Personal Loan Analysis

Laura Ahumada, Erin McClure-Price, Duy Nguyen
Southern Methodist University, Dallas, Texas, United States

1 Introduction

There are many reasons why the average person would choose to obtain a personal loan, whether it is cash for an emergency or unexpected bill, or perhaps to help with the cost of moving to a new location. Regardless of the reason, an estimated 20 million personal loans will be obtained in 2022 (VanSomeren & Tarver 2022). With an average return of 9.34% ("Consumer Credit - G.19"), it is clear why banks are interested in providing personal loans, and Thera Bank is no exception.

Thera Bank was interested in investigating how to use a marketing campaign to better coerce their customers into taking personal loans. To this end, they collected their customers' demographic data, including age, income, education, family size, and zip codes, in addition to the monetary information they had readily available. Given the wealth of information, including a total of 14 features from 5,000 customers, the information in the data set is ideal for an in depth analysis. We chose to use this dataset, called "Bank Personal Loan Modeling", which is featured many times on the Kaggle website.

Our second project for the Applied Statistics course in the Southern Methodist University Master of Science of Data Science program was to perform an exploratory data analysis on a data set and provide an easily interpretable logistic regression model using a feature selection method of our choosing, followed by another logistic regression model that is more complex– complete with either variable transformations, interaction terms, or polynomials–then another model with LDA or QDA applied only to the available continuous predictors that Thera Bank collected. The R code for all analyses can found in the Appendix [*see Section 7.0*]

2 Personal Loan Data Set

The data set used in this project describes customer demographic information (age, income, etc.), the customer's relationship with the bank (mortgage, securities account, etc.), and the customer response to the last personal loan campaign (Personal Loan). Among these 5000 customers, only 480 (= 9.6%) accepted the personal loan that was offered to them in the earlier campaign. There are no missing values in the dataset. There are a mix of numerical and

categorical variables, where the former include which made data pre-processing necessary. More details regarding the data set can be found in the Appendix [*see Table 6.1*].

3 Exploratory Data Analysis

3.1 Redundant Variables

The first thing that we noticed was that ID could be dropped from the model since it is only a unique identifier and would not affect the results of any model. In addition, the zip code variable contained 467 unique areas, which we agreed was far too many to include as a categorical factor; as such, zip code was removed from further analysis.

3.2 Summary Statistics

Analysis of the remaining variables' summary statistics showed that the Age variable was normally distributed with an average of 45 years old, a minimum age of 23, and a maximum age of 67 [*see Section 6.2*]. The Income variable was mostly normal but slightly right-skewed with an average of $73,000; the median was $64,000, the minimum was $8,000, and the maximum was $224,000. The Family variable was evenly distributed across family sizes of 1, 2, 3, and 4 – or about 25% of the total each. The Average Spending (CCAvg) per $1,000 ranges from 0 to $10,000 and the median is $1,500. We noted that for Education, 41% of customers had completed undergraduate degrees, 28% had completed graduate degrees, and 30% were deemed advanced/professional; this variable appeared to be evenly distributed. The Mortgage variable was highly skewed with a mean of 56.5 and a median of 0. Making it a potential variable that should be logged in the second part of the models; the boxplot created logging mortgage did show a better distribution [*see Section 6.6*]. Both the target variable, Personal Loan, and the variable for Securities Account were unbalanced with 90% of customers in the "no" category (0) and only 10% in the "yes" category (1). The CD Account variable showed increased unbalance with 94% "no" and 6% "yes". The Online variable was more balanced, with 40% "no" and 60% "yes", and the Credit Card variable showed that 70% did not have a credit card with the bank and 30% had a credit card with the bank. The Experience variable presented an interesting potential error through the presence of negative values as low as -3. We looked further into the 52 people with negative total experience and found that they were between 23 and 29 years old and all had a salary with a median of 65,000. We concluded that these individuals could just be young people that just started working, therefore we updated the negative values to be 1.

3.3 Correlations

After transforming Family, Education, Securities.Account, CD.Account, Online and CreditCard into factors, we created a correlation plot using the continuous variables and found that the correlation between Age and Experience had a value of 1 [*see Section 6.14*], however, neither variable had a strong correlation with Personal Loan. Ultimately, we chose to combine the two variables into a new column "Experience2" by dividing Experience by Age.

3.5 Significant Variables

Then, we checked the relationships between the response variable and all other predictors [*see Section 6.7*]. We found that increases in Income, Mortgage, Average Credit Card (CCAvg) and having a CD Account equated with increased likelihood of taking out a loan. Interestingly, we noted that customers who had education past undergraduate school tended to have an increased number of loans, and the number of family members did not seem to greatly affect whether a customer would take out a personal loan, although there was a slight increase for n=3 and n=4. The predictors that appeared to have little to no relationship with Personal Loan, at first, were Security Account, Online, and Credit Card. However, we created a specific plot and these predictors showed that relationships did exist [*see Section 6.8*], and so they were retained for logistic regression analysis. The pair plot also displayed how CCAvg and Income had a 64% correlation which was also viewed in the heatmap, as the pairplot did display its distribution, 62% relationship was with no loan and only 2% relationship with getting a loan. Some Loess plots were created to see further information and so that after a specific amount of spending CCAvg the probabilities of getting a loan somewhat stopped increasing and flatten. For income and mortgage both did again confirm the idea that there is a positive relationship, the higher the mortgage or higher the income the more they would say yes to loan [*see Section 6.3.1*]. The rest of the loess single variable plots did not show much, however when looking at two explanatory variables with Personal Loan we found there could be some interactions in the combination of: Mortgage with education, Family with Income, CCAvg with Family, Income with Education, and Mortgage with Family. The combination of the factors within each pair of explanatory variables did seem to have a different influence in personal loan.

4 Objective 1

4.1 Problem Statement

For Objective 1 we were tasked to create an interpretable logistic regression (LR) model. The first step was to create a Full model with all variables except the two we had removed, ID, Zip Code, Age, and Experience, but including our new variable Experience2 that was created by dividing Age by Experience.

4.2 Building The Model

We created a 70/30 train test split of the data set and began by creating a model with all the variables to see which appeared as significant and if they matched those observed in the EDA. The variables with importance were Income, Family, CCAvg, Education, Securities.Account, CD.Account, Online, CreditCard, interestingly it did not have Mortgage as a significant factor. We used the model with all the variables for 3 model selection methods (StepWise, Forward and Lasso) to see if it would select other variables, specially mortgage . The Forward and stepwise ended up selecting the same attributes as those viewed as significant for the exception of forward actually keeping Mortgage and Experience2 which were not statistically significant. Lasso kept the same attributes as stepwise but included Experience2. We also built 2 other models. The Intuition model was created simply through normal assumptions that certain variables would be related to a customer's comfort with taking out a personal loan and the variables selected were: Income, Family, Education, CD Account, and Credit Card. The EDA model was created based on the actual variables found to have the strongest relationship with the personal loan during the EDA which were: Income, CD.Account, Mortgage, Education, Family and CD.Account.

A function was created to pass through different thresholds for each model to get the best specificity and we saw that 0.3 was the best base cutoff for comparing the specificity, accuracy and sensitivity of models, followed in order 0.5 and 0.7 [*see Section 6.4*]. The results of model performance using the cutoff of 0.3 can be found in Table 1.

Table 1. Performance criteria values from the six models at a threshold of 0.3.

Criteria	**Full Model**	**Stepwise Selection**	**Forward Selection**	**LASSO**	**Intuition**	**EDA**
AIC	820.398	818.074	820.398	819.837	836.879	844.481
BIC	906.645	892.000	906.645	892.00	892.324	899.926
Accuracy	0.957	0.957	0.957	0.956	0.956	0.953
Sensitivity	0.975	0.975	0.975	0.976	0.976	0.974
Specificity	**0.800**	**0.806**	**0.800**	**0.794**	**0.774**	**0.774**
Predictors	All available	Income, Family, CCAvg, Education, Securities Account, CD Account, Online, Credit	Income, Family, CCAvg, Education, Mortgage, Securities Account, CD Account,	Income, Family, CCAv, Education, Securities Account, CD Account, Online, Credit Card,	Income, Family, Education, CD Account, Credit Card	Income, Family, Education, Mortgage, CD Account

		Card	Online, Credit Card, Experience2	Experience2		

Judging by overall performance criteria, the Stepwise selection produced the model with the best metrics. An AIC of 818, BIC of 892 and Specificity of 80%. The specificity was selected because the reference set by the algorithm was No and the scope of the project is to get the best prediction for those customers that would Yes, making Specificity the target. Since this is cut-off dependent, we also used the ROC as a criterion and found that the Stepwise selection still created the best performing model [*see Section 6.5 and Section 6.15*].Output for the winning Stepwise selection model can be found below [*see Section 6.16*]. Interestingly, we suspected that being online and having a credit card would increase the odds of a customer getting a personal loan, seeing that it decreased the odds (based on the coefficients) was a surprise. Otherwise, the results of the model mostly made sense. Increasing family size will increase financial strain making it more likely that a customer will take a loan, and having an investment account in the form of a CD account intuitively indicates that a person would be willing to take a risky venture in the form of a loan.

Best Performing Model Equation (Stepwise Selection)

$$log(\frac{p(x)}{1-p(X)}) = \hat{\beta}_0 + \hat{\beta}_1 Income + \hat{\beta}_2 Family2 + \hat{\beta}_3 Family3 + \hat{\beta}_4 Family4 + \hat{\beta}_5 CCAvg + \hat{\beta}_6 Education2 + \hat{\beta}_7 Education3 + \hat{\beta}_8 Securities.Account1 + \hat{\beta}_9 CD.Account1 + \hat{\beta}_{10} Online1 + \hat{\beta}_{11} CreditCard1$$

4.3 Assumptions

Linearity

By binding the logit value $log(\frac{p(x)}{1-p(X)})$ within our dataframe with only continuous variables and plotting them side by side, we can confirm the linear association of such variables with the personal loan offer acceptance outcome in logit scale. Smoothed scatter plots [*see Section 6.20*] show that variable CCAvg look the most linear, however see that most of the data is on bottom given an idea that it could be logged and adding power terms could work, Income looks fairly linear up until the top part were it seems to curve having it logged and adding a 3rd power term could help here, because the plot seems to be cubic. As seen in the EDA when checking the distribution, Mortgage does need a log-transformation, because its plot is non-linear. *Experience2* is performing strangely according to the unusual spot of clumped data where it should be clumped near the top left and spread to the bottom right (*see Section 6.20*). It coincides with what we saw in the EDA of it not having a relationship with *Personal.Loan* which is why we did not include it in our model. When taking the categorical variable into account we did see in the EDA

that there is a relationship with personal loan [*see Section 6.7*]. The variables mentioned above will be addressed in Objective 2 of this project.

Influential Values

We applied Cook's Distance algorithm to check the model for high leverage outliers [*see Section 6.17*], and found that while there were somewhat high Cook's D and leverage data points, 350, 2249, 2540, we did not feel comfortable removing them because, assumably, the data were accurate and should not be removed.

Multicollinearity

Variables with a VIF value that is greater than 5 or 10 indicate problematic collinearity. However, none of our variables demonstrate high VIF values [see Section 6.19].

4.4 Interpretation

Complete Model of Objective 1

$$log(\frac{p(x)}{1-p(X)}) = -12.548 + 0.064\ Income - 0.206\ Family2 + 1.922\ Family3 + 1.397\ Family4 + 0.127\ CCAvg + 4.018\ Education2 + 4.177\ Education3 - 0.714\ Securities.Account1 + 3.545\ CD.Account1 - 0.817\ Online1 - 0.887\ CreditCard1$$

Coefficients and Confidence Intervals

Interpretation of our model was done by taking the natural logarithm of each variable, resulting in the odds ratios of our winning model. Those of which are displayed below [*see Section 6.21*] along with their confidence intervals.

While holding other variables constant,

- there is enough evidence that an increase of $1,000 in a customer's **income** is associated with about a $100(1.066 - 1)$ = **6.6% increase** in the odds of them accepting a personal loan offer (p-value < 2e-16).
- customers with a family size of 1 are not significant for our study, meaning that their attributes do not concern their likelihood of accepting a personal loan offer.
- there is not sufficient evidence that customers with a **family size of 2** are associated with about a $100(0.814 - 1)$ = **18.6% decrease** in the odds of them accepting a personal loan offer (p-value = 0.4639).
- there is enough evidence that customers with a **family size of 3** are associated with about a $100(6.834 - 1)$ = **583.4% increase** in the odds of them accepting a personal loan offer(p-value = 7.97e-11).

- there is enough evidence that customers with a **family size of 4** are associated with about a $100(4.042 - 1) = $ **304.2% increase** in the odds of them accepting a personal loan offer (p-value = 1.537404e-06).
- there is enough evidence that an increase of $1,000 in a customer's **average spending on credit cards per month** is associated with about a $100(1.135 - 1) = $ **13.5% increase** in the odds of them accepting a personal loan offer (p-value = 0.021412).
- customers with **highschool** education are not significant for our study, meaning that their attributes do not concern their likelihood of accepting a personal loan offer.
- there is significant evidence that customers with **undergraduate** education level are associated with about a $100(55.564 - 1) = $ **5456.4% increase** in odds of them accepting a personal loan offer (p-value < 2e-16).
- there is significant evidence that customers with **graduate** education level are associated with about a $100(65.186 - 1) = $ **6418.6% increase** in odds of them accepting a personal loan offer (p-value < 2e-16)..
- there is significant evidence that customers with a **securities account** are associated with about a $100(0.489 - 1) = $ **51.1% decrease** in odds of them accepting a personal loan offer(p-value 0.0408).
- there is significant evidence that customers with a **certificate of deposit (CD) account** are associated with about a $100(34.648 - 1) = $ **3364.8% increase** in odds of them accepting a personal loan offer (p-value < 2e-16).
- there is significant evidence that customers that use **(Online) internet banking facilities** are associated with about a $100(0.442 - 1) = $ **55.8% decrease** in odds of them accepting a personal loan offer (p-value = 5.204664e-05).
- there is significant evidence that customers that use a **credit card issued by the bank** are associated with about a $100(0.412 - 1) = $ **58.8% decrease** in odds of them accepting a personal loan offer (p-value = 5.98e-04).

4.5 Conclusions

Stepwise selection with a classification cutoff of 0.3 led to the best model based on performance criteria. It correctly predicted that customers would accept a personal loan 80% of the time. We are aware this leaves a 20% error margin within the model where a customer that would have accepted a loan as No instead, essentially missing that opportunity. This is attributed to having little data on customers that do not accept loans. There was data for 480 people that did not accept a loan versus 4,520 data entries of people that did accept a loan, meaning that we have 9 times more information on those that would accept a loan. Based on the available data, our model actually performed well. Something to point out is that 80% was the specificity score because the reference selected by the algorithm was “No”, making

sensitivity the chances for the model to correctly predict that a customer would not accept a loan when they did not want a loan. Therefore, making specificity our target value. Of course due to this metric being based on cutoffs we also took into account the AIC, BIC and ROC. Out of all the models it did have the lowest AIC of 818 and lowest BIC of 892. For the ROC curve we focused on having the lowest false positive and that also pointed to the Stepwise model. The variables selected were Income, family, CCAvg, Education, Securities.Accoung, CD.Accoung, Online and Credit Card. All of these variables concur with EDA analysis as being significant. The only variable that the model did not pick was seen as important in the EDA was mortgage. We will continue to add in the second set of the models to see if it does end up as significant once we add complexity. The variables that influenced the most according to the model were Income, Family of 3, Family of 4, Undergraduate education (Education 2), Graduate (Education 3) and having a certificate deposit account (CD.Account). All these variables had a positive influence of a minimum 3,364.8% increase in odds of a customer receiving a loan.

5 Objective 2

5.1 Building Competing Models

In Objective 1 we saw that Income, Family, CCAvg, CD.Account, Education and Credit Card were significant, therefore we leveraged those variables and kept *Mortgage* based on the EDA analysis, for these second set of models we added complexity.

