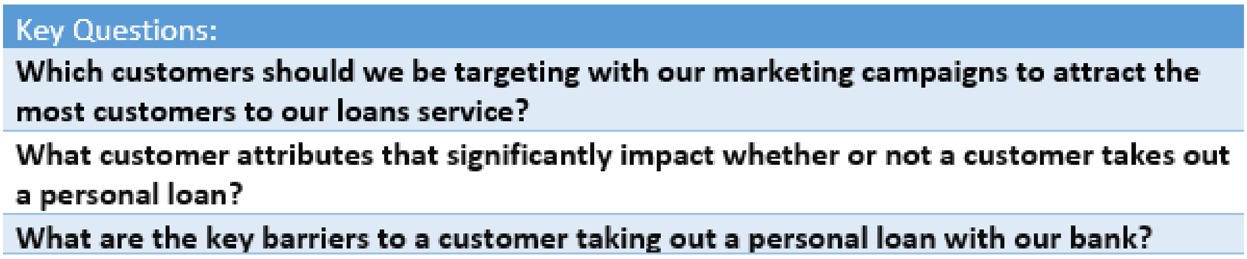
Data Analytics Project

Section 1 - The Problem

I have chosen to address a real-world problem currently impacting both large and small banks across the world. This report focuses on a new bank based in California who has started to amass customers which is, of course, a good thing, however, most of the customers it has attracted have been liabilities, customers. This means that it is attracting people to its general services, e.g. Credit Cards, etc, however, the bank wants to attract more people to its personal loans service as this is where it would produce greater revenue. Recently, the bank had a marketing campaign which attracted more liability customers to their personal loans. The bank now wants to know who the people are they should be targeting with their marketing campaigns to significantly increase the number of people using their loans service.

The decision maker for this problem will likely be the manager/owner of this private bank. It is likely that they do not understand complex data analytics/statistics thus information provided will need to be clear, easy-to-interpret and actionable. Below are the key questions that this manager would need to know from this analysis.



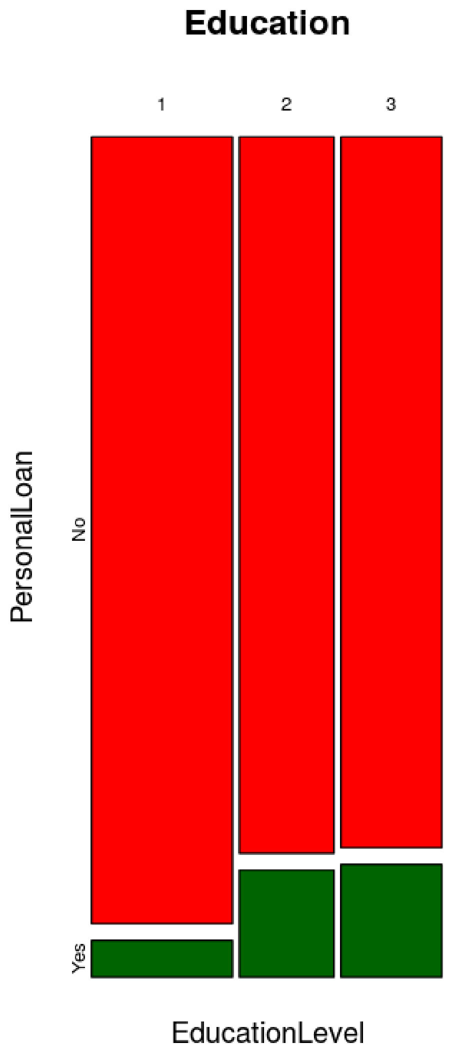
Key Questions:
Which customers should we be targeting with our marketing campaigns to attract the most customers to our loans service?
What customer attributes that significantly impact whether or not a customer takes out a personal loan?
What are the key barriers to a customer taking out a personal loan with our bank?

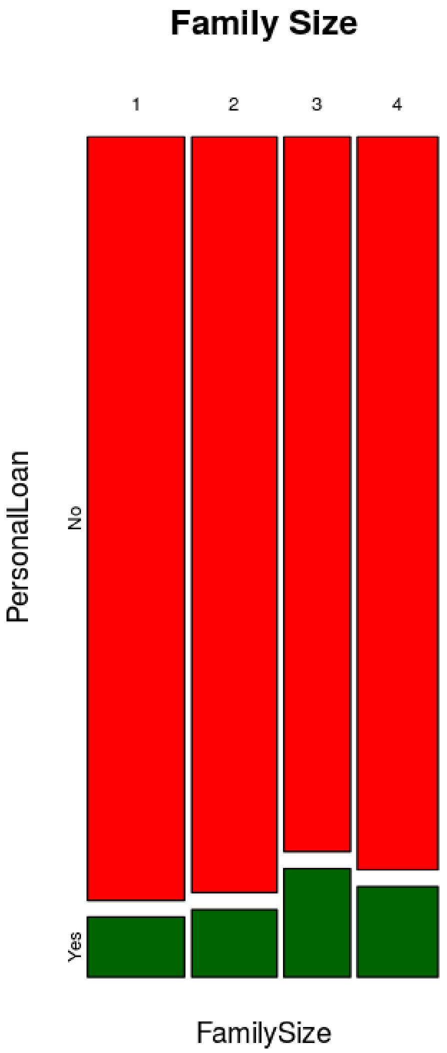
The Data

The data was compiled and collected by the bank itself. It contains information on current customers at this bank i.e. age, income etc. It also states whether or not this customer has a personal loan from this the bank. The data is highly reliable and relevant for the problem since it is directly sourced for the bank itself and the data is actually looking at the bank's real clients. This data dates 7 months back and thus is relatively new and is certainly still quite relevant for the bank. The data set already contains a variable "Personal Loans" which is a perfect fit for our target attribute. This is a binary variable where 1 means that the customer has taken out a personal loan with this bank. The ultimate goal of this analysis will be to create a model that can accurately relate the target attribute to some of the key explanatory attributes.

