	-3.316622e-01
V7	6.120548e-01

CONCLUSION

From our *first model* where we try to model how many days an individual has to pay back a loan, we can see that there is an inverse relationship between term and the length of the loan. As term increases, the loan length (paid off-effective date) decreases. Interestingly as the principal amount and due date of the loan increase, the loan length also increases for both.

From our second model, some key findings are that as principal and due date increasee, the probability of loan being paid off goes down, and also men are more likely to pay off the loan than women.

One of the limitations of our models is the relatively small number of predictors and also sample size made available to us. A dataset containing some information regarding customers financial stability at the time of taking out the loan could help improve the model's accuracy. Three possible variables that we believe will help the model's accuracy are the customers income, credit score, and past loan defaults. As mentioned previously, a model split into training and testing will help us evaluate the model's accuracy better. With the limitations aside, we believe that this model can provide great contributions in the world of loan lending. The information that our two models provide will allow companies to select customers wisely and accurately calculate earnings for a given time period.