Interaction Term Models

We performed 5 interactions models, 4 that included all the same variables but a single different interaction each and the last one included the 4 interactions. The interactions selected were the following: Mortgage with education, Family with Income, CCAvg with Family, Income with Education and Mortgage with family. Each of these interactions was based on LOESS plots [*see Section 6.3.2*]. In those graphs we can see how not only these had an influence in personal loan but that depending on the combination of the two variables the relationship with personal loan would vary, become more or less significant. For instance both, (family with mortgage) and (family with income) had similar results, where family 1 and 2 would would require a larger income or bigger mortgage to start to see a relationship with personal loan but the slope was small and not significant enough to ever have a loan, while family 3 and 4 would start the relationship with personal loan a little earlier with a higher slope and do lead to having a loan. In each of these models we checked the interaction was significant in the model and in the ANOVA type 3 test. When comparing all the models, the one that included all of the interactions; Mortgage with Education, Family with Income, CCAvg with Family, Income with Education and Mortgage with family and the rest of the variables: Securities.Account,

CD.Account,Credit Card performed the best. All of the interactions did show significance in the model, but checking Mortgage and Family through ANOVA showed no significance and neither did Education with Mortgage. Interestingly, when comparing the models, the one that performed the best was the one that included all of the interaction terms [*see Section 6.9.1*]. It had the lowest AIC (631.036) as well as the lowest BIC (772.728) and the best specificity (88%). The ROC provided support for it being the best model [*see Section 6.9.2*].

Log-transformed Models

We had seen in the EDA that logging Income and Mortgage could be useful so we created a model with the same variables leveraged from the Objective 1 analysis and EDA. It included Family, CCAvg, CD.Account, Education, Credit Card, and log transformed mortgage and income, and all showed significance. This was very good because in Objective 1 none of the models selected for Mortgage even though it appeared to be significant from the EDA. he model did well but we went ahead and created a model mixing all of the interaction terms with the logged variables which ultimately led to a better result than including just the logged variables, increasing specificity from 80% to 87%. It also had the lowest AIC (600) and BIC (766) values overall [*see Section 6.1*0], and ROC curves provided visual confirmation that it was the best performing model so far [*see Section 6.10*].

LDA and QDA Models

The LDA model and QDA model performed poorly, however, between the two, the LDA model outperformed the QDA model. This makes sense since the groups have a clear division of who gets loans and who does not which fits for LDA modeling. Visually, the LDA and QDA models also do not perform as well as the logistic regression models [*see Section 6.12*]. Neither the LDA nor QDA have AIC or BIC, but specificity was only 55% and its ROC line was the closest to the diagonal when compared to the rest of the models created.

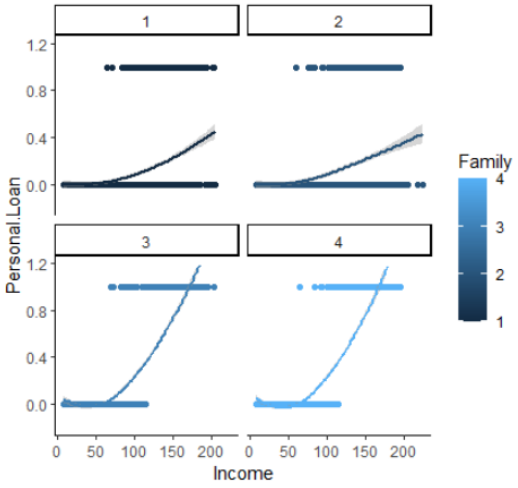
Polynomial Models

Giving predictors power in our models was an added attempt to make the numeric variable fit the data better due to their distribution while adding complexity. The only continuous variables that are able to be made into polynomials in our dataset were Income and CCAvg. Logistic regression models were created with these 2 variables raised to the 2nd power then the 3rd power individually, as well as both together, adding to a total of six models. According to the performance criteria and ROC curves [see Section 6.11], the model with Income raised to the 3rd power was the best performing model within the polynomial models. It led to an AIC of 648, BIC of 746, and specificity of 86%, and the ROC curve confirmed it.

5.2 Main Analysis

Out of the four competing model categories above, the model with Income raised to the 3rd power and the model with log transformed and interaction terms performed the best. One way to further improve the polynomial logistic regression model is to use the log-transform Mortgage variable created during the EDA, and to see whether the stepwise feature selection would still select this new variable.

Using our eda analysis, model results and intuition a final contending model was created with interaction terms Income*Family, along with the two features mentioned earlier, log-transforming Mortgage and 3rd power Income, because family sizes 1 and 2 respond much more differently to the likelihood of accepting a personal loan offer than family sizes 3 and 4 on the scale of their Income. This specific interaction is much more clear than any other pairs of interactions as seen from the loess curves [*see Section 6.3.2*]. We passed the model through stepwise feature selection to check if variables remained and it kept all variables. This model dominates both the stepwise model, the winning model from Objective 1, and the LDA model. It had the best specificity achieved so far of 90%, the lowest AIC of 584, BIC of 695 and the ROC showed that indeed the best model when compared to the earlier top models [see section 6.11.2].



5.3 Comparing Models Table

Table 2. Performance criteria for the models created for Objective 2.
We did not include BIC or AIC for these due to LDA not having AIC or BIC. View full table of all criterions including ROC graph of best models in [see section 6.11.2].

Model	Accuracy	Sensitivity	Specificity
Stepwise	0.9613	0.99703	0.65161
Complex	0.972	0.97844	0.91613
LDA	0.896	0.945	0.445

5.4 Conclusion

After the creation and analysis of several models with additions of logged variables, interaction terms, polynomials and finally a polynomial with log transformed and interaction terms, we were able to obtain the best overall performing model. The specificity of the Complex model showed an increase of 10% when compared to the best performing model from Objective 1. The results of the Complex model inferred that we could now correctly predict 90% of the time the people that wanted a loan which leaves only a 10% error margin for false positives in the form of wrongly predicting that customers would say No to a personal loan when actually wanting it. Ideally we would want an even lower percentage of false positives, however, a prediction of 90% is still pretty good given the fact that there is so little information in this data on those that actually want a loan and so much data on those that will say no to a loan. In addition, LDA performed better than QDA even though the explanatory variables were standardized. Neither model outperformed the Complex logistic regression model nor the simple model from Part 1. The reasons for this could be due to slightly non-normal distributions of some of the variables such as Mortgage and Income.

Throughout this process we were able to see how, just like in Objective 1, the predictors Income, Family, Education, CD Account remained as the most important variables, with the only difference being that now the interaction of Family and Income became significant. The complexity added to variables that were already the most significant seemed to help the model. All these variables were noted in the EDA as potentially important predictors. The variable that did not have as much influence but remained in the model was log transformed Mortgage. Income to the 3rd power along with CCAvg, CreditCard and Online also helped the model but were not as influential. The utility of adding complexity to an otherwise simple model is clear, as model performance was increased. While the Complex model from Objective 2 did outperform the model from Objective 1, it did so by a narrow margin and at the loss of interpretability. However, our main purpose was to be able to create a model with greater prediction powers, which the Complex model succeeded in doing, so that fits our goal.

6 Appendix

6.1 Data Descriptions

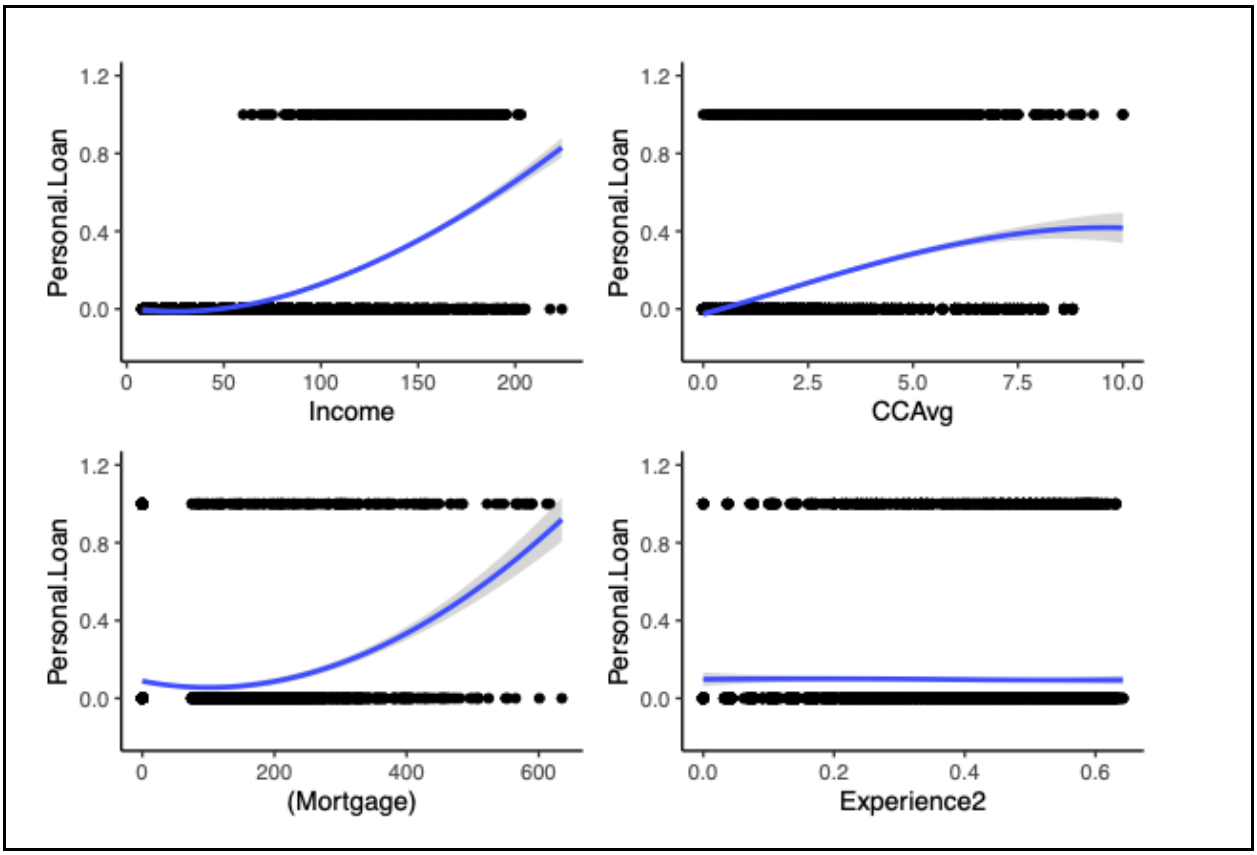
Attribute	Attribute Information
ID	Customer ID
Age	Customer's age in completed years
Experience	Number of years of professional experience
Income	Annual income of each customer ($1000)
Zip.Code	Home address zip code
CCAvg	Average spending on credit cards per month ($000)
Education	Education level where: 1: Undergrad 2: Graduate 3: Advanced/Professional
Mortgage	Value of house mortgage if any ($1000)
Personal.Loan	Did this customer accept the personal loan offered in the previous year?
Securities.Account	Does the customer have a securities account with the bank?
CD.Account	Does the customer have a certificate of deposit (CD) with the bank?
Online	Does the customer use internet banking facilities
Credit.Card	Does the customer use a credit card issued by the bank?
Family	Family size of the customer

6.2 Summary Statistics

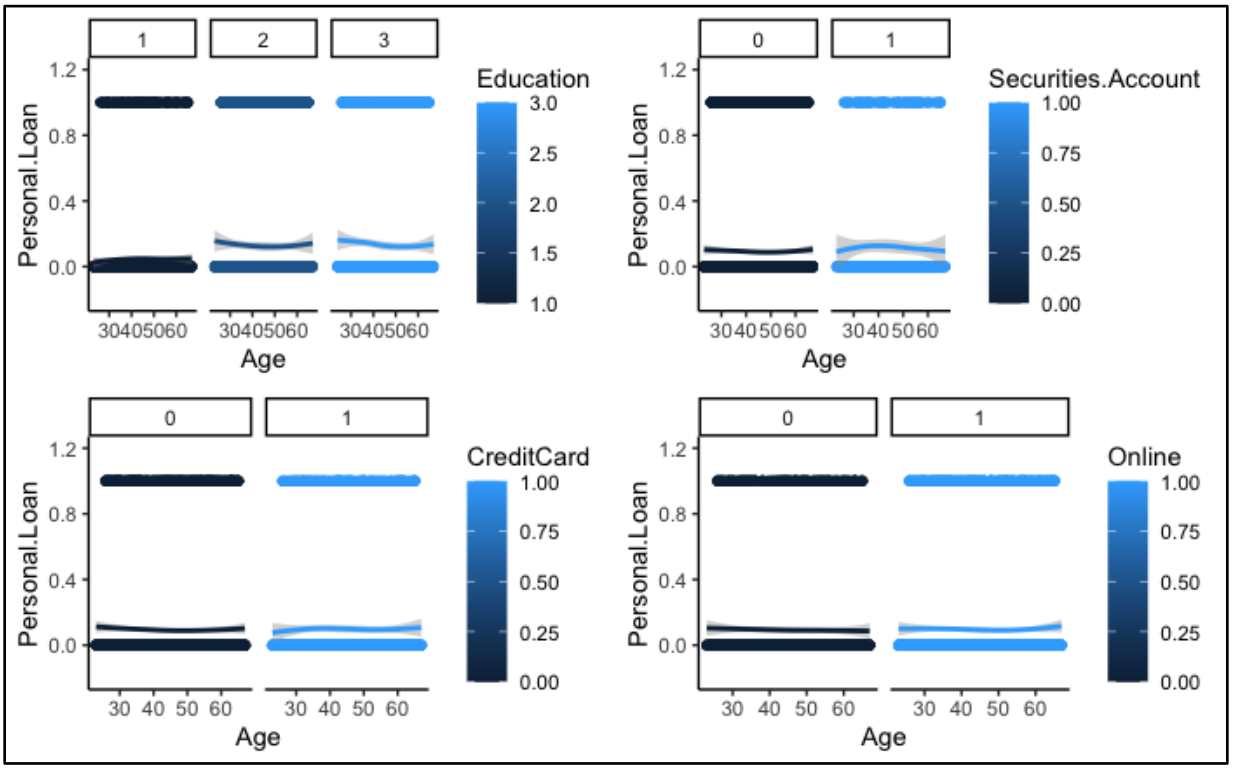
Categorical variables with only two levels where 0 indicates "No" and 1 indicates "Yes" are Personal Loan, Securities Account, CD Account, Online, CCAvg and Credit Card. Family has four levels indicating family size, and Education has three levels indicating undergraduate school (1), graduate school (2), or advanced/professional (3).

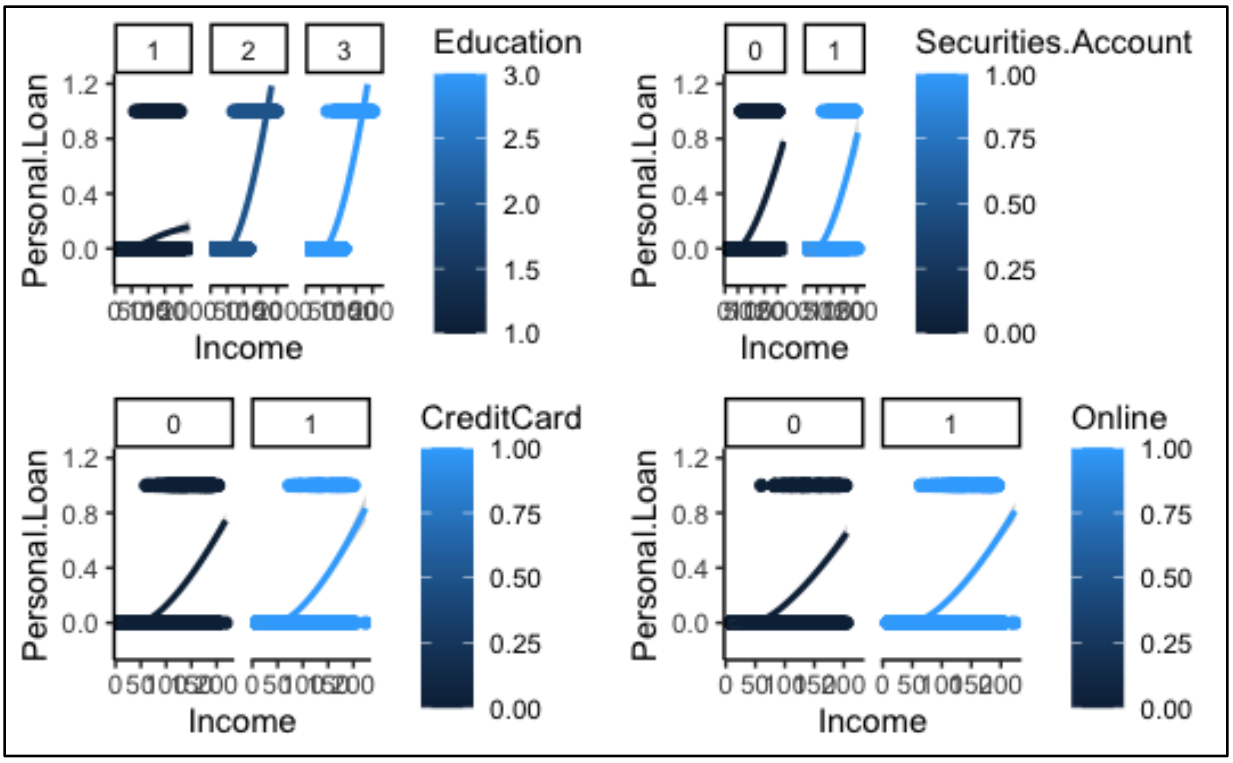
Predictor (Continuous)	**Minimum**	**1st Quantile**	**Median**	**Mean**	**3rd Quantile**	**Maximum**
Age	23.0	35.0	45.0	45.3	55.0	67.0
Experience	-3.0	10.0	20.0	20.1	30.0	43.0
Income	8.0	39.0	64.0	73.8	98.0	224.0
CCAvg	0.0	0.7	1.5	1.9	2.5	10.0
Mortgage	0.0	0.0	0.0	56.5	101.0	635.0
Predictors (Categorical)	**Levels**					
		0	**1**	**2**	**3**	**4**
Family	Frequency		1472	1296	1010	1222
	Proportion		0.294	0.259	0.202	0.244
Education	Frequency		2096	1403	1501	
	Proportion		0.419	0.281	0.300	
Personal Loan	Frequency	4520	480			
Securities Account	Frequency	4478	522			
CD Account	Frequency	4698	302			
Online	Frequency	2016	2984			
Credit Card	Frequency	3530	1470			