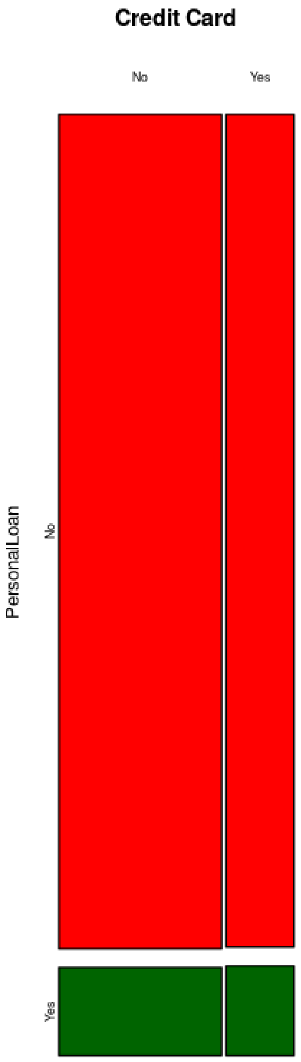
Section 2: Data Exploration

The dataset contains 11 variables and 5000 rows where each row represents a different customer at the at this bank. Initially, I began by exploring the overall data looking at the summary and highlighting any potential insights. I first began looked at how some of the categorical variables were distributed. Using proportion tables I explored the relationship between these variables and the target attribute.

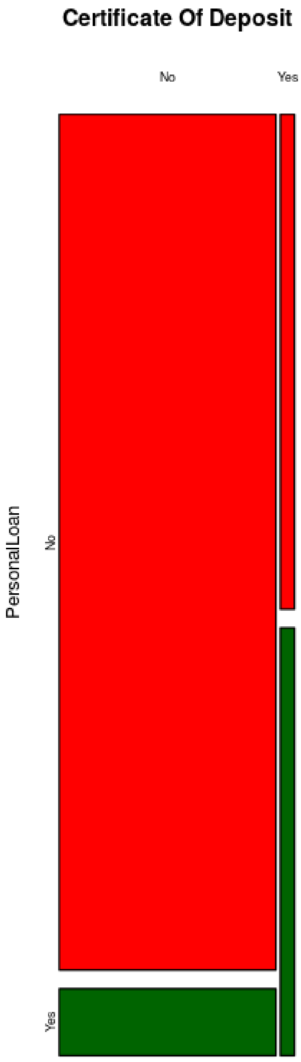




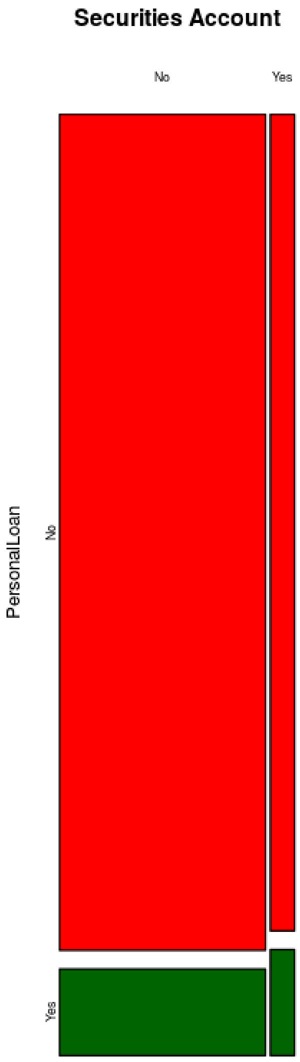
As we see above people with a higher education level take out a higher proportion of personal loans. The increase is more significant for from undergrad (1) to (2) grad. This makes sense as the many undergrad students already have lots of debt making it less likely that they would take out more loans. We also see that there are significant changes in the distribution of loans as family size fluctuates. As family size increase people take out more personal loans, up till size 3 at which point there is a decrease.



Credit Card

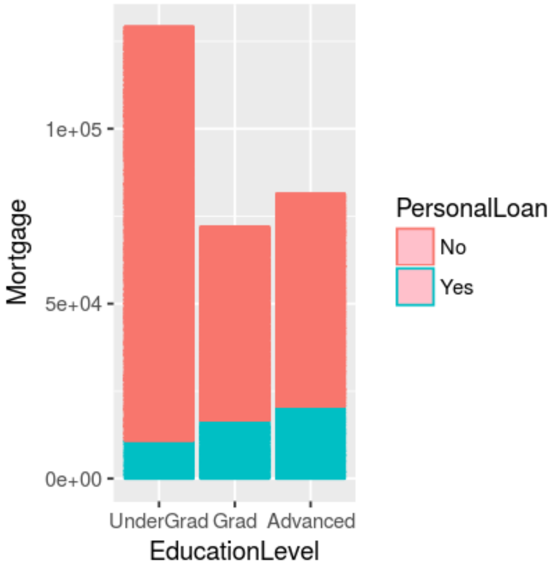


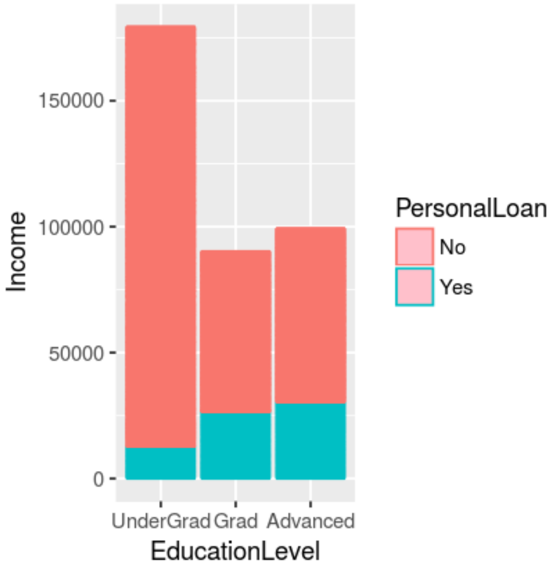
Certificate Of Deposit



Securities Account

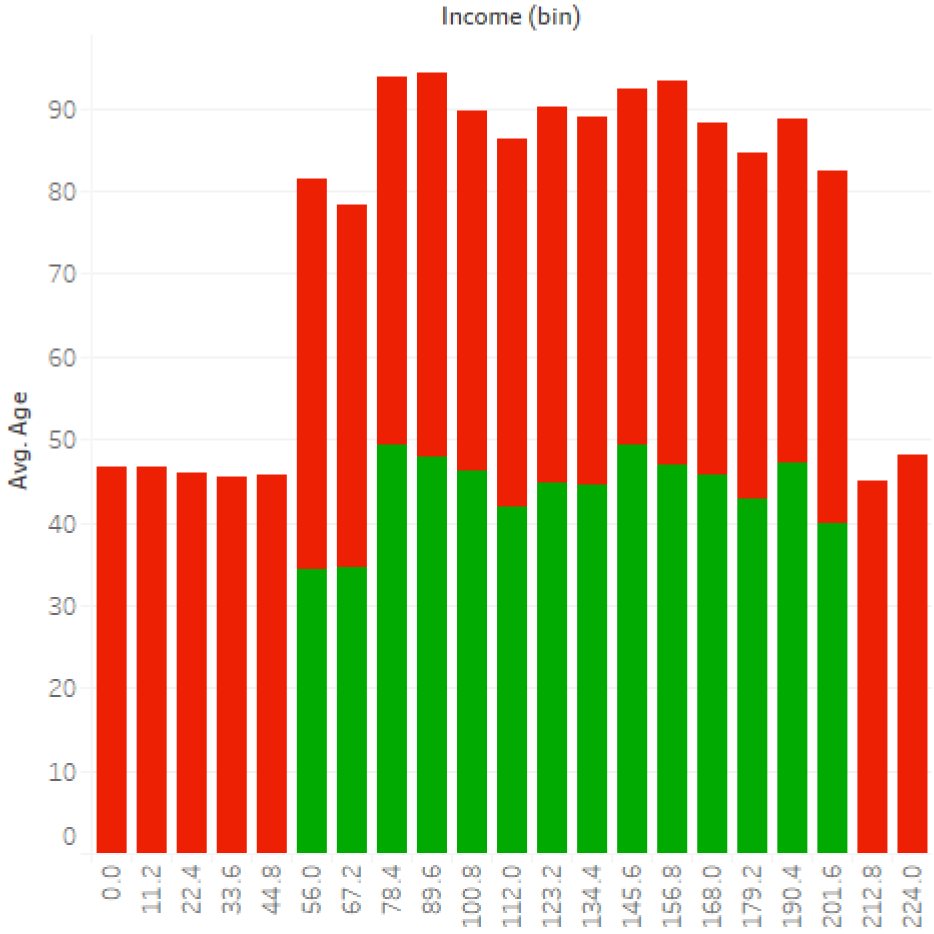
I also used these tables to explore the relation between the being a customer of a different service at this bank, e.g. credit cards, and how this impacts the likely hood of also being a personal loans customer. We see that for “credit card” and “securities account” the difference is minute, however, for we that people who hold a certificate of deposit are significantly more likely to take out a loan.





I then explored the relationship between income variables and education. We see that both “Mortgage” and “Income” have an almost identical relationship with education level and personal loans. We in fact see that this that customers who are undergraduates have the most income. Another interesting insight that may be drawn is the fact that people with higher incomes tend to take out less loans, this is strange as you would expect that higher income individuals would be more viable for paying back loans. To explore this further I looked at how personal loans changed with bother the age and income of a customer.

How Income and Age influence personal Loans



Average of Age for each Income (bin). Color shows details about Personal Loans.