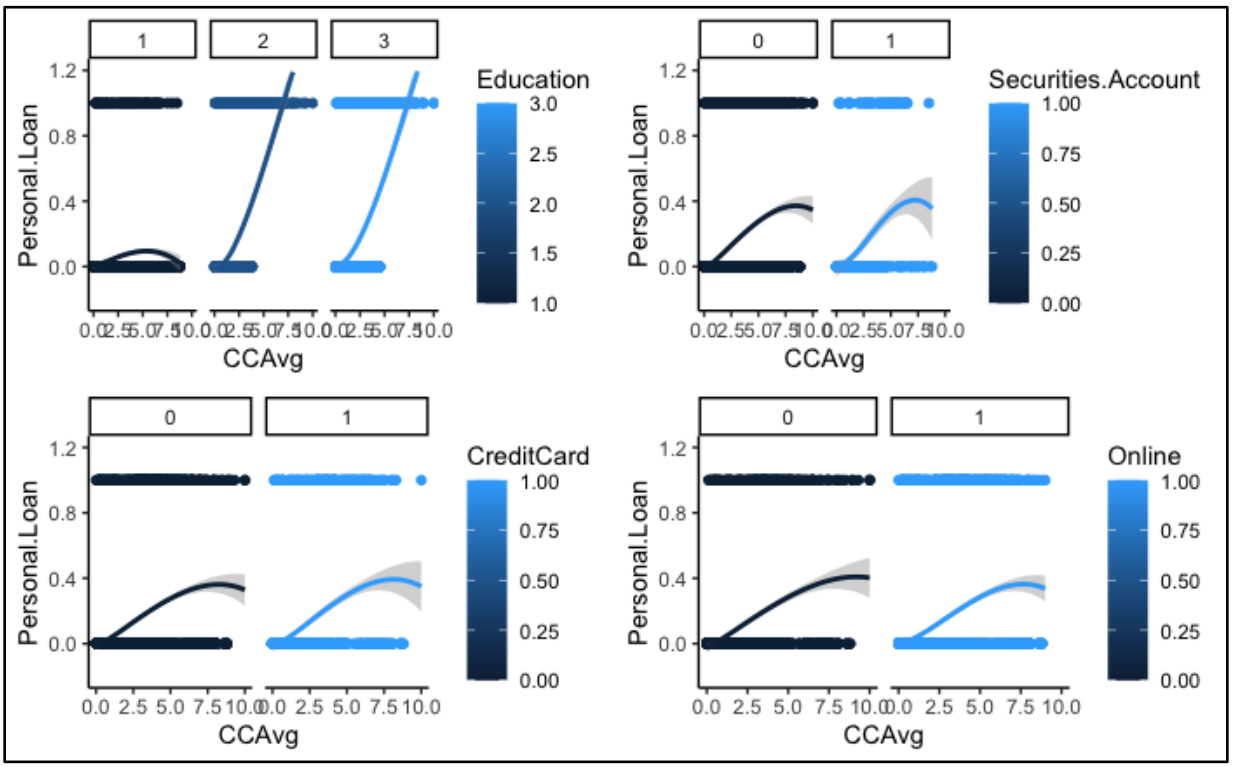
6.3.1 LOESS Curves: Single Variables

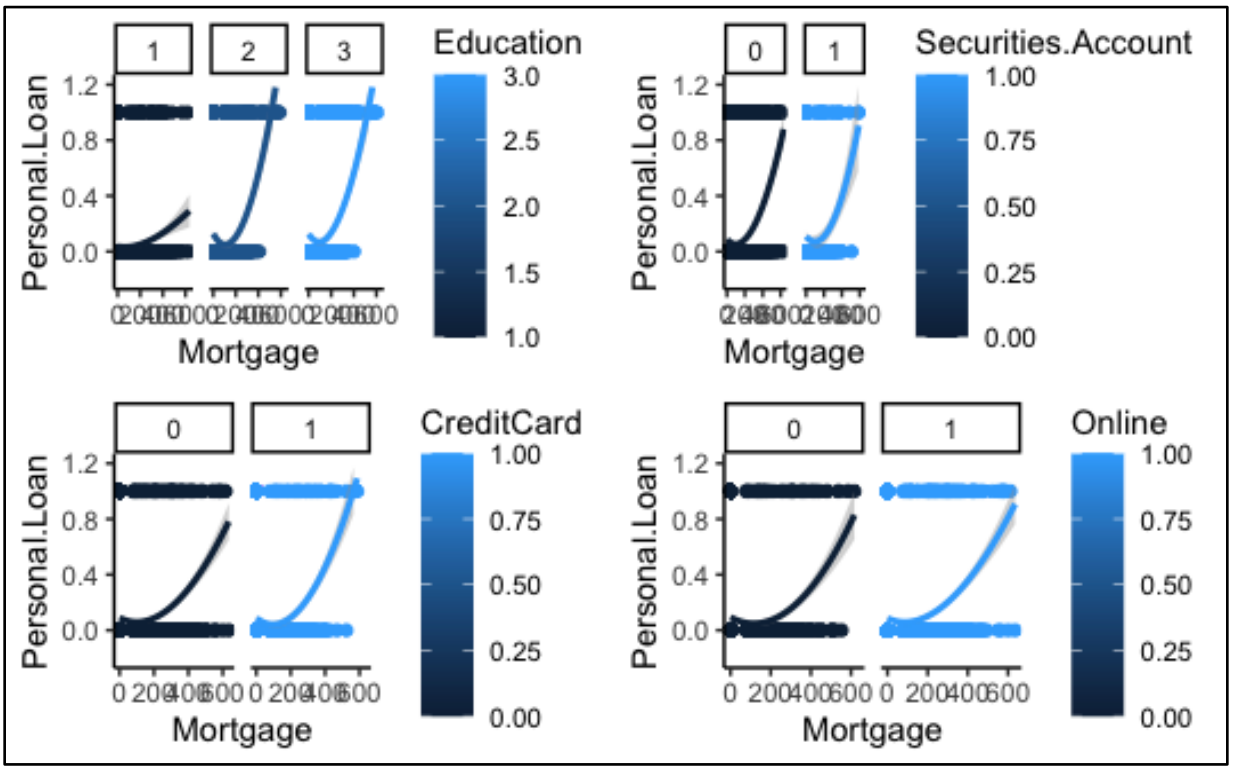


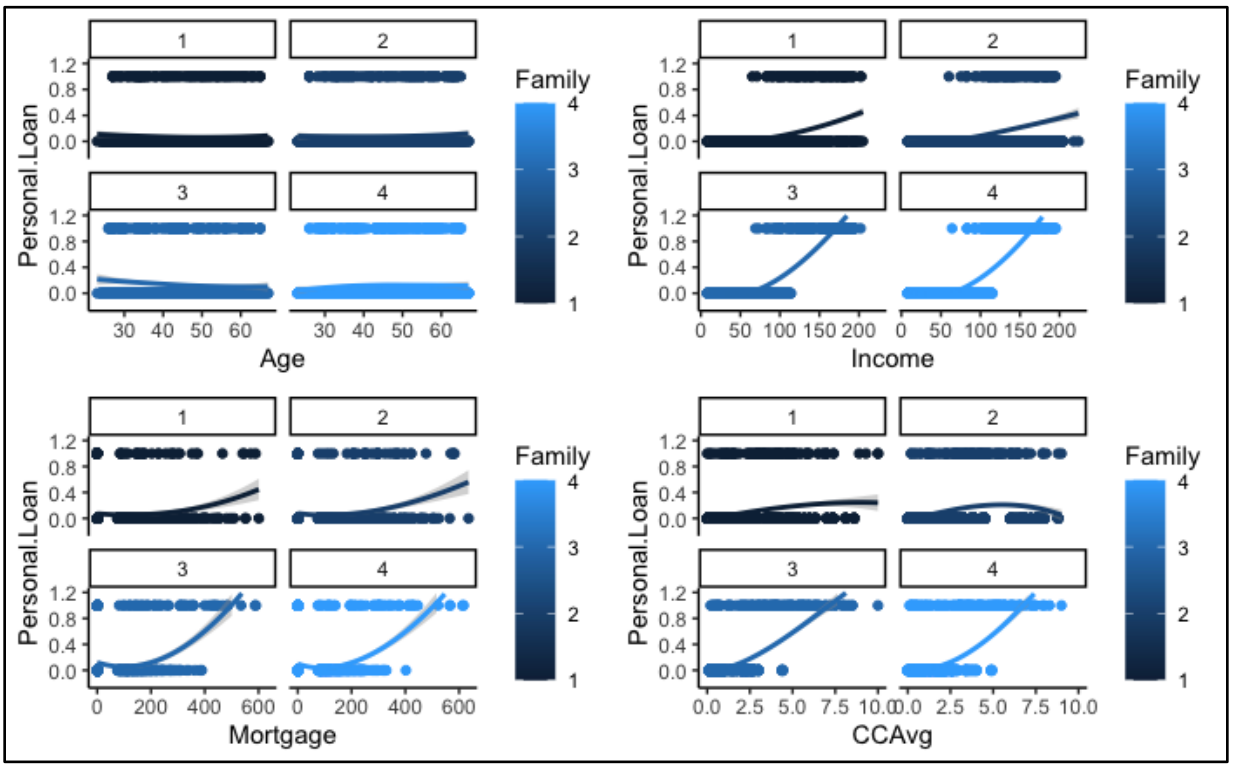
6.3.2 LOESS Curves: Two Variables and Response Variable






1
2
3
Education
3.0
2.5
2.0
1.5
1.0
0
1
Securities.Account
1.00
0.75
0.50
0.25
0.00
CreditCard
Online
Personal.Loan
1.2
0.8
0.4
0.0
0.0 2.5 5.0 7.5 10.0
CCAvg


1
2
3
Education
3.0
2.5
2.0
1.5
1.0
0
1
Securities.Account
1.00
0.75
0.50
0.25
0.00
CreditCard
Online
Personal.Loan
1.2
0.8
0.4
0.0
0 200400600
Mortgage


1
2
3
4
Personal.Loan
1.2
0.8
0.4
0.0
Family
4
3
2
1
30 40 50 60
Age
0 50 100 150 200
Income
0 200 400 600
Mortgage
0.0 2.5 5.0 7.5 10.0
CCAvg

6.4 Comparing Thresholds Function

Accuracy, Sensitivity, and Specificity information for all models from Objective 1 based on three different thresholds.

Model	Threshold	Accuracy	Sensitivity	Specificity
Full	0.3	0.957	0.975	**0.806**
	0.5	0.962	0.995	0.677
	0.7	0.955	0.999	0.574
Stepwise	0.3	0.957	0.975	**0.806**
	0.5	0.962	0.995	0.677
	0.7	0.955	0.999	0.574
LASSO	0.3	0.822	2.106	**0.284**
	0.5	0.841	2.164	0.226
	0.7	0.853	2.200	0.168
Intuition	0.3	0.957	0.975	**0.806**
	0.5	0.962	0.995	0.677
	0.7	0.955	0.999	0.574

6.5 Competing Models Criteria

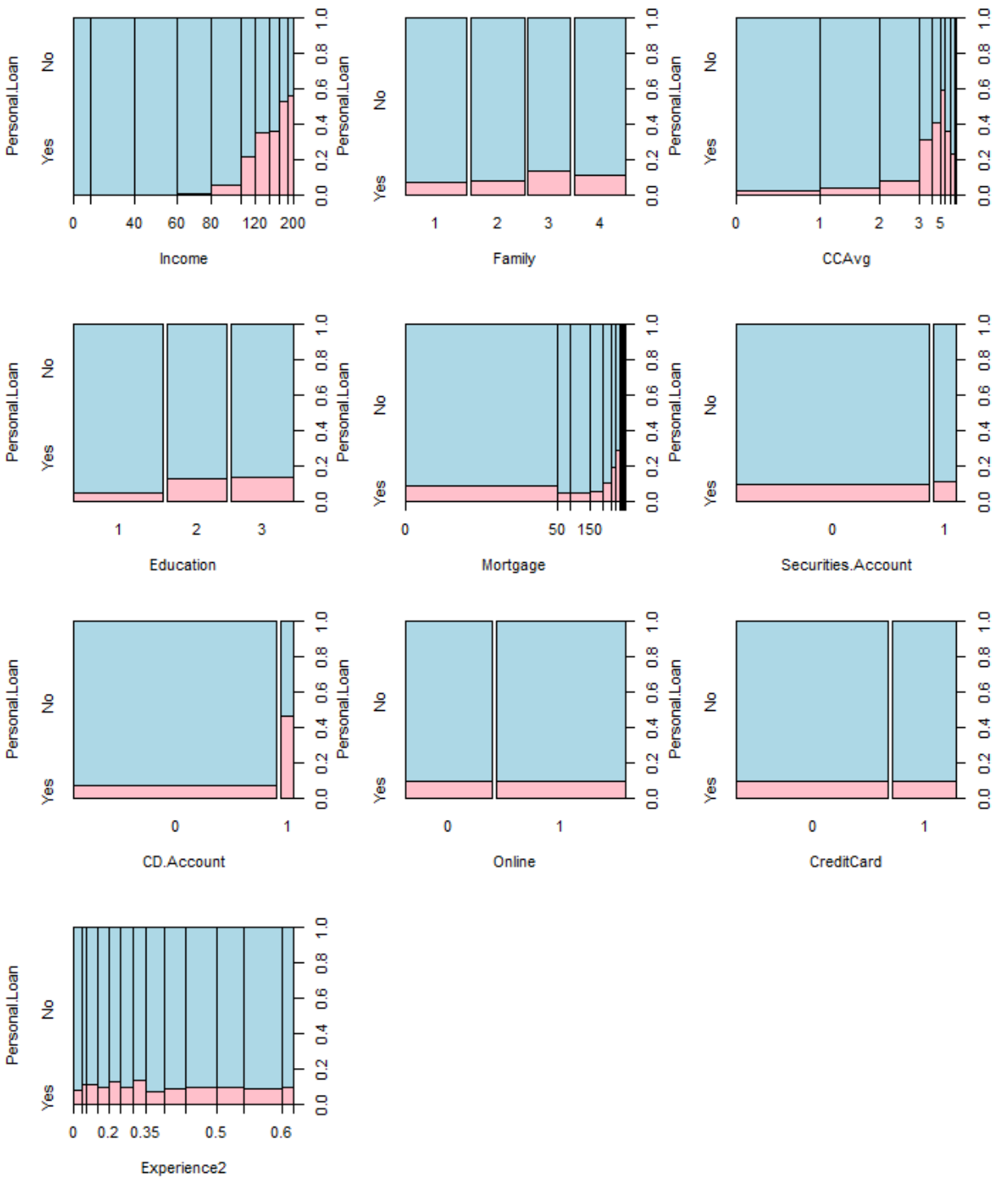
Accuracy, Sensitivity, and Specificity information for all models from Objective 1 based on three different thresholds. Values are rounded to the 3rd decimal (0.001).