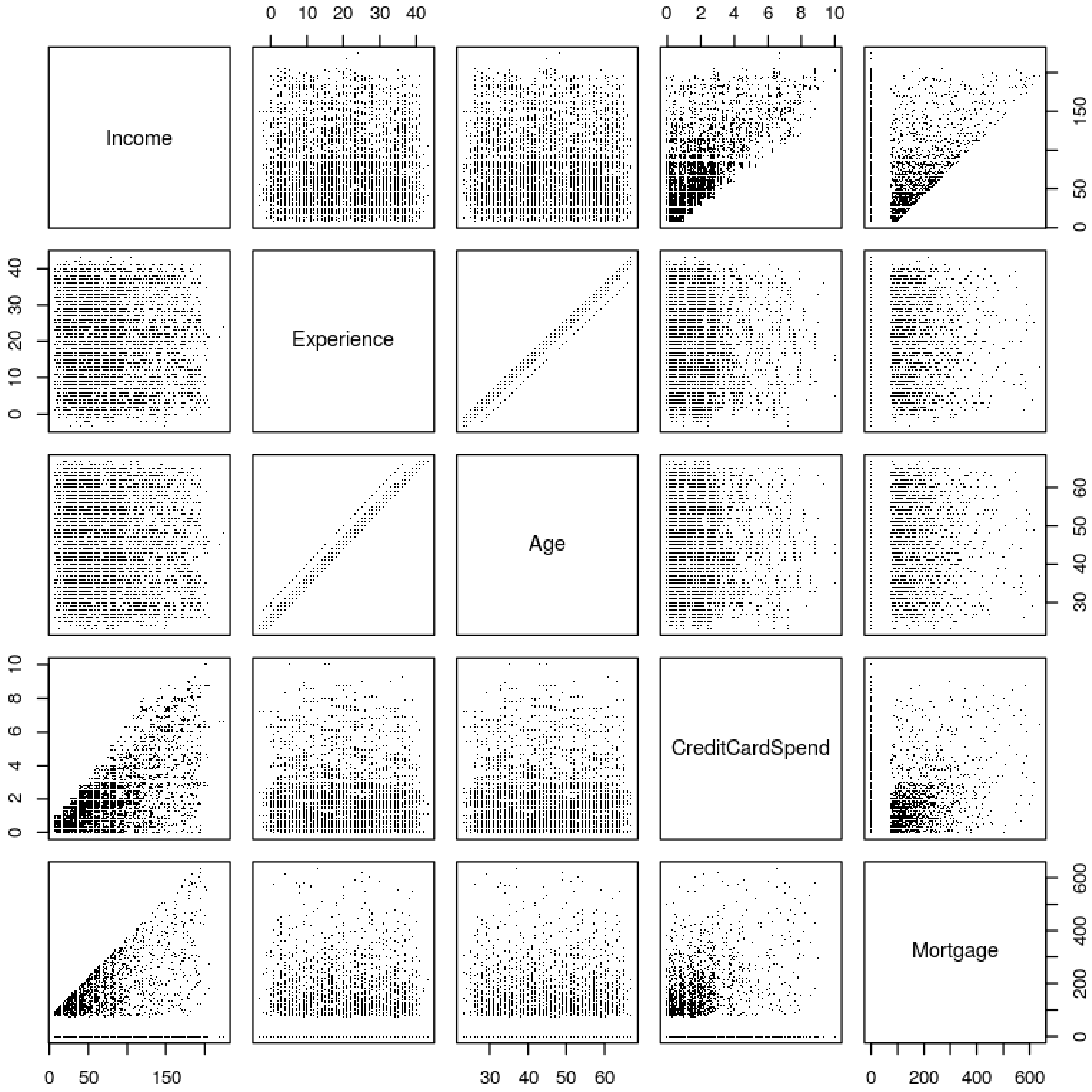
Personal Loans

🟥 No

🟩 Yes

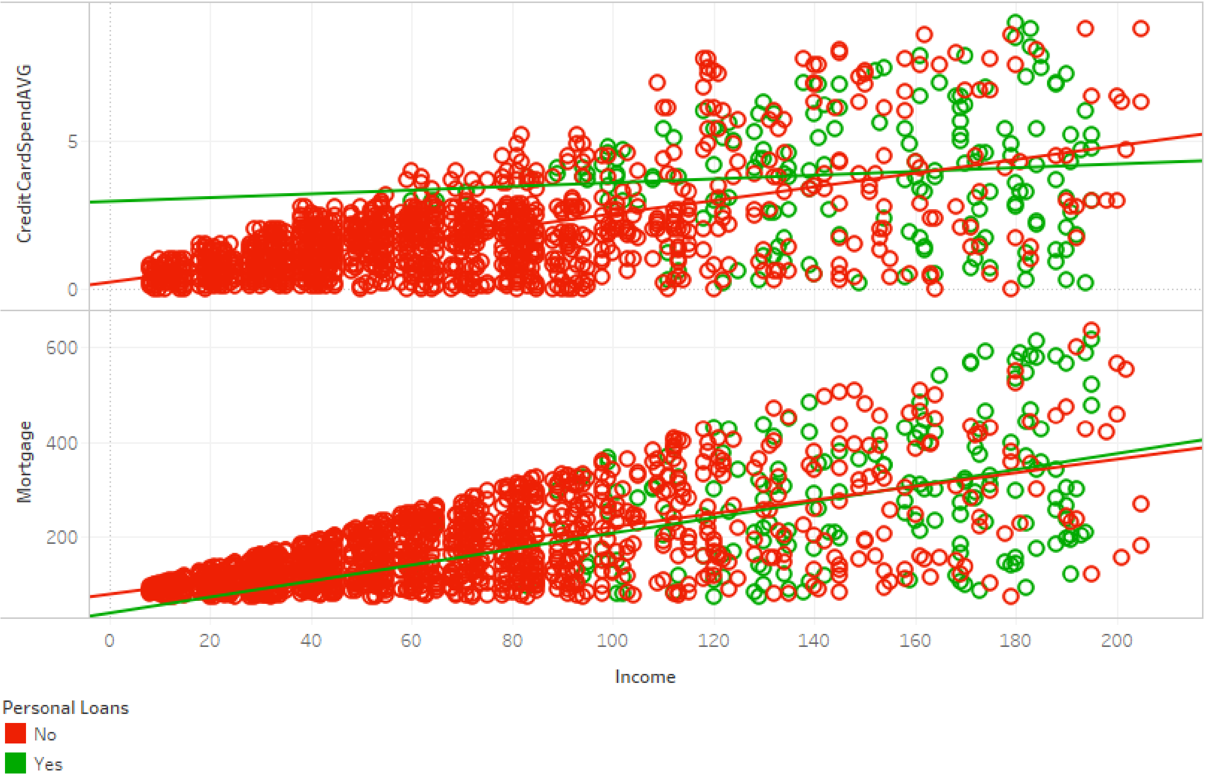
We see that both very low and very high-income individuals tend not to take out loans at this bank. Furthermore, we observe that whilst it is mostly mid-income people who are taking out loans these are usually younger people (sub 50) as older individuals are not taking out personal loans. This could be an interesting group that the bank may wish to target.

Correlation Analysis

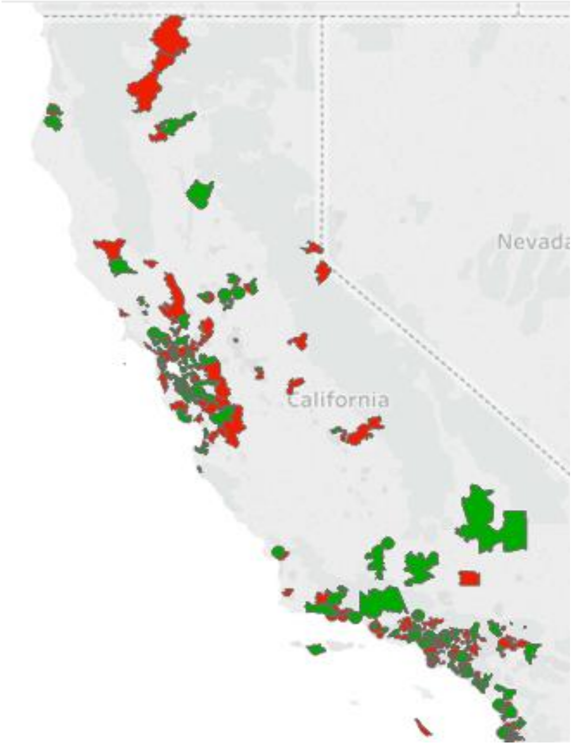


I first began by plotting some key variables together to see how different variables are correlated. We see some expected things, such as the fact that “Age” and “Experience” have very strong correlations, and income variables such as mortgage and credit card spend are strongly correlated. However, we make some interesting observations, for one, experience is not correlated within income. After this I performed correlation tests to see which of these variables had a statistically significant relation. I then plotted the significant variables in more clearly, as we can see the Income is strongly positively correlated with both credit card spending and mortgages. This will be important in the modelling phase as highly correlated variable impact regression model output.

Credit Card Spending, Mortgage and Income



Where are our customers?



I then used tableau to identify what key areas in California that this bank should target to attract the most customers. As we can see in the south, we have more people have personal loans. Whereas in central and north California are places where less customers take out loans with this bank. Thus, this could be an area of interest for the banks marketing campaign.

Section 3: Data Preparation

To prepare the data set I first started by renaming the columns as many of the columns had ambiguous unclear friends. This made the outputs later on more user friendly. I then used the “factor” function in R to rename the values of the categorical variables. Initially these were binary (0,1) however, changing these to yes/no made the data clearer and easier to interpret and later model. The same was done using a calculated field in Tableau to ensure that the plots were user friendly.

I also used the “str” function in R which told me how R was interpreting my data. Some of the variables were being misinterpreted by R. For example, some factor variables were being assumed to be strings. Using the different functions in R (e.g. as. factor, as.numeric) I reformatted the variables to ensure that they were all correctly understood by R. This was important for the modelling phase as mistreating the columns could lead to misleading or incorrect models being created. With regard to the target attribute no calculation was required as it was already available in the original data set. Also, did not contain main empty values, thus I chose to omit any NA values as there were so few thus this wouldn’t impact the data.

For forming the train, test and validation data sets these were broken down into 3 equal approximately equal size data sets. This allowed me to use the training data for the original models and then the validation only for the ensemble model to test how well it performs on new data it has not seen before.

Section 4: Prediction Models

The Models

My chosen models are classification tree, logistic regression and support vector machines. These are all supervised learning methods which fit the problem at hand. Starting with classification trees, this is a machine learning algorithm that partitions the data based on rules that are derived from the data to find the best partitions for the data. This is a great model for the problem as it will provide an easy to interpret tree that will classify our data. Logistic regression is a program that will find a relationship between the explanatory variable and the target, by providing an odds ratio. This tells how the likelihood of someone getting a personal loan change as you change the explanatory variable. Finally, support vector machines, analyses classification, and regression data to find the non-linear ways of splitting and grouping data using a hyperplane. This is a good fit for the problem as here we are dealing with multiple variables.