Model	AIC	BIC	Threshold	Accuracy	Sensitivity	Specificity
Stepwise	818.0738	892	0.3	0.9613	0.6516	0.9970
LASSO	818.0738	892	0.44	0.9633	0.6774	0.9963

Model	AIC	BIC	Threshold	Accuracy	Sensitivity	Specificity
Intuition	836.8792	892.3239	0.5	0.96	0.6516	0.9955
EDA	844.4809	899.9256	0.5	0.96	0.6516	0.9955

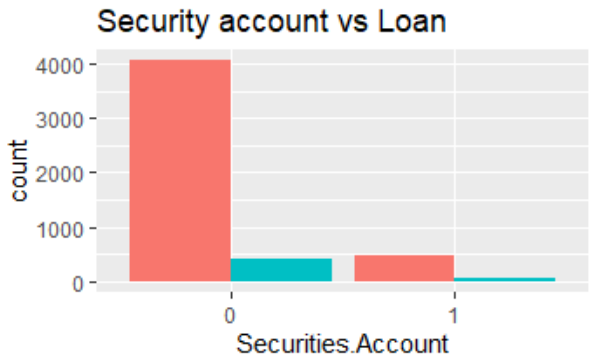
6.6 Logging Mortgage Boxplots

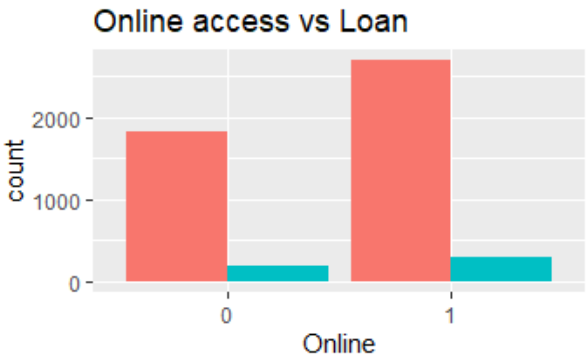
6.7 Correlation Plots

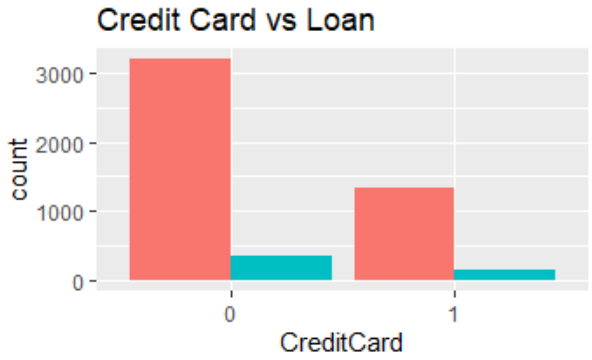


Correlation plots of the response variable, Personal Loan, vs all available predictors.

6.8 Plot of Specific Predictors





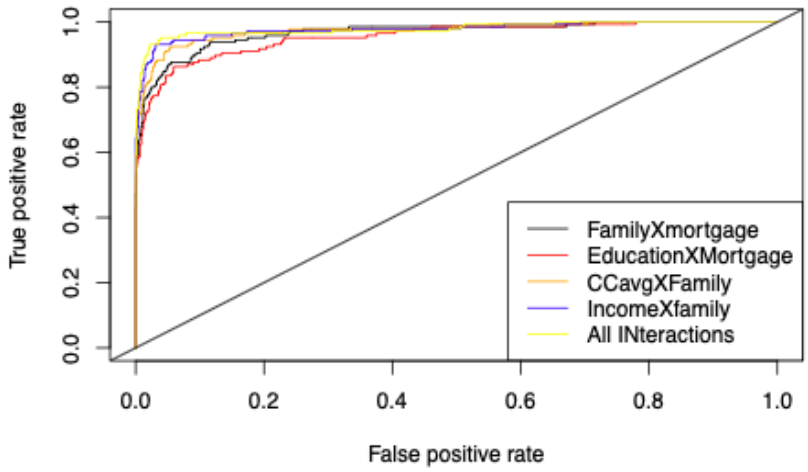


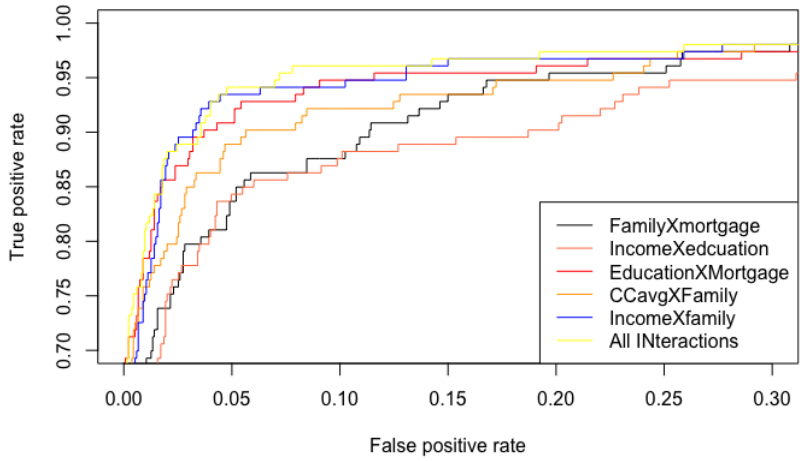
Analysis of Section 6.7 indicated that these three predictors, Security Account, Online, and Credit Card, had little to no relationship with Personal Loan. Upon further inspection, we noted that there were relationships and so we retained the predictors for logistic regression analysis.

6.9.1 Criteria Table For Interaction Models

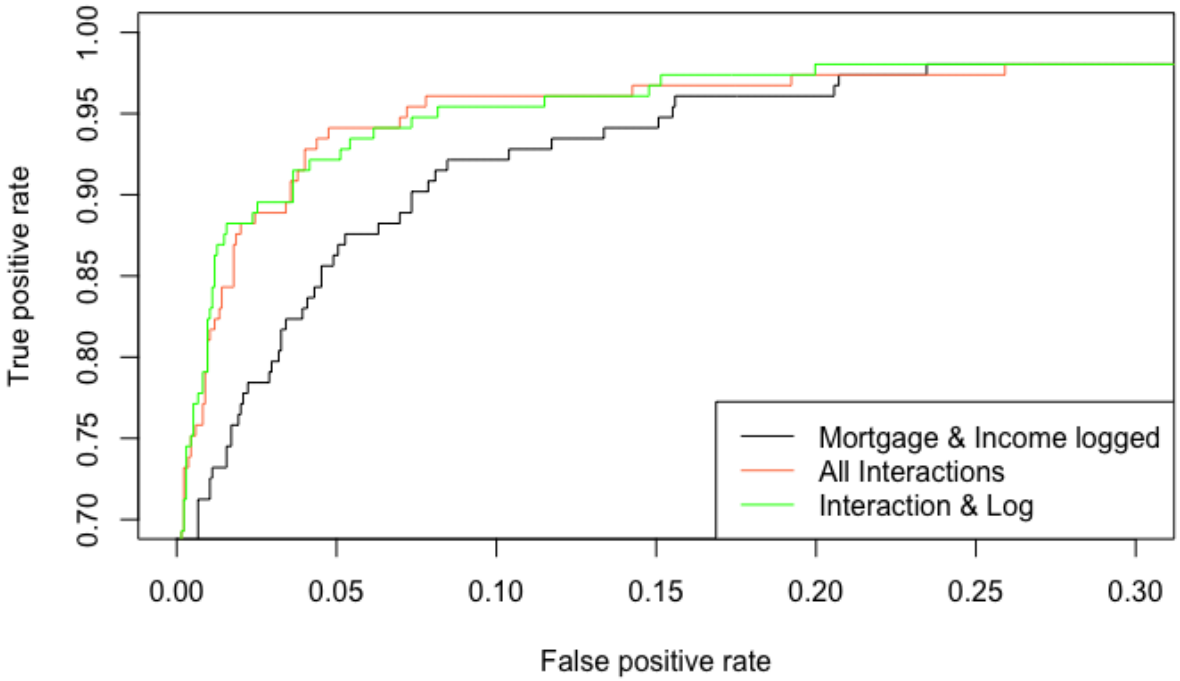
	Criterion	Famxmortgage	EduXMortgage	CCAvgxFam	FamxIncome	AllInter
1	AIC	844.759	916.902	768.813	655.916	631.036
2	BIC	937.167	984.668	855.061	748.324	772.728
3	Accuracy	0.961	0.953	0.962	0.973	0.976
4	Sensitivity	0.980	0.973	0.977	0.984	0.986
5	Specificity	0.781	0.774	0.822	0.870	0.884

6.9.2 ROC Curves of Interaction Models





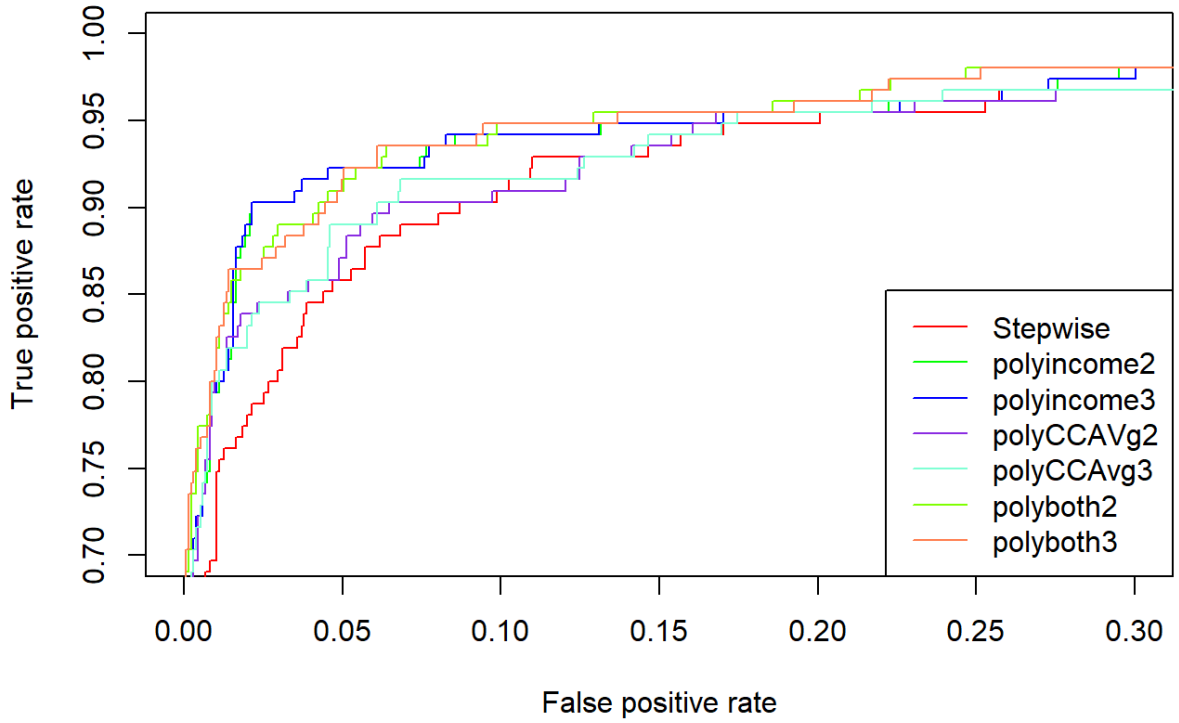
6.10 Criteria and ROC Curves of Log-transformed Models



6.11 Criteria and CROC Curves of Polynomial Models

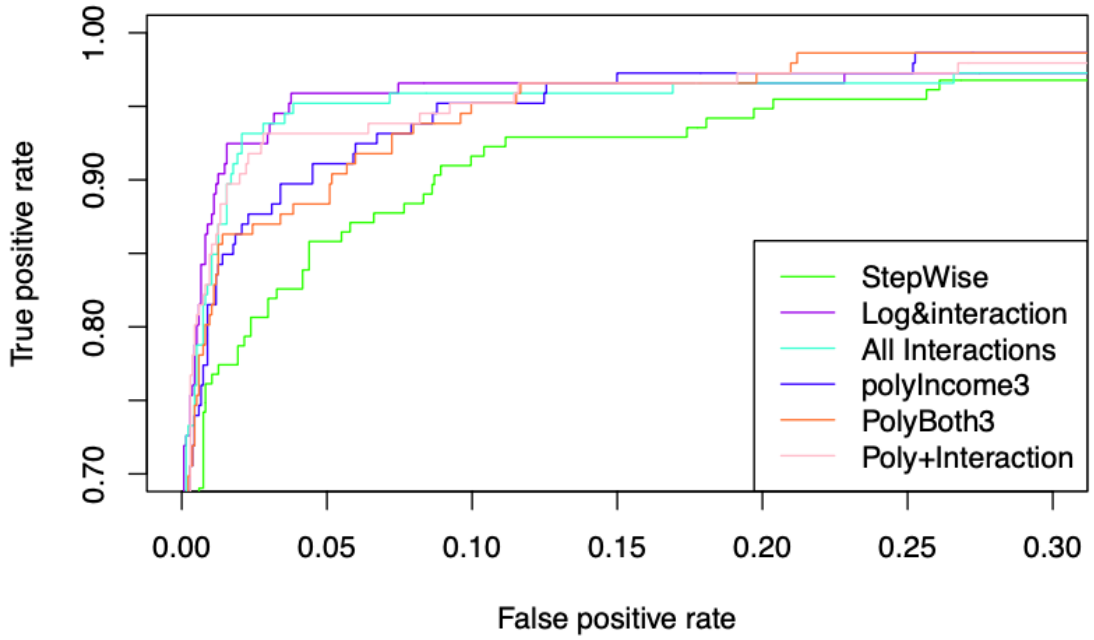
All polynomial models are predicted with a probability cutoff of 0.3.

	Criterion	polyIncome2	polyIncome3	polyCCAvg2	polyCCAvg3	PolyBoth2	PolyBoth3
1	AIC	694.203	695.393	784.087	780.157	661.317	663.552
2	BIC	774.290	781.640	870.334	872.564	747.564	762.120
3	Accuracy	0.968	0.969	0.955	0.957	0.965	0.966
4	Sensitivity	0.979	0.979	0.971	0.975	0.976	0.977
5	Specificity	0.870	0.870	0.808	0.795	0.863	0.863

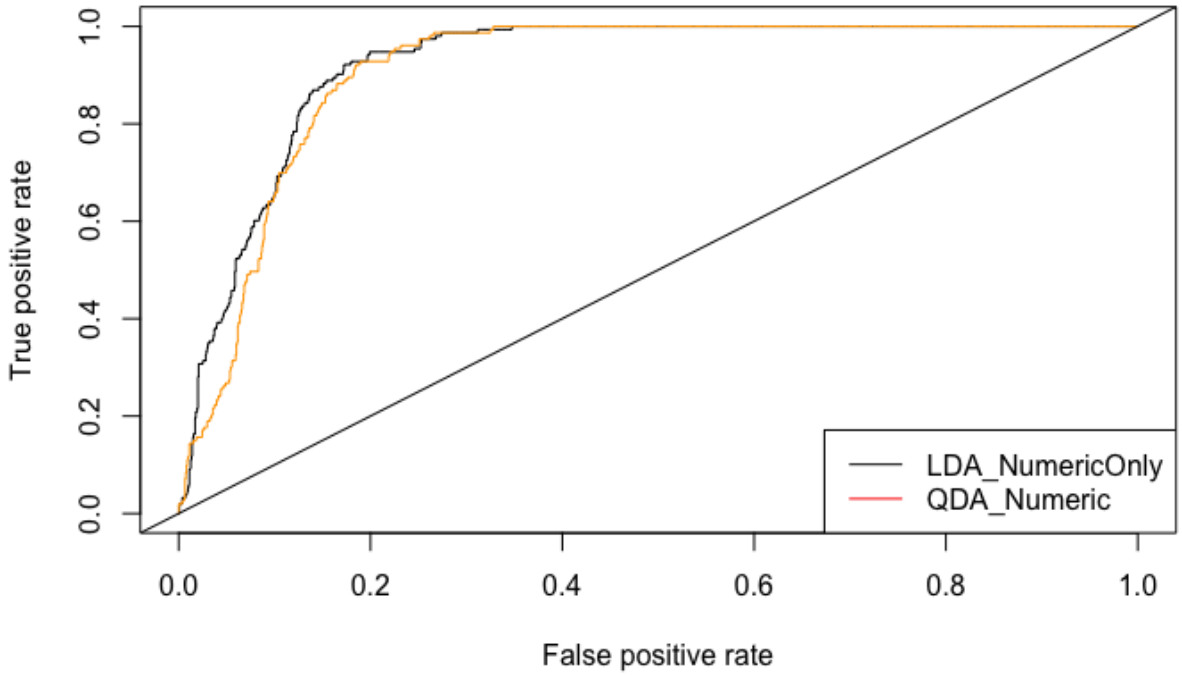


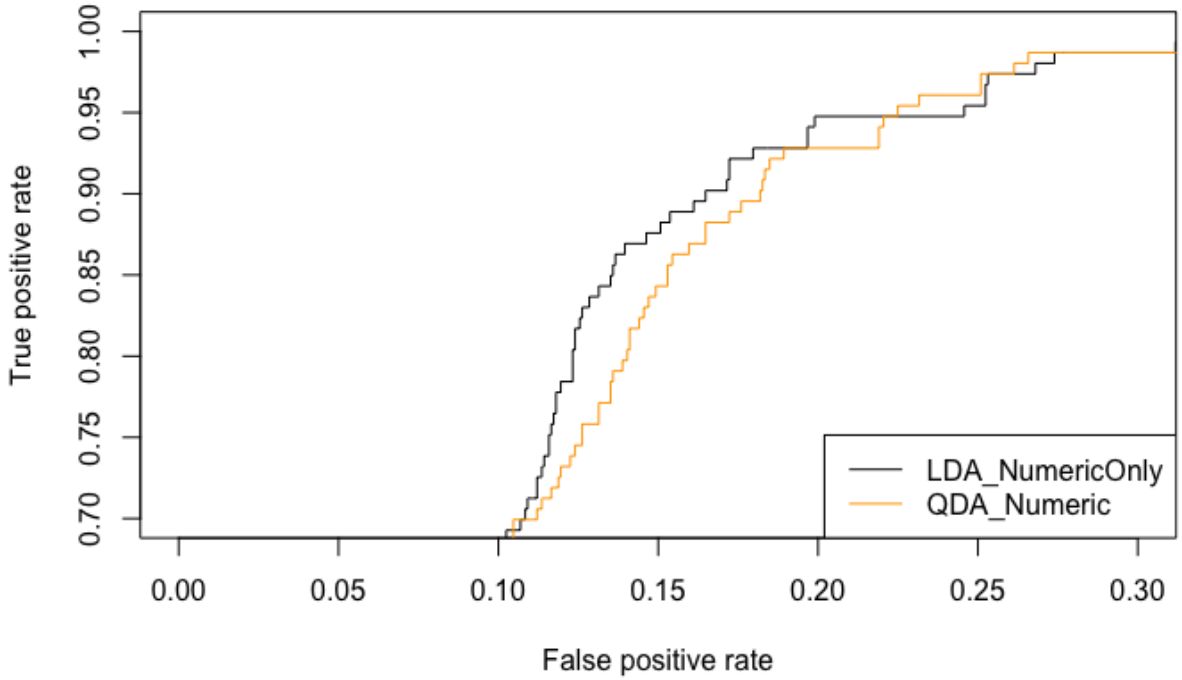
6.11.2 ROC Curves and criterion of best models