Variable Selection

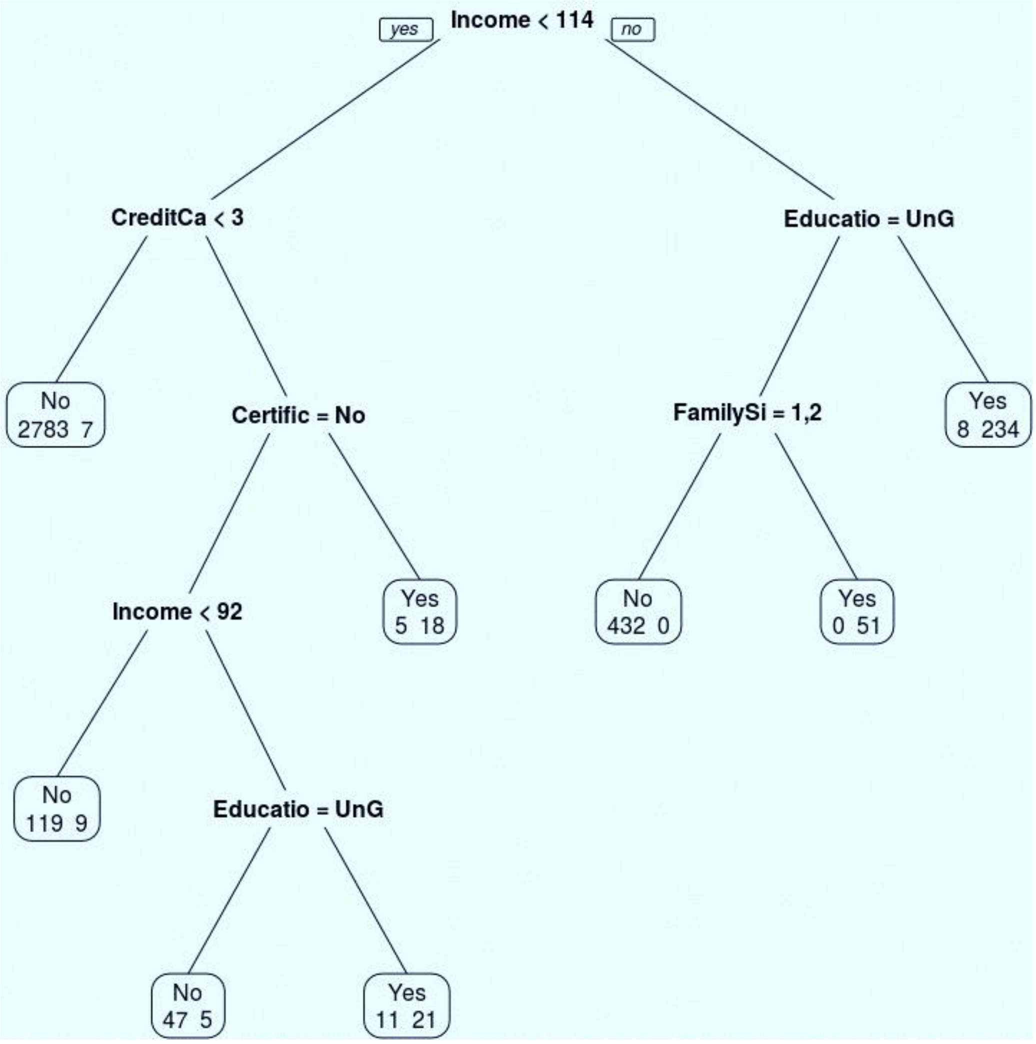
1. First, I created and pruned a tree with all the variables to identify which variables were important for creating the tree.
2. This identified some key variables such as income (the root) node and education as being important predictions.
3. I then used the correlation analysis (Section 2) to remove variables that were highly correlated as this would create inconsistencies in the model outputs.
4. After performing a logistic regression with all the variables, I identified the ones that were statistically significant.

For each model, I initially trained the model with all the variables and then began to adjust variables based on these factors to improve the accuracy and model outputs. For example, for logistic regression I used the deviance for each variable to the Null which, the higher the deviance the more important the predictor.

Chosen Variables:
Income, Family Size, Credit Card Spend, Education Level, Certificate of Deposit, Online

Model Insights

Classification Tree



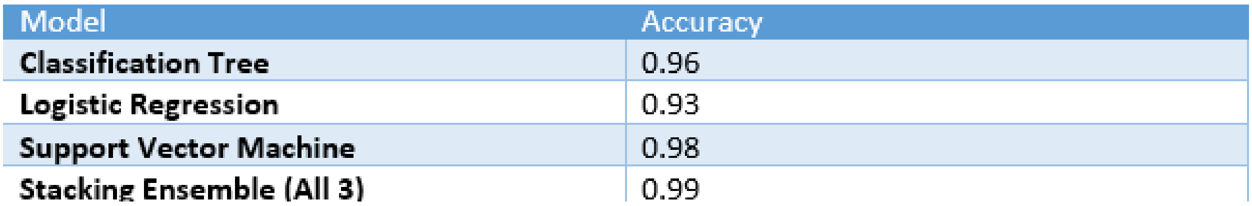
This model tells us a few things first that there is a large portion of people that are not taking out personal loans because of income. We see that people with <114k income and who spend less than 3k on their credit card are unlikely to take out a loan. This is a potential group that the bank may want to target to attract and convert more customers. We also see that a significant portion of people who take out loans are no undergraduates. This means that undergraduates could be another key area that the bank may wish to target.

Logistic Regression Model

From this model we see some evidence to support the classification tree. The odds ratio of the education level is 4.3 and is statistically significant to 1%. This means that if someone has an education level of graduate or advanced, they are 73 times more likely to have an account with this bank than someone who is an undergraduate. We also see another significant predictor in certificate of deposit account. Having a CD account meant you were 20 times more likely to be a personal loans customer at this bank.