Criterion <chr>	Step_Wise <dbl>	LogAndInteraction <dbl>	LDA <dbl>	QDAn... <dbl>	polyIncome3 <dbl>	PolyBoth3 <dbl>	finalmodel2 <dbl>	AllInteractions <dbl>
AIC	818.074	595.571	0.000	0.000	671.141	647.923	584.356	606.312
BIC	892.000	761.905	0.000	0.000	757.388	746.491	695.246	729.523
Accuracy	0.957	0.972	0.899	0.882	0.969	0.971	0.969	0.973
Sensitivity	0.975	0.980	0.957	0.931	0.981	0.981	0.978	0.983
Specificity	0.806	0.906	0.412	0.469	0.863	0.881	0.887	0.887

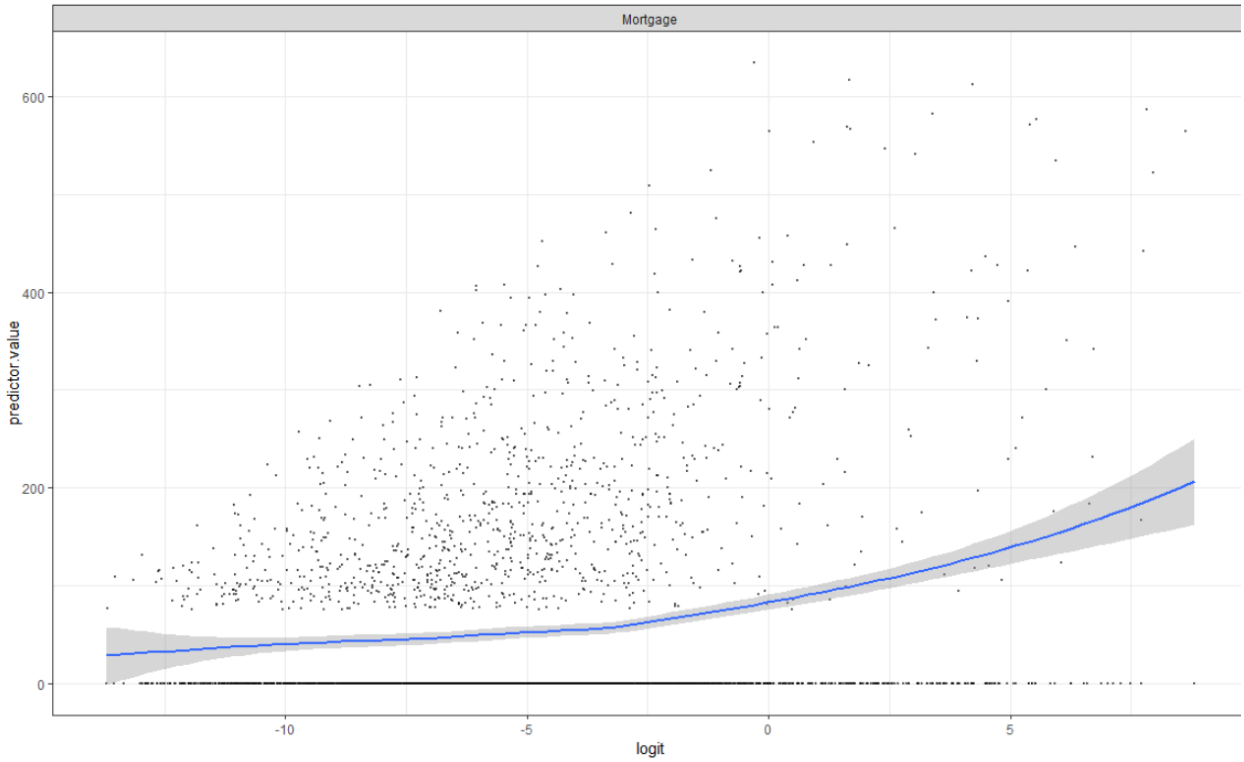


6.12 CROC Curves of LDA and QDA Models

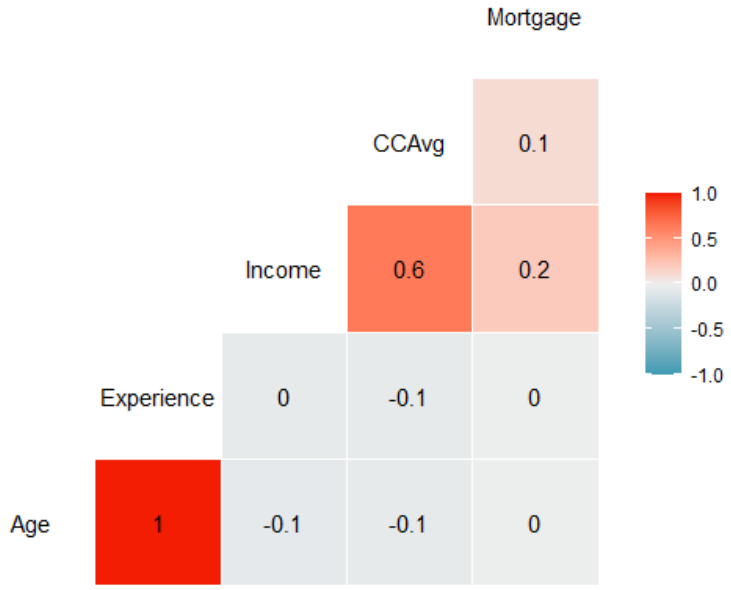




6.13 Mortgage Analysis

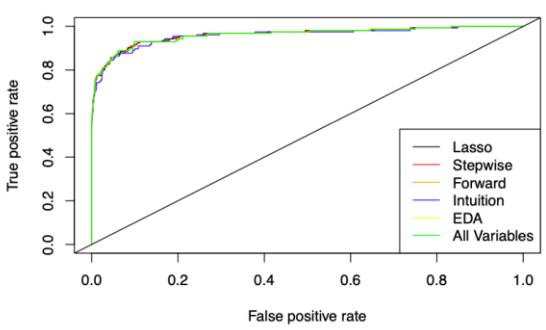


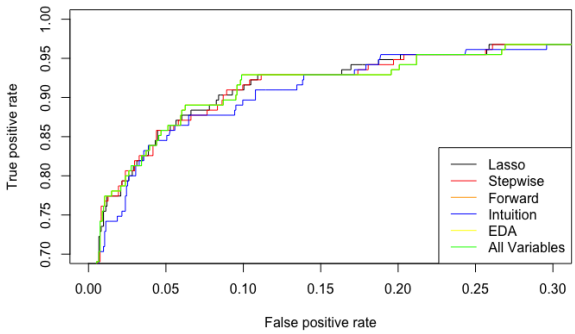
6.14 Correlation Plot of Continuous Variables



The Age and Experience variables had a correlation of 1, indicating multicollinearity was present.

6.15 ROC Curves Describing Model Performance of Part 1 models





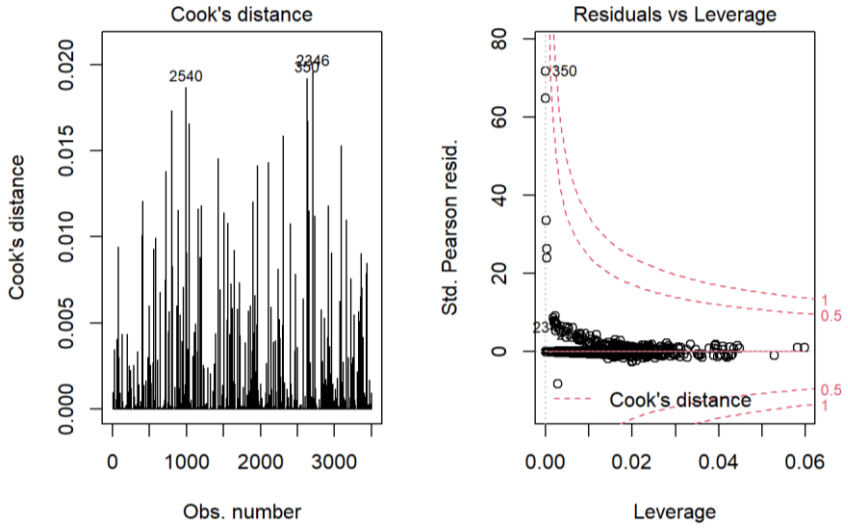
ROC curves for the models created for Objective 1 showed that all models performed with remarkable similarity, though the Intuition model did slightly underperform in comparison to the others. The second plot below is the zoomed in version

6.16 Output for the Best Performing Model

```
Call:
glm(formula = Personal.Loan ~ Income + Family + CCAvg + Education +
    Securities.Account + CD.Account + Online + CreditCard, family = "binomial",
    data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9123  -0.1813  -0.0649  -0.0194   4.1340

Coefficients:
                     Estimate Std. Error z value             Pr(>|z|)
(Intercept)        -12.547610   0.663989 -18.897 < 0.0000000000000002 ***
Income               0.063807   0.003717  17.167 < 0.0000000000000002 ***
Family2             -0.206161   0.281441  -0.733              0.463853
Family3              1.921856   0.295621   6.501      0.0000000000797 ***
Family4              1.396735   0.290604   4.806      0.0000015374042 ***
CCAvg                0.126789   0.055110   2.301              0.021412 *
Education2           4.017534   0.334135  12.024 < 0.0000000000000002 ***
Education3           4.177245   0.333415  12.529 < 0.0000000000000002 ***
Securities.Account1 -0.713607   0.348807  -2.046              0.040771 *
CD.Account1          3.545241   0.403945   8.777 < 0.0000000000000002 ***
Online1             -0.817114   0.201944  -4.046      0.0000520466362 ***
CreditCard1         -0.886710   0.258336  -3.432              0.000598 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2163.69  on 3499  degrees of freedom
Residual deviance:  794.07  on 3488  degrees of freedom
AIC: 818.07

Number of Fisher Scoring iterations: 8
```

Model parameters and coefficient values for the Stepwise model from Objective 1.

6.17 Outlier Analysis Plots



Cook's distance algorithm (Left) and a plot of Residuals vs Leverage (Right) showed the presence of some high leverage data points.

6.18 Standardized Residuals vs Indexes

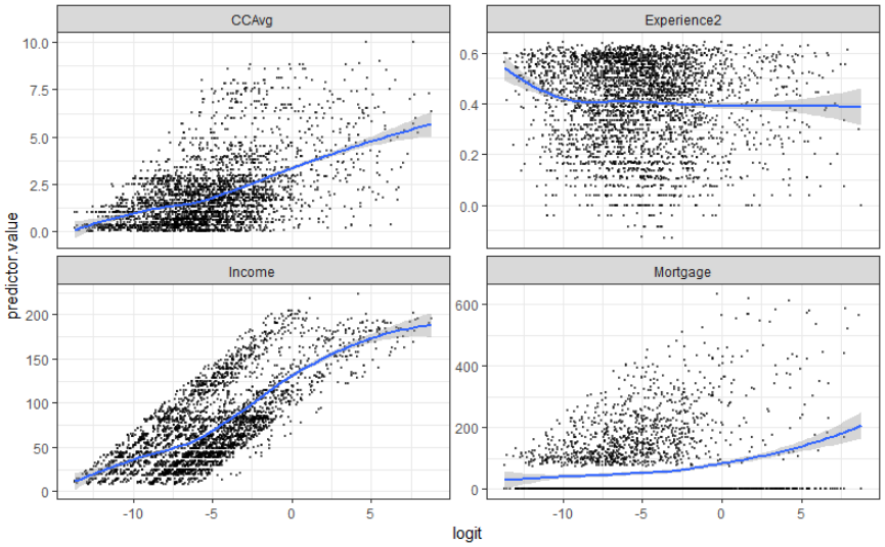
The data showed that 90% of people will not take a personal loan offer, and 10% will, we can see this reflected by this standardized residuals versus index plot.

6.19 VIF Values from Objective 1 Model Predictors

	GVIF	Df	GVIF^(1/(2*Df))
Income	2.940809	1	1.714879
Family	1.529409	3	1.073381
CCAvg	1.516750	1	1.231564
Education	2.323075	2	1.234570
Securities.Account	1.291648	1	1.136507
CD.Account	1.936714	1	1.391659
Online	1.143566	1	1.069376
CreditCard	1.383602	1	1.176266

VIF analysis of the winning stepwise model, showing that there is no multicollinearity present in the model.

6.20 Scatter Plots of Predictor vs Logit Values



Checking the linear relationship between continuous predictor variables and the logit of the outcome by visually inspecting the scatter plots between each predictor and the logit values.

6.21 Odds Ratios with Confidence Intervals

	Odds Ratio	2.5 %	97.5 %
(Intercept)	" 0.0000035533836"	" 0.0000009670612"	" 0.0000130566048"
Income	" 1.0658871357374"	" 1.0581505454691"	" 1.0736802915192"
Family2	" 0.8137023043692"	" 0.4687078501608"	" 1.4126314289566"
Family3	" 6.8336324635477"	" 3.8284108534361"	" 12.1978895250861"
Family4	" 4.0419826931018"	" 2.2868175478207"	" 7.1442621677030"
CCAvg	" 1.1351775688223"	" 1.0189519542112"	" 1.2646603281260"
Education2	" 55.5639227354387"	" 28.8653266780218"	"106.9570264763547"
Education3	" 65.1860299588602"	" 33.9118239040492"	"125.3019747277599"
Securities.Account1	" 0.4898738322890"	" 0.2472743352565"	" 0.9704863681571"
CD.Account1	" 34.6480306876299"	" 15.6978347366024"	" 76.4746253654833"
Online1	" 0.4417045599866"	" 0.2973287142396"	" 0.6561859281298"
CreditCard1	" 0.4120091864564"	" 0.2483194731810"	" 0.6836015216606"

Odds ratios and their confidence intervals for the best performing stepwise model from Objective 1.

7.0 RMD code

````
---
title: "LoanLogisticRegModel"
author: "Laura Ahumada, Erin McClure-Price, Duy Nguyen"
date: ""
output: "pdf_document"
editor_options:
  chunk_output_type: inline
---

```{r, setup, include=FALSE}
knitr::opts_chunk$set(warning = FALSE, message = FALSE, comment=NA)
```
```{r, echo=FALSE, warning=FALSE, message=FALSE}
library(ROCR)
library(glmnet)
#Load Libraries
library(Hmisc)
library(gridExtra)
library(dplyr)
library(tidyr)
library(ggplot2)
library(ggthemes)
library(class)
library(caret)
theme_set(theme_classic()) #Set the theme for plots
theme_update(plot.title = element_text(hjust = 0.5))
options(scipen=999)

```
````

EDA
Looking into the data

```{r, include=FALSE}
PersonalLoan=read.csv("~/Documents/Statistics 2/StatsProject2/Statistcs2-project-2/Bank_Personal_Loan_Modelling.csv")

# 14 attributes
dim(PersonalLoan)[2]
print("There are 14 Columns")
# 5000 data size
nrow(PersonalLoan)
print("There 5000 entries")
# no duplicates
any(duplicated(PersonalLoan))
print("There are no duplicates")

#No missing values
print("No missing values")
(sapply(PersonalLoan, function(x) all_miss=sum(is.na(x))))

attach(PersonalLoan)
head(PersonalLoan)
```

Statistics

+ ID can be dropped since it is not a useful predictor and just a unique identifier and will affect the results of the model
+ Age mean is 45 years old, 23 has is the lowest and 67 is the highest. There seems to be a normal distribution.
+ ***There are negative values in Experience which is odd because there can't be negative years of experience. This will be looked at next***.
+ The mean income is 73 while the median is 64 showing skewness which seems normal representation of society. The minimum income is 8,000 and the max is 224,000. These are common Salaries
+ ZIP.Code should not be numberic they should be changed to categorical. There are 467 distinct ZIP codes
+ There seems to be an equal distribution for families size 1,2,3,4. Each are compose of about 25%
+ The CCAVG, Average Spending per 1000 goes from 0 to 10,000, however the median is 1,5000. Here we can see that there are outliers.
+ For Education, 41% have have up to highschool, 28% up to under grad studies and interestingly 30% have up to grad school. That seems like a reasonable distribution.
+ For mortgage we can see how skewed it is, the mean is 56.5 and the median is 0. Showing the extreme outliers. **This needs to be looked at**
+ For the target variable, personal loan, we can see quite a difference, 90% without a loan and only 10% with loan. This makes sense because not many people take loan from banks

+ Securities account also has a big difference where 90% does not have a security account while 10% does
+ For CD.Account (Ceritificate deposit) once again we see the 94% does not have it while 6% has it.
+ As for online (online banking capability), 40% doe not have it while 60% has it. That distribution is a little more balanced and makes sense.
+ As for Credit card 70% does not have it while 30% has it.

```{r, warning = F, echo=FALSE}
summary(PersonalLoan)
```

Looking into questions obtained in the statistical analysis

+ ***Looking at the entries with negative experience***
+ The 52 people with negative total experience are between 23 to 29 years old and salary median of 65,000. Sound like these could just young people that just started working. We can change their negative values to 1 since they have an income, thus are are working.

```{r, echo=FALSE}
library(dplyr)
#Verifying who are the people with negative experience
summary(PersonalLoan %>% select (Age,Income,Personal.Loan) %>% filter(Experience<0) )

#Updating values
PersonalLoan$Experience[PersonalLoan$Experience<0]=1
```

```{r, include=FALSE}
# Creating graphs because we will change the variables to categories and this won't work
PersonalLoan = PersonalLoan %>% mutate(Experience2 = Experience/Age)

a=ggplot(PersonalLoan,aes(x=Income,y=Personal.Loan))+geom_point()+
  geom_smooth(method="loess",size=1,span=2)+
  ylim(-.2,1.2)

b=ggplot(PersonalLoan,aes(x=CCAvg,y=Personal.Loan))+geom_point()+
  geom_smooth(method="loess",size=1,span=2)+
  ylim(-.2,1.2)


c=ggplot(PersonalLoan,aes(x=(Mortgage),y=Personal.Loan))+geom_point()+
  geom_smooth(method="loess",size=1,span=2)+
  ylim(-.2,1.2)

d=ggplot(PersonalLoan,aes(x=Experience2,y=Personal.Loan))+geom_point()+
```

```
  geom_smooth(method="loess",size=1,span=2)+
  ylim(-.2,1.2)

e=ggplot(PersonalLoan,aes(x=Income,y=Personal.Loan))+geom_point()+
  geom_smooth(method="loess",size=1,span=2)+
  ylim(-.2,1.2)
f=ggplot(PersonalLoan,aes(x=Education,y=Personal.Loan))+geom_point()+
  geom_smooth(method="loess",size=1,span=2)+
  ylim(-.2,1.2)

#ggplot(PersonalL,aes(x=Online,y=Personal.Loan))+geom_point()+
geom_smooth(method="loess",size=1,span=2)+ facet_grid(~Personal.Loan) +

#ggplot(PersonalL,aes(x=CreditCard,y=Personal.Loan))+geom_point()+
geom_smooth(method="loess",size=1,span=2)+ facet_grid(~Personal.Loan)

```

+ Setting categories as factors
+ Creating a new variable called Exprience 2 due to Age and experience having a correlation of 97%. That way we can get rid of Age and Experience

```{r,echo=FALSE}
# Identification Columns (ID and ZIP.Code)
# Droppingzip and ID as there are too many unique values and will affect the model
PersonalLoan = PersonalLoan[-c(1,5)]

# making sure yes is our targer variable
PersonalLoan$Personal.Loan<-
factor(ifelse(PersonalLoan$Personal.Loan==1,"Yes","No"),levels=c("No","Yes"))  #last level is the success

# To categorical
factor_vars = c("Family", "Education",
                "Securities.Account", "CD.Account", "Online", "CreditCard")
PersonalLoan[factor_vars] = lapply(PersonalLoan[factor_vars], as.factor)

# ALREADY CREATED TO DO PLOTS
#PersonalLoan = PersonalLoan %>% mutate(Experience2 = Experience/Age)

# getting rid of Age and Experience
PersonalLoan = PersonalLoan[-c(1,2)]
print(describe(PersonalLoan))
```

+ ***Checking mortgage distribution***
+ Mortgage may need to be logged as it is very skewed.

```{r, echo=FALSE}
```

```
a1=PersonalLoan %>% ggplot(aes(x=(Mortgage), color="blue")) + geom_boxplot()+
theme_classic()+ theme(legend.position="none") + ggtitle("Distribution of Mortgage")

b1=PersonalLoan %>% ggplot(aes(x=log(Mortgage), color="blue")) + geom_boxplot()+
theme_classic() +theme(legend.position="none")+ ggtitle("Distribution of Mortgage logged")
grid.arrange(a1,b1, ncol=2)
```

+ Creating another data set with Mortgage logged since logging did improve the distribution. The zero's were replaced to with 1 in order to not get infitiy when logging. Also, removing age since it has a 99% correlation with Experience, therefore only one is needed

```{r}
#Creating a new data set with the modified attributes
PersonalLoan2=mutate(PersonalLoan)
#Updating values
PersonalLoan2$Mortgage[PersonalLoan2$Mortgage==0]=1
PersonalLoan2$MortgageLogged=log(PersonalLoan2$Mortgage)
#PersonalLoan2=PersonalLoan2 %>% select(-Mortgage)

```

Relationships and correlations

+ Experience and Age has a Correlation of 99%.Too high. However, they seem to have no relationship with Loan as both; yes loan and no loan, have the same correlation with experience and age of 0.99
+ CCAvg and Income has a 64% correlations and it does seems to have a relationship with Loan since there is 0.62 with no loan and .02 with yes loan.
+ The rest of the explanatory variables do not seem to have relationship between each other

```{r echo=FALSE}
library(GGally)
ggpairs(PersonalLoan,columns = c(1,2,3,5,6,7,8),mapping=aes(colour=Personal.Loan))

library(corrplot)
numData=PersonalLoan %>%
select(colnames(PersonalLoan)[!grepl('factor|logical|character',sapply(PersonalLoan,class))])
corrplot(cor(numData), type = "upper", order = "hclust",
         tl.col = "black", sig.level = 0.05,insig = "blank")
```

Checking realtionship between Loan (response variable) and the rest of the predictors

Very high Relationship present
+ ***Income***, the more income the more changes to get a loan

+ ***Mortgage***, people high mortgages seem to ask for loans more so than those with lower mortgages

Relationship present
+ ***CD.Account*** (certificate of deposit) seem to have a relationship with Personal Loan. Those with personal loan tend to have CD.Account more so than those with no CD.Account
+ ***Personal education***, people with up to highschool tend do not get loans as much as does with an education of higher than highschool
+ !!**there seems to be relationship with loan and education but that of education 2 and 3 seem to be very close so maybe making 2 and 3 one single category would be more helpful**!!

Very small relationship
+ ***Security account*** and Personal Loan seem to have slight relationship

No relationship with response
+ ***Online and Credit card*** don't seem to have a relationship

Looking closer at each realationship to see if anything was missed

```{r, warning=FALSE, message=FALSE, echo=FALSE}
#+ ***Checking relationships closer with loess***
grid.arrange(a,b,c,d, ncol=2)
grid.arrange(e,f, ncol=2)

```

+ More detailed graph on each variable with response

```{r, echo=FALSE, warning=FALSE}
library(GGally)
library(car)
#ggpairs(numData, aes(alpha = 0.4))
#pairs(lifeExp[,-c(1,3)],col=Status)


# This determines green is yes.
par(mfrow=c(2,3))
plot(Personal.Loan~.,data=PersonalLoan, col= c("pink","lightblue"))

```



```{r, include=FALSE}
# Looking deeper at relationship between Loan and Securities Account
# There actually is a relationship
#Loan vs securites
a=ggplot(PersonalLoan, aes(x=Securities.Account, fill=Personal.Loan)) +
```

geom_bar(position="dodge")+theme(legend.position="none")+ ggtitle("Security account vs Loan")

b=ggplot(PersonalLoan, aes(x=Online, fill=Personal.Loan)) + geom_bar(position="dodge")+theme(legend.position="none")+ ggtitle("Online access vs Loan")

c=ggplot(PersonalLoan, aes(x=CreditCard, fill=Personal.Loan)) + geom_bar(position="dodge")+theme(legend.position="none")+ ggtitle("Credit Card vs Loan")

grid.arrange(a,b,c, ncol=2)

```

# Model: Objective 1

```{r}
# Train Test Split
set.seed(123)
index<-sample(1:dim(PersonalLoan)[1],round(.70 * dim(PersonalLoan)[1]))
train<-PersonalLoan[index,]
test<-PersonalLoan[-index,]

# Split Predict for lasso
dat.test.x = model.matrix(Personal.Loan ~ Income + Family + CCAvg-1 + Education + Securities.Account-1 + CD.Account-1 + Online-1 + CreditCard -1 + Experience2, test)

dat.train.x = model.matrix(Personal.Loan ~ Income + Family + CCAvg-1 + Education + Securities.Account-1 + CD.Account-1 + Online-1 + CreditCard -1 + Experience2, train)
dat.train.y = train$Personal.Loan

```

Performing model with all variables, some feature selection methods (forward, stepwise, LASSO) and another based on EDA
+ With the full model with all attributes it showed that the only important were Income, Family,CCavg, Education,Securites.Account, CD.Account, Online, CreditCard

+ Once Stepwise was added to the full model it selected all of those that appeared as significant in the full model:Income, Family, CCAvg, Education, Securities.Account, CD.Account, Online, CreditCard. It also selected Experience2 but that one was not significant as we had seen in the EDA.

+ When the forward model was added to the full model it selected the same thing as stepwise but included CCAVG which was significant and Mortgage which was not significant .

+ As for LASSO it selected all of the attributes by Stepwise and included CCAvg.
```

```{r, echo=FALSE, warning=FALSE, message=FALSE}
# FUNCTIONS TO PREDICT RESULTS

library(MASS)
library(tidyverse)
library(car)

results <- function(model,bool){

  if (bool==FALSE){
    AIC_result=round((AIC(model)),3)
    BIC_result=round((BIC(model)),3)
    fit.pred<-predict(model,newdata=test,type="response")
    class.model<-factor(ifelse(fit.pred>0.3,"Yes","No"),levels=c("No","Yes"))
    accuracy=round((mean(class.model==test$Personal.Loan)),3)
    Sensitivitylis=round((sensitivity(class.model, test$Personal.Loan)),3)
    specificitlis= round((specificity(class.model, test$Personal.Loan)),3)
  }else{
    AIC_result=0
    BIC_result=0
    fit.pred.lasso = predict(model, newx = dat.test.x, type = "response")
    class.model<-factor(ifelse(fit.pred.lasso>0.3,"Yes","No"),levels=c("No","Yes"))
    accuracy=round((mean(class.model==test$Personal.Loan)),3)
    Sensitivitylis=round((sensitivity(class.model, test$Personal.Loan)),3)
    specificitlis= round((specificity(class.model,test$Personal.Loan)),3)
  }
  all_=c(AIC_result,BIC_result,accuracy,Sensitivitylis,specificitlis)
  return(all_)

}


ROC_predict <- function(model, bool){
  fit.pred<-predict(model,newdata=test,type="response")
  results.model<-prediction(fit.pred, test$Personal.Loan,label.ordering=c("No","Yes"))
  roc.model= performance(results.model, measure = "tpr", x.measure = "fpr")
  return(roc.model)

}

```



```{r, echo=FALSE, warning=FALSE, message=FALSE}

# RUNNING ALL MODELS
```

```
all_results=data.frame(Criterion=c("AIC","BIC","Accuracy","Sensitivity", "Specificity"))

#FULL_MODEL
simpleLG<-glm(Personal.Loan~.,family="binomial",data=train)
all_results$Full_Model=results(simpleLG,FALSE)
roc_full=ROC_predict(simpleLG)
summary(simpleLG)


# MODEL SELECTION

#STEP_WISE
stepWiseAIC<-simpleLG %>% stepAIC(trace=FALSE, direction = "both")
all_results$Step_Wise=results(stepWiseAIC,FALSE)
roc_stepWiseAIC=ROC_predict(stepWiseAIC)

# FORWARD
ForwardModel <- step(simpleLG, direction = "forward", trace = FALSE)
all_results$Forward_Model=results(ForwardModel,FALSE)
roc_ForwardModel=ROC_predict(ForwardModel)

#########################################################

# LASSO
cvfit = cv.glmnet(dat.train.x, dat.train.y, family = "binomial", type.measure = "class", nlambda =
1000)
#coef(cvfit, s = "lambda.min")
# Optimal penalty
cvfit$lambda.min

# Having found the optinal penalty we creat the mode
# For final model predictions go ahead and refit lasso using entire data set
LASSOmodel<-glmnet(dat.train.x, dat.train.y, family = "binomial",lambda=cvfit$lambda.min)
#CF <- as.matrix(coef(LASSOmodel, LASSOmodel$lambda.1se))
#CF[CF!=0,]
#coef(LASSOmodel)

#Predict
all_results$LASSO_model=results(LASSOmodel,TRUE)

#predict
fit.pred.lasso <- predict(LASSOmodel, newx = dat.test.x, type = "response")
results.lasso<-prediction(fit.pred.lasso, test$Personal.Loan,label.ordering=c("No","Yes"))
roc.lasso = performance(results.lasso, measure = "tpr", x.measure = "fpr")


lassoAIC<-glm(Personal.Loan~Income+Family+CCAvg+Education+
Securities.Account+CD.Account+ Online+ CreditCard,family="binomial",data=train)
#results(lassoAIC,FALSE)
#########################################################
```

```
#MODEL BASED ON INTUTION
intuition = glm(formula = Personal.Loan ~ CreditCard + Family + CD.Account + Education +
Income,
          family = "binomial", data = train)
all_results$Intuition=results(intuition,FALSE)
roc_Intuition=ROC_predict(intuition)


eda=glm(Personal.Loan~Income+ CD.Account+Mortgage+ Education + Family+ CD.Account,
family = "binomial", data = train)
all_results$EDA=results(eda,FALSE)
roc_EDA=ROC_predict(eda)
```

+ Using different threholds
```{r, echo=FALSE, warning=FALSE, message=FALSE}
threshold <- function(model,bool){

  if (bool==FALSE){
    fit.pred<-predict(model,newdata=test,type="response")
  }else{
     #for lasso
    fit.pred = predict(model, newx = dat.test.x, type = "response")
  }

  ModelList=c()
  limits=c(0.3,0.5,0.7)
  for (i in  limits){
      lim=i
      class.model<-factor(ifelse(fit.pred>i,"Yes","No"),levels=c("No","Yes"))
      #confusionMatrix(class.model, test$Personal.Loan)
      accuracy=mean(class.model==test$Personal.Loan)
      Sensitivity=(sensitivity(class.model, test$Personal.Loan))
      specificit= (specificity(class.model,test$Personal.Loan))
      ModelList=append(ModelList,list(c(lim,accuracy,Sensitivity,specificit)))
  }

  return(ModelList)

}
```

Changing the threshold
```{r, echo=FALSE}
# Comparing stepwise with different threashold
saving=threshold(simpleLG,FALSE)
names(saving) <- c("0.3", "0.5", "0.7")
print("All_Attributes")
print("Threashhold | Accuracy| Sensitivity| Specificy")
```

```
print(saving)
print("--------------------------------")

# Comparing simpleLG with different threashold
saving=threshold(stepWiseAIC,FALSE)
names(saving) <- c("0.3", "0.5", "0.7")
print("StepWiseAIC")
print("Threashhold | Accuracy| Sensitivity| Specificy")
print(saving)
print("--------------------------------")

# Comparing ,ForwardModel with different threashold
saving=threshold(ForwardModel,FALSE)
names(saving) <- c("0.3", "0.5", "0.7")
print("Threashhold | Accuracy| Sensitivity| Specificy")
print("ForwardModel")
print(saving)
print("--------------------------------")

# Comparing LASSOmodel with different threashold
saving=threshold(LASSOmodel,TRUE)
names(saving) <- c("0.3", "0.5", "0.7")
print("Threashhold | Accuracy| Sensitivity| Specificy")
print("LASSO")
print(saving)
print("--------------------------------")

# Comparingintuition,eda with different threashold
saving=threshold(intuition,FALSE)
names(saving) <- c("0.3", "0.5", "0.7")
print("Intuition")
print("Threashhold | Accuracy| Sensitivity| Specificy")
print(saving)
print("--------------------------------")

# Comparing eda with different threashold
saving=threshold(eda,FALSE)
names(saving) <- c("0.3", "0.5", "0.7")
print("EDA")
print("Threashhold | Accuracy| Sensitivity| Specificy")
print(saving)
```

Choosing 0.3 threshold based of the threashold results.