Ensemble Model

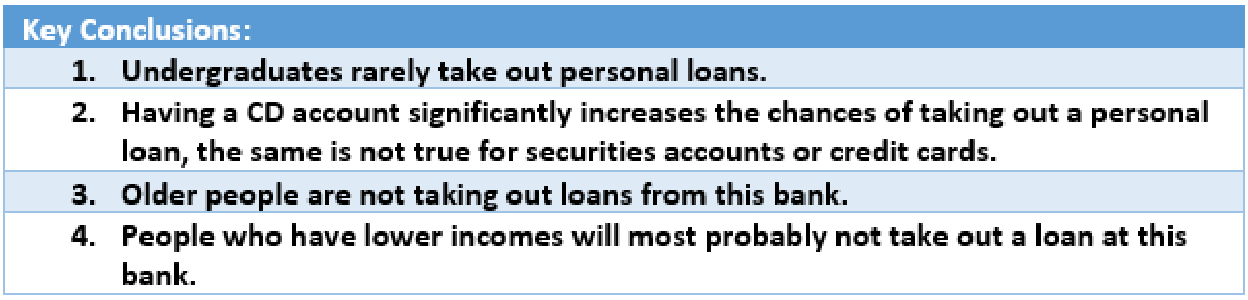
This model takes the three models previously mentioned and it trains the data set using each model. After these it takes these intermediate predictions and uses this in a new model (in this case a random forest was used) to create a new model which as learned from all the previous model. This model is then used to create new predictions.



Model	Accuracy
Classification Tree	0.96
Logistic Regression	0.93
Support Vector Machine	0.98
Stacking Ensemble (All 3)	0.99

Above we see the accuracy of each model. The best model was the stacking ensemble model, this, in theory, is to be expected as stacking different models will generally improve the accuracy of predictions. Also, of the three models support vector machines performed best, this could be for many reasons, one is that support vector machines are less sensitive to outliers thus extreme values would impact other models more strongly (Drakos, 2019).

Section 5: Problem Conclusions and Recommendations



Key Conclusions:
1. Undergraduates rarely take out personal loans.
2. Having a CD account significantly increases the chances of taking out a personal loan, the same is not true for securities accounts or credit cards.
3. Older people are not taking out loans from this bank.
4. People who have lower incomes will most probably not take out a loan at this bank.

The key variables look at are income variables, education level and certificate of deposit. From the conclusions that have been drawn from this analysis, I recommend that this bank focus on the groups of people who have not been taking out loans. For example, older people tend to have disposable income thus they would be able to repay loans, therefore, it may be the case that the marketing campaigns are not being catered to older generations. Also, people with lower incomes (<114k) tend to avoid taking out loans, however, many of them have enough money to afford these loans, thus a marketing campaign targeted at them may be more fruitful. I believe focusing on groups have thus far not engaged with personal loans has the most potential in increasing the proportion of people taking out loans at this bank.

Words: 1997

References

Drakos, G. (2019). *Support Vector Machine vs Logistic Regression*. [online] Towards Data Science. Available at: https://towardsdatascience.com/support-vector-machine-vs-logistic-regression-94cc2975433f [Accessed 23 Mar. 2019].

Mathew, A. (2019). *Evaluating Logistic Regression Models*. [online] R-bloggers. Available at: https://www.r-bloggers.com/evaluating-logistic-regression-models/ [Accessed 23 Mar. 2019].

Data Source:

Kaggle.com. (2019). *Bank_Loan_modelling*. [online] Available at: https://www.kaggle.com/itsmesunil/bank-loan-modelling#Bank_Personal_Loan_Modelling.xlsx [Accessed 23 Mar. 2019].

Supplementary Code Appendix

Data Description Code

```
## Explore the correlations than do some plots do explore highly correlated variables.

pairs(~ Income + Experience + Age + CreditCardSpend + Mortgage, data = bank.df,

   pch=".")

library(datasets)

library(ggplot2)

Plot1 <- ggplot(bank.df[which(bank.df$Mortgage>0),], aes(x = Income , y = Mortgage, colour = PersonalLoan)) +

    geom_point() + geom_smooth(method=lm)

Plot1 <- Plot1 + ggtitle("Relation between Income and Mortages")

Plot1

### p-value < 2.2e-16 very significant correlation, rno = 0.65, ryes = 0.74,


Plot3 <- ggplot(bank.df[which(bank.df$Mortgage>0),], aes(x = Mortgage , y =CreditCardSpend , colour = PersonalLoan)) +

    geom_point() + geom_smooth(method=lm)

Plot3 <- Plot3 + ggtitle("Relation between Income and Mortages")

Plot3

### p-value < 2.2e-16

install.packages("ggpubr")

library("ggpubr")

## Excluding mortgage values of 0 as these are simply people that did not take out a mortgage.
## Correlation test for Income variable

cor1 <- cor.test(bank.df[which(bank.df$Mortgage>0),]$Mortgage[bank.df$PersonalLoan =="No"],

          bank.df[which(bank.df$Mortgage>0),]$Income[bank.df$PersonalLoan =="No"], method = "pearson")

cor1

cor3 <-cor.test(bank.df[which(bank.df$Mortgage>0),]$Mortgage[bank.df$PersonalLoan =="Yes"],

          bank.df[which(bank.df$Mortgage>0),]$Income[bank.df$PersonalLoan =="Yes"], method = "pearson")

cor3

attach(bank.df)

t.test(Income[PersonalLoan == "No"],Income[PersonalLoan == "Yes"])

p<-ggplot(bank.df, aes(x=FamilySize, y=Mortgage)) +

  geom_bar(stat="identity", fill="blue")

p

p4 < - ggplot(bank.df, aes(x=FamilySize, y=Mortgage)) +

  geom_bar(stat="identity", fill="blue")

### People who have a family size of have the lowest mortages taken out,


p1<-ggplot(bank.df, aes(x=EducationLevel, y=Mortgage, color = PersonalLoan)) +

  geom_bar(stat="identity", fill = "Pink")

bank.df$EducationLevel <- factor(bank.df$EducationLevel,

                labels = c("UnderGrad", "Grad", "Advanced"))

p2<-ggplot(bank.df, aes(x=EducationLevel, y=Income, color = PersonalLoan)) +

  geom_bar(stat="identity", fill = "Pink")


library("gridExtra")

grid.arrange(p1,p2, ncol = 2, nrow = 1)


par(mfrow=c(1,2))

plot(table(EducationLevel,PersonalLoan), color = c("Red", "darkgreen"), main = "Education")

plot(table(FamilySize,PersonalLoan), color = c("Red", "darkgreen") , main = "Family Size")
```

```
#### We see that as education level increases the proportion of people taking out personal loans also increases.

par(mfrow=c(1,3))

plot(table(bank.df$CreditCard,PersonalLoan), color = c("Red", "DarkGreen"), main = "Credit Card")

plot(table(bank.df$CertificateOfDeposit,PersonalLoan), color = c("Red", "darkgreen"), main= "Certificate Of Deposit")

plot(table(bank.df$SecuritiesAccount,PersonalLoan), color = c("Red", "darkgreen"), main = "Securities Account")
```