Criterion Comparison of all models

```{r}
all_results
```

The ROC of models

```{r, echo=FALSE}
plot(roc.lasso)
plot(roc_stepWiseAIC, col="red", add=TRUE)
plot(roc_ForwardModel,col="orange", add = TRUE)
plot(roc_Intuition,col="blue",add=TRUE)
plot(roc_full,col="yellow",add=TRUE)
plot(roc_full,col="green",add=TRUE)
legend("bottomright",legend=c("Lasso", "Stepwise","Forward","Intuition","EDA", "All
Variables"),col=c("black","red","orange","blue","yellow","green"),lty=1,lwd=1)
abline(a=0, b= 1)


```


Verify Proportions in test and train manually
+ Distribution in train and test do represent that of the whole data
```{r, echo=FALSE}
# Holding the upcoming predictions accountable
# all propotions match
print("All data")
prop.table(table(PersonalLoan$Personal.Loan))
print("Train")
prop.table(table(train$Personal.Loan))
print("Test")
prop.table(table(test$Personal.Loan))

# This means that,
# it is preffered that our predictions are 90% no loan and 10% yes loan.
```

+ Assumptions via PLOTS of selected model, Stepwise and checking VIF
+ Plots look normal and there seems to be multicolinarity among variables based on VIF
```{r, echo=FALSE}
par(mfrow = c(1, 2))
#Cook's Distance Plot
plot(stepWiseAIC, 4)
#Standardized Residuals vs Leverage
plot(stepWiseAIC, 5)
par(mfrow = c(1, 1))
# vifs
vif(stepWiseAIC)

```

Conclusion from Part 1
+ The best model was step setting the threshold to 0.3
it gave an sensitivity of 94 and specificity of 72
+ Due to the imbalance of amount of people with loan and without loan we do see we do see that the model favors no loan due to it but 72 compared to the 55 specificity was a great increase. This model is about trying to predict those who will say yes to Loan therefore Specificity is important.
+ The attributes found useful were : Income, Family,CCavg, Education,Securites.Account, CD.Account, Online, CreditCard
+ The threshold was set to 0.3 and it lead to a Sensitivity of 0.96 and specificy of 0.71
+ These were variables seen in the EDA as related to the loan.
+ Coefficients results:
For every unit increase in income the odd of getting a loan are e^1.06 times higher
For every unit increase in Family the odd of getting a loan are e^0.698209 times higher
For every unit increase in CCAvg the odd of getting a loan are e^0.120635 times higher
For every unit increase in Education the odd of getting a loan are e^1.713690 times higher
For every unit increase in Securities.Account 1 the odd of getting a loan are e^-0.937183 times less likely
For every unit increase in CD.Account1 the odd of getting a loan are e^3.840892 times higher
For every unit increase in Online1 the odd of getting a loan are e^-0.673230 times less likely
For every unit increase in CreditCard1 the odd of getting a loan are e^ -1.122701 times higher

##3

MODEL: Part 2

+ Checking graphs for interactions
+ Across all of the plots only one age with mortgage showed to maybe have an interaction
+
```{r, echo=FALSE}

library(sjPlot)    #For effect plotting
library(sjmisc)    #For effect plotting
library(ResourceSelection)  #Hosmer Lemeshow test
#names(PersonalLoan)

PersonalL=read.csv("~/Documents/Statistics 2/StatsProject2/Statistcs2-project-2/Bank_Personal_Loan_Modelling.csv")

a=ggplot(PersonalL,aes(x=Age,y=Personal.Loan,colour=Education))+geom_point()+
  geom_smooth(method="loess",size=1,span=1.5)+
  ylim(-.2,1.2)+
  facet_wrap(~Education)
```

```
b=ggplot(PersonalL,aes(x=Age,y=Personal.Loan,colour=Family))+geom_point()+
  geom_smooth(method="loess",size=1,span=1.5)+
  ylim(-.2,1.2)+
  facet_wrap(~Family)

c=ggplot(PersonalL,aes(x=Age,y=Personal.Loan,colour=CCAvg))+geom_point()+
  geom_smooth(method="loess",size=1,span=1.5)+
  ylim(-.2,1.2)+
  facet_wrap(~CCAvg)

d=ggplot(PersonalL,aes(x=Age,y=Personal.Loan,colour=Online))+geom_point()+
  geom_smooth(method="loess",size=1,span=1.5)+
  ylim(-.2,1.2)+
  facet_wrap(~Online)

e=ggplot(PersonalL,aes(x=Age,y=Personal.Loan,colour=CreditCard))+geom_point()+
  geom_smooth(method="loess",size=1,span=1.5)+
  ylim(-.2,1.2)+
  facet_wrap(~CreditCard)

f=ggplot(PersonalL,aes(x=Age,y=Personal.Loan,colour=Securities.Account))+geom_point()+
  geom_smooth(method="loess",size=1,span=1.5)+
  ylim(-.2,1.2)+
  facet_wrap(~Securities.Account)


grid.arrange(a,f,e,d, ncol=2)

#Education,Family,CCAvg,Online,CreditCard, Securities.Account
a=ggplot(PersonalL,aes(x=Income,y=Personal.Loan,colour=Education))+geom_point()+
  geom_smooth(method="loess",size=1,span=1.5)+
  ylim(-.2,1.2)+
  facet_wrap(~Education)

b0=ggplot(PersonalL,aes(x=Income,y=Personal.Loan,colour=Family))+geom_point()+
  geom_smooth(method="loess",size=1,span=1.5)+
  ylim(-.2,1.2)+
  facet_wrap(~Family)

c=ggplot(PersonalL,aes(x=Income,y=Personal.Loan,colour=CCAvg))+geom_point()+
  geom_smooth(method="loess",size=1,span=1.5)+
  ylim(-.2,1.2)+
  facet_wrap(~CCAvg)

d=ggplot(PersonalL,aes(x=Income,y=Personal.Loan,colour=Online))+geom_point()+
  geom_smooth(method="loess",size=1,span=1.5)+
  ylim(-.2,1.2)+
  facet_wrap(~Online)

e=ggplot(PersonalL,aes(x=Income,y=Personal.Loan,colour=CreditCard))+geom_point()+
```

```
  geom_smooth(method="loess",size=1,span=1.5)+
  ylim(-.2,1.2)+
  facet_wrap(~CreditCard)

f=ggplot(PersonalL,aes(x=Income,y=Personal.Loan,colour=Securities.Account))+geom_point(
)+
  geom_smooth(method="loess",size=1,span=1.5)+
  ylim(-.2,1.2)+
  facet_wrap(~Securities.Account)


grid.arrange(a,f,e,d, ncol=2)

#Education,Family,CCAvg,Online,CreditCard, Securities.Account
a=ggplot(PersonalL,aes(x=Mortgage,y=Personal.Loan,colour=Education))+geom_point()+
  geom_smooth(method="loess",size=1,span=1.5)+
  ylim(-.2,1.2)+
  facet_wrap(~Education)

b1=ggplot(PersonalL,aes(x=Mortgage,y=Personal.Loan,colour=Family))+geom_point()+
  geom_smooth(method="loess",size=1,span=1.5)+
  ylim(-.2,1.2)+
  facet_wrap(~Family)

c=ggplot(PersonalL,aes(x=Mortgage,y=Personal.Loan,colour=CCAvg))+geom_point()+
  geom_smooth(method="loess",size=1,span=1.5)+
  ylim(-.2,1.2)+
  facet_wrap(~CCAvg)

d=ggplot(PersonalL,aes(x=Mortgage,y=Personal.Loan,colour=Online))+geom_point()+
  geom_smooth(method="loess",size=1,span=1.5)+
  ylim(-.2,1.2)+
  facet_wrap(~Online)

e=ggplot(PersonalL,aes(x=Mortgage,y=Personal.Loan,colour=CreditCard))+geom_point()+
  geom_smooth(method="loess",size=1,span=1.5)+
  ylim(-.2,1.2)+
  facet_wrap(~CreditCard)

f=ggplot(PersonalL,aes(x=Mortgage,y=Personal.Loan,colour=Securities.Account))+geom_poi
nt()+
  geom_smooth(method="loess",size=1,span=1.5)+
  ylim(-.2,1.2)+
  facet_wrap(~Securities.Account)

grid.arrange(a,f,e,d, ncol=2)

#Education,Family,CCAvg,Online,CreditCard, Securities.Account
a=ggplot(PersonalL,aes(x=CCAvg,y=Personal.Loan,colour=Education))+geom_point()+
  geom_smooth(method="loess",size=1,span=1.5)+
```

```
  ylim(-.2,1.2)+
  facet_wrap(~Education)

b2=ggplot(PersonalL,aes(x=CCAvg,y=Personal.Loan,colour=Family))+geom_point()+
  geom_smooth(method="loess",size=1,span=1.5)+
  ylim(-.2,1.2)+
  facet_wrap(~Family)

c=ggplot(PersonalL,aes(x=CCAvg,y=Personal.Loan,colour=CCAvg))+geom_point()+
  geom_smooth(method="loess",size=1,span=1.5)+
  ylim(-.2,1.2)+
  facet_wrap(~CCAvg)

d=ggplot(PersonalL,aes(x=CCAvg,y=Personal.Loan,colour=Online))+geom_point()+
  geom_smooth(method="loess",size=1,span=1.5)+
  ylim(-.2,1.2)+
  facet_wrap(~Online)

e=ggplot(PersonalL,aes(x=CCAvg,y=Personal.Loan,colour=CreditCard))+geom_point()+
  geom_smooth(method="loess",size=1,span=1.5)+
  ylim(-.2,1.2)+
  facet_wrap(~CreditCard)

f=ggplot(PersonalL,aes(x=CCAvg,y=Personal.Loan,colour=Securities.Account))+geom_point(
)+
  geom_smooth(method="loess",size=1,span=1.5)+
  ylim(-.2,1.2)+
  facet_wrap(~Securities.Account)

grid.arrange(a,f,e,d, ncol=2)
grid.arrange(b,b0,b1,b2, ncol=2)

plot_model(simpleLG,type="pred",terms=c("Education","Family","Online"))

plot_model(simpleLG,type="pred",terms=c("Education","Family","Income"))

plot_model(simpleLG,type="pred",terms=c("Education","CCAvg","Income"))

#Education income personal loan
#CCavg and education
#Mortgage and Credit Card
#Education and Mortgage
#Education and CCAvg
#income and family
#Family and CCAvg
#Family and income
#mortgage and family
```

+ **Running the models**
+ In part 1 we saw that, Income, Family, CCAvg, CD.Account, Education and Credit Card were significant, therefore we leverage those varibles and kept mortgage (just like in part 1) based on the EDA analysis, for Part 2

Interaction

+ All variables:Income, Family, CCAvg, CD.Account, Education, and Credit Card and the in the interaction of family and mortgage were significant.
+ When passing the model model through stepwise and forward it kept the all variables including the interaction and thus kept the same significant variables showing signs that the interaction is indeed useful.
+ When Anova was applied, it did show the interaction as significant
+ The Hoslem test however showed that the model was a poor fit, yes again this is not reliable due to the size of the data (This I asked in class and Dr.Turner said we couldn't relly on this metric with big data)

Logged
+ We logged morgage
+ All variables were significant: Income, Family, CCAvg, CD.Account, Education, and Credit Card including the log mortgage and log income
+ Once the models were mixed, having logged income and the interaction of logged mortgage with family lead to the interaction no longer being significant however all other variable remained significant.

Polynomials
+ When income was set to poly 2 all variables were significant
+ The income was set to 3 polys, the first poly income was significant but the rest were no longer significant. However, the rest of the variables are still significant
+ CCAvg set to poly2 showed all variables as significant just like poly 3
+ when setting both CCavg and income to poly 2 all variables showed importance however when setting both CCAvg and income to poly 3, income ^2 and income^3 did not show significance in the model while all the rest of the attributes did show significance.
+ Interesting when MortagaLogged or mortgage was added it reduced the specificity and sensitivity
+ Therefore the attributes used where income, Family, CCAvg, CD.Account, Education, Credit Card . we brought up Securites.Account agian as well as online since we had seem that they had importance in some models and EDA. Of course the difference within the models was having Incompo Poly or CCAvg or both to ^2 and ^3 polynomial.

LDA and QDA
+ LDA mode and QDA model did poorly however between the two, LDA model out performed QDA. This makes sense since the groups have a clear division of who gets loans and who doesn't which would be LDA.

```{r, echo=FALSE}
```

```
########## Interaction and LOGS ######################

# Train Test Split
index<-sample(1:dim(PersonalLoan2)[1],round(.70 * dim(PersonalLoan2)[1]))
train<-PersonalLoan2[index,]
test<-PersonalLoan2[-index,]

# RUNNING ALL MODELS
#setting up the data frame for criterion
all_results2=data.frame(Criterion=c("AIC","BIC","Accuracy","Sensitivity", "Specificity"))
all_resultsX=data.frame(Criterion=c("AIC","BIC","Accuracy","Sensitivity", "Specificity"))

#INTERACTION  Family and mortgage
Interaction1<-
glm(Personal.Loan~Securities.Account+CD.Account+CreditCard+Education+Income+CCAvg
+Family*Mortgage,data=train,family="binomial")   #complex
#summary(ComplexInteractionl)
all_resultsX$Famxmortgage=results(Interaction1,FALSE)
roc_interaction=ROC_predict(Interaction1)


#INTERACTION with educationMORTGAGE
Interactionl2<-
glm(Personal.Loan~Securities.Account+CD.Account+CreditCard+Education+Income+CCAvg
+Education*Mortgage,data=train,family="binomial")   #complex
#summary(ComplexInteractionl)
all_resultsX$EduXMortgage=results(Interactionl2,FALSE)
roc_EducationxMorgage=ROC_predict(Interactionl2)

#INTERACTION with CCAVG&Family
Interactionl3<-
glm(Personal.Loan~Securities.Account+CD.Account+CreditCard+Education+Income+CCAvg
+Family*CCAvg,data=train,family="binomial")   #complex
#summary(ComplexInteractionl)
all_resultsX$CCAvgxFam=results(Interactionl3,FALSE)
roc_CCAvgFam=ROC_predict(Interactionl3)

#INTERACTION INCOME AND EDUCATION
Interactionl4<-
glm(Personal.Loan~Securities.Account+CD.Account+CreditCard+Family+Income*Education+
CCAvg+CCAvg,data=train,family="binomial")   #complex
#summary(ComplexInteractionl)
all_resultsX$IncomeXedu=results(Interactionl4,FALSE)
roc_incomeXedu=ROC_predict(Interactionl4)

#INTERACTION  Family and income
Interactionl5<-
glm(Personal.Loan~Securities.Account+CD.Account+CreditCard+Education+CCAvg+Income*
Family+Mortgage,data=train,family="binomial")   #complex
```

```
#summary(ComplexInteractionl)
all_resultsX$FamxIncome=results(Interactionl5,FALSE)
roc_IncomeFamily=ROC_predict(Interactionl5)

#INTERACTION all
ComplexInteractionl<-
glm(Personal.Loan~Securities.Account+CD.Account+CreditCard+Education*Mortgage+
                Family*Income+CCAvg*Family,data=train,family="binomial")   #complex
#summary(ComplexInteractionl)
all_results2$AllInteractions=results(ComplexInteractionl,FALSE)
all_resultsX$AllInteractions=results(ComplexInteractionl,FALSE)
roc_AllInteraction=ROC_predict(ComplexInteractionl)

```


```{r, include=FALSE}

#MODEL SELCT TO FILTER

#Performing stepwise to see if it keeps the interaction
stepComplexinteraction<-ComplexInteractionl %>% stepAIC(trace=FALSE, direction="both")
  #Anova(stepComplexinteraction,type=3)
  #hoslem.test(stepComplexinteraction$y,fitted(stepComplexinteraction))
#all_results2$stepInteraction=results(stepComplexinteraction,FALSE)
roc_interactionStepwise=ROC_predict(stepComplexinteraction)

interaction_Forward <- step(ComplexInteractionl, direction = "forward", trace = FALSE)
  #summary(interaction_Forward)
  #Anova(interaction_Forward,type=3)
  #hoslem.test(interaction_Forward$y,fitted(interaction_Forward))
#all_results2$InteractionForward=results(interaction_Forward,FALSE)
roc_interactionForward=ROC_predict(interaction_Forward)
```

```{r, echo=FALSE}
##########################################

# LOGGING INOCME AND USING  MORTGAGELOGGED
Personal.Loan~Securities.Account+CD.Account+CreditCard+Education+Income+CCAvg+Fa
mily*Mortgage

ComplexLog<-glm(Personal.Loan~Education+log(Income)+CCAvg+ Family+
MortgageLogged+Securities.Account+CD.Account+CreditCard,data=train,family="binomial")
#complex
#stepComplexlog<-ComplexLog %>% stepAIC(trace=FALSE) # Kept all variables
all_results2$LogIncomeMortgage=results(ComplexLog,FALSE)
roc_logIncomeMortgage=ROC_predict(ComplexLog)

# LOGGING AND INTERACTION
```

```
ComplexLogInteraction<-
glm(Personal.Loan~MortgageLogged*Education+log(Income)*Family+CCAvg*
Family*MortgageLogged+Securities.Account+CD.Account+CreditCard,data=train,family="bino
mial")   #complex
#stepComplexLoginteraction<-ComplexLogInteraction %>% stepAIC(trace=FALSE)
#Anova(ComplexLogInteraction,type=3)
#hoslem.test(stepComplexLoginteraction$y,fitted(stepComplexLoginteraction))
all_results2$LogAndInteraction=results(ComplexLogInteraction,FALSE)
roc_logAndInteraction=ROC_predict(ComplexLogInteraction)
all_results2


```


```{r, echo=FALSE}
########## LDA and QDA ######################

## Scaling
train_lda=mutate(train)
train_lda$income=scale(train[c(1)])
train_lda$CCAvg=scale(train[c(3)])
train_lda$Mortgage=scale(train[c(5)])

test_lda=mutate(test)
test_lda$income=scale(test[c(1)])
test_lda$CCAvg=scale(test[c(3)])
test_lda$Mortgage=scale(test[c(5)])
#function to create criterion for each model
ldaqda_predict <- function(model){
  pred_model<-predict(model,newdata=test_lda)$class
  accuracy=round((mean(pred_model==test_lda$Personal.Loan)),3)
  Sensitivit=round((sensitivity(pred_model, test_lda$Personal.Loan)),3)
  specificit= round((specificity(pred_model, test_lda$Personal.Loan)),3)
  return(c(0,0,accuracy,Sensitivit,specificit))
}


#############################################################
################### LDA ######################


#Keeping only numeric variables
lda_<- lda( Personal.Loan ~ (Income) + (CCAvg)  + (Mortgage),
    data = train_lda)
all_results2$LDA=ldaqda_predict(lda_)
fit.p<-predict(lda_,newdata=test_lda)
results.model<-prediction(fit.p$posterior[,2],
test_lda$Personal.Loan,label.ordering=c("No","Yes"))
roc.lda_= performance(results.model, measure = "tpr", x.measure = "fpr")
```

```
#pred<-predict(lda_,newdata=test)$class
#plot(pred)

#LDA ALL VARIABLES
#lda_2<- lda( Personal.Loan ~ Income + CCAvg  + Family + Education +Securities.Account +
CD.Account + CreditCard + MortgageLogged + Experience2 + Online, data = train)
#lda_2 <- lda(Personal.Loan ~ ., data = train, CV = TRUE)
#all_results2$LDA_catNUm=ldaqda_predict(lda_2)

#fit.p<-predict(lda_2,newdata=test)
#results.model<-prediction(fit.p$posterior[,2],
test$Personal.Loan,label.ordering=c("No","Yes"))
#roc.lda_2= performance(results.model, measure = "tpr", x.measure = "fpr")

############################################################
################### QLA ######################

# qda_mod2 <- qda(Personal.Loan ~ ., data = train, CV = TRUE)
qda_1<- qda( Personal.Loan ~ Income + CCAvg  + Mortgage, data = train_lda)
# myldawPrior<- lda(Personal.Loan ~ X1 + X2, data = PersonalLoan2, prior = c(.20,.80))
all_results2$QDAnum=ldaqda_predict(qda_1)
fit.p<-predict(qda_1,newdata=test_lda)
results.model<-prediction(fit.p$posterior[,2],
test_lda$Personal.Loan,label.ordering=c("No","Yes"))
roc.qda_1= performance(results.model, measure = "tpr", x.measure = "fpr")

# All variables
#qda_2<- qda( Personal.Loan ~ Income + CCAvg  + Family + Education +Securities.Account
+ CD.Account + CreditCard + MortgageLogged + Experience2 + Online, data = train)
#all_results2$QDAall=ldaqda_predict(qda_2)
#fit.p<-predict(qda_2,newdata=test)
#results.model<-prediction(fit.p$posterior[,2],
test$Personal.Loan,label.ordering=c("No","Yes"))
#roc.qda_2= performance(results.model, measure = "tpr", x.measure = "fpr")


```


```{r, echo=FALSE}
# POLYNOMYAL MODELS
all_results3=data.frame(Criterion=c("AIC","BIC","Accuracy","Sensitivity", "Specificity"))

model.poly.income2 = glm(formula = Personal.Loan ~ poly(Income, 2) + Family + CCAvg +
Education + Securities.Account + CD.Account + Online + CreditCard, family = "binomial", data
= train)
all_results3$polyIncome2=results(model.poly.income2,FALSE)
roc_PolyIncome2=ROC_predict(model.poly.income2)

model.poly.income3 = glm(formula = Personal.Loan ~ poly(Income, 3) + Family + CCAvg +
```

```
Education + Securities.Account + CD.Account + Online + CreditCard, family = "binomial", data
= train)
all_results3$polyIncome3=results(model.poly.income3,FALSE)
roc_PolyIncome3=ROC_predict(model.poly.income3)

model.poly.CCAvg2 = glm(formula = Personal.Loan ~ Income + Family + poly(CCAvg, 2) +
Education + Securities.Account + CD.Account + Online + CreditCard+Mortgage, family =
"binomial", data = train)
all_results3$polyCCAvg2=results(model.poly.CCAvg2,FALSE)
roc_CCAvg2=ROC_predict(model.poly.CCAvg2)

model.poly.CCAvg3 = glm(formula = Personal.Loan ~ Income + Family + poly(CCAvg, 3) +
Education + Securities.Account + CD.Account + Online + CreditCard+Mortgage, family =
"binomial", data = train)
all_results3$polyCCAvg3=results(model.poly.CCAvg3,FALSE)
roc_CCAvg3=ROC_predict(model.poly.CCAvg3)

model.poly.both2 = glm(formula = Personal.Loan ~ poly(Income, 2) + Family + poly(CCAvg,
2) + Education + Securities.Account + CD.Account + Online + CreditCard, family = "binomial",
data = train)
all_results3$PolyBoth2=results(model.poly.both2,FALSE)
roc_PolyBoth2=ROC_predict(model.poly.both2)

model.poly.both3 = glm(formula = Personal.Loan ~ poly(Income, 3) + Family + poly(CCAvg,
3) + Education + Securities.Account + CD.Account + Online + CreditCard, family = "binomial",
data = train)
all_results3$PolyBoth3=results(model.poly.both3,FALSE)
roc_PolyBoth3=ROC_predict(model.poly.both3)


#extra model
extrememodel= glm(formula = Personal.Loan ~ poly(log(Income), 3) +
Family*MortgageLogged + poly(CCAvg, 3) + Education + Securities.Account + CD.Account +
Online + CreditCard, family = "binomial", data = train)
all_results3$ExtremexModel=results(extrememodel,FALSE)
roc_extreme=ROC_predict(model.poly.both3)

model_InteractionPoly = glm(formula = Personal.Loan ~ poly(Income, 3) + Family +
Income*Family + CCAvg + Education + Securities.Account + CD.Account + Online +
CreditCard + MortgageLogged, family = "binomial", data = train)
all_results3$InteractinPolyLog=results(model_InteractionPoly,FALSE)
roc_InteractionPoly=ROC_predict(model_InteractionPoly)

```


Criterion of best models

```{r, echo=FALSE}
```

```
all_combined=all_results %>% cbind(all_results2,all_results3,all_resultsX)
all_combined[c(1,3,11,13,16,11,29,22)]
```

+ ***The ROC of the best model from the first part vs the best models of the second set***

```{r, echo=FALSE}

#plot(roc_PolyIncome2, col = "green", xlim = c(0, 0.3), ylim = c(0.7, 1.0))
plot(roc_stepWiseAIC, col = "chartreuse", xlim = c(0, 0.3), ylim = c(0.7, 1.0))
plot(roc_logAndInteraction, col = "blueviolet", add = TRUE, xlim = c(0, 0.3), ylim = c(0.7, 1.0))
plot(roc_AllInteraction, col = "aquamarine", add = TRUE, xlim = c(0, 0.3), ylim = c(0.7, 1.0))
plot(roc_PolyIncome3, col = "blue", add = TRUE, xlim = c(0, 0.3), ylim = c(0.7, 1.0))
plot(roc_PolyBoth3, col = "coral", add = TRUE, xlim = c(0, 0.3), ylim = c(0.7, 1.0))
plot(roc_InteractionPoly,col="pink",add=TRUE, xlim = c(0, 0.3), ylim = c(0.7, 1.0))
legend("bottomright", legend = c( "StepWise", "Log&interaction", "All Interactions",
"polyIncome3", "PolyBoth3", "Poly+Interaction"),
       col = c( "chartreuse", "blueviolet", "aquamarine", "blue", "coral","pink","greenyellow"),
       lty=1, lwd=1)
abline(a=0, b= 1)


print("Closer view")
plot(roc_stepWiseAIC)
plot(roc_logAndInteraction, col = "blueviolet", add = TRUE, xlim = c(0, 0.3), ylim = c(0.7, 1.0))
plot(roc_AllInteraction, col = "aquamarine", add = TRUE, xlim = c(0, 0.3), ylim = c(0.7, 1.0))
plot(roc_PolyIncome3, col = "blue", add = TRUE, xlim = c(0, 0.3), ylim = c(0.7, 1.0))
plot(roc_PolyBoth3, col = "coral", add = TRUE, xlim = c(0, 0.3), ylim = c(0.7, 1.0))
plot(roc_InteractionPoly,col="pink",add=TRUE, xlim = c(0, 0.3), ylim = c(0.7, 1.0))
legend("bottomright", legend = c( "StepWise", "Log&interaction", "All Interactions",
"polyIncome3", "PolyBoth3", "Poly+Interaction"),
       col = c( "black", "blueviolet", "aquamarine", "blue", "coral","pink","greenyellow"),
       lty=1, lwd=1)
abline(a=0, b= 1)

```


EXTRA: The ROC of each set of models

+ ***Best of logs and interactions***

```{r, echo=FALSE}

plot(roc_interaction,col="black",xlim = c(0, 0.3), ylim = c(0.7, 1.0))
plot(roc_incomeXedu, col="red", add=TRUE,xlim = c(0, 0.3), ylim = c(0.7, 1.0))
plot(roc_EducationxMorgage, col="coral", add=TRUE,xlim = c(0, 0.3), ylim = c(0.7, 1.0))
plot(roc_CCAvgFam,col="orange", add = TRUE,xlim = c(0, 0.3), ylim = c(0.7, 1.0))
plot(roc_IncomeFamily,col="blue",add=TRUE,xlim = c(0, 0.3), ylim = c(0.7, 1.0))
plot(roc_AllInteraction,col="yellow",add=TRUE,xlim = c(0, 0.3), ylim = c(0.7, 1.0))
plot(roc_logAndInteraction,col="green",add=TRUE,xlim = c(0, 0.3), ylim = c(0.7, 1.0))
```

```
legend("bottomright",legend=c("FamilyXmortgage","IncomeXedcuation",
"EducationXMortgage","CCavgXFamily","IncomeXfamily","All INteractions","Interaction &
Log"),col=c("black","coral","red","orange","blue","yellow","green"),lty=1,lwd=1)
abline(a=0, b= 1)
```

+ ***LDA & QDA***

```{r, echo=FALSE}

plot(roc.lda_)
#plot(roc.lda_2, col="red", add=TRUE)
plot(roc.qda_1,col="orange", add = TRUE)
#plot(roc.qda_2,col="blue",add=TRUE)
legend("bottomright",legend=c("LDA_NumericOnly","QDA_Numeric"),col=c("black","orange"),l
ty=1,lwd=1)
abline(a=0, b= 1)
```

+ ***Lets get closer look at all the log and interaction , interaction with log and logging alone***
```{r, echo=FALSE}

plot(roc_logIncomeMortgage,col="black",xlim = c(0, 0.3), ylim = c(0.7, 1.0))
plot(roc_AllInteraction,col="coral",add=TRUE,xlim = c(0, 0.3), ylim = c(0.7, 1.0))
plot(roc_logAndInteraction,col="green",add=TRUE,xlim = c(0, 0.3), ylim = c(0.7, 1.0))
legend("bottomright",legend=c("Mortgage & Income logged","All Interactions","Interaction &
Log"),col=c("black","coral","green"),lty=1,lwd=1)
abline(a=0, b= 1)
```

+ ***Poly models ***
```{r, echo=FALSE}

plot(roc_PolyIncome2, col = "green", xlim = c(0, 0.3), ylim = c(0.7, 1.0))
plot(roc_PolyIncome3, col = "blue", add = TRUE, xlim = c(0, 0.3), ylim = c(0.7, 1.0))
plot(roc_CCAvg2, col = "blueviolet", add = TRUE, xlim = c(0, 0.3), ylim = c(0.7, 1.0))
plot(roc_CCAvg3, col = "aquamarine", add = TRUE, xlim = c(0, 0.3), ylim = c(0.7, 1.0))
plot(roc_PolyBoth2, col = "chartreuse", add = TRUE, xlim = c(0, 0.3), ylim = c(0.7, 1.0))
plot(roc_PolyBoth3, col = "coral", add = TRUE, xlim = c(0, 0.3), ylim = c(0.7, 1.0))
legend("bottomright", legend = c( "polyincome2", "polyincome3", "polyCCAVg2",
"polyCCAvg3", "polyboth2", "polyboth3"),
       col = c( "green", "blue", "blueviolet", "aquamarine", "chartreuse", "coral"),
       lty=1, lwd=1)
abline(a=0, b= 1)
```

These criterion of all models

```{r, echo=FALSE}
print("First basic models")
all_results
print("Interaction, Logginb and LDA/QDA results")
all_resultsX
print("LDA/QDA results")
all_results2
print("Polynomials resuls")
all_results3

```

References

[1] Kaggle (2021). *Bank Personal Loan Modelling (Supervised Learning)*. Data retrieved July 18, 2022, from https://www.kaggle.com/code/somnathpathak/bank-personal-loan-modelling-supervised-learning/data

[2] Forbes. *Personal Loan Originations To Surpass Pre-Pandemic Levels In 2022*. Diphtheria: Data retrieved Jan 31, 2022 from https://www.forbes.com/advisor/personal-loans/personal-loan-originations-projections/

[3] Consumer Credit - G.19. *More education is what makes people live longer, not more money.* Data retrieved July 8, 2022, from https://www.federalreserve.gov/releases/g19/current/