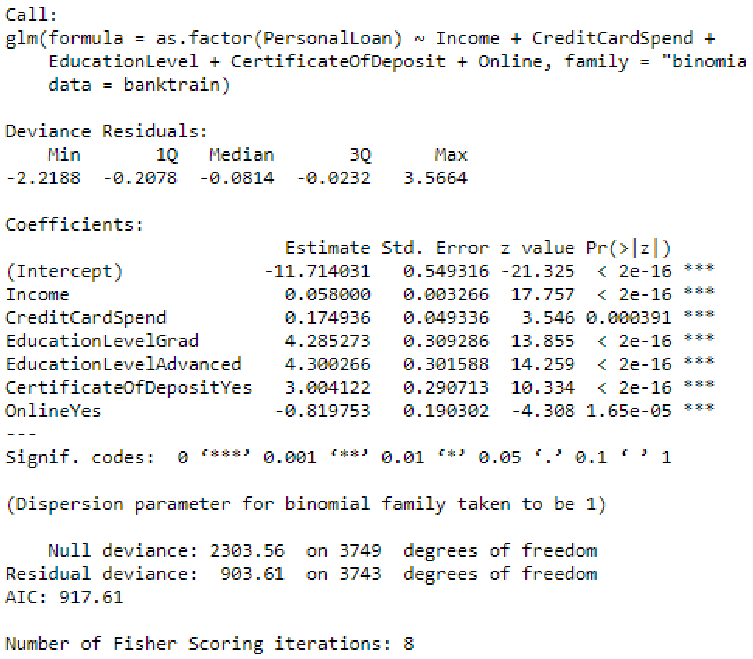
Data Organisation and Cleaning:

```
#### making training, validation and test data

set.seed(123)

bank.df1 <- bank.df1[sample(nrow(bank.df)),]

split <- floor(nrow(bank.df)/3)

trainingData <- bank.df1[0:split,]

validationData <- bank.df1[(split+1):(split*2),]

testingData <- bank.df[(split*2+1):nrow(bank.df),]


## Making Binary variables YES/NO to make it clear in plots and models.

bank.df$CreditCard <- factor(bank.df$CreditCard,

                    labels = c("No","Yes"))

bank.df$CertificateOfDeposit <- factor(bank.df$CertificateOfDeposit,

                    labels = c("No","Yes"))

bank.df$SecuritiesAccount <- factor(bank.df$SecuritiesAccount,

                    labels = c("No","Yes"))

bank.df$Online <- factor(bank.df$Online,

                    labels = c("No","Yes"))

bank.df$EducationLevel <- factor(bank.df$EducationLevel,

                    labels = c("UnderGrad", "Grad", "Advanced"))

### Making column names clear to make model outputs and data description easiliy interpretable.
```

```
colnames(bank.df) <- c("ID","Age", "Experience", "Income", "ZIPCode", "FamilySize", "CreditCardSpend","EducationLevel", "Mortgage",
                   "PersonalLoan", "SecuritiesAccount", "CertificateOfDeposit", "Online", "CreditCard")


#since there are only a small number of missing values we omit the rows with NA values
x <- na.omit(bank.df)
#Also mortage has alot of skewed zero values this is because people without a martgage would receivie a default of 0, making it
# less accurate.
# Ensuring R understand the different data types using as.factor fucntion in r
as.factor(bank.df$PersonalLoans)



##### Making sure R understands what each varible represents i.e. strings, factors, numbers etc
bank.df$FamilySize <- as.factor(bank.df$FamilySize)
bank.df$EducationLevel <- as.factor(bank.df$EducationLevel)
bank.df$PersonalLoan <- as.factor(bank.df$PersonalLoan)
bank.df$SecuritiesAccount <- as.factor(bank.df$SecuritiesAccount)
bank.df$CertificateOfDeposit <- as.factor(bank.df$CertificateOfDeposit)
bank.df$Online <- as.factor(bank.df$Online)
bank.df$CreditCard <- as.factor(bank.df$CreditCard)
bank.df$Experience <- as.numeric(bank.df$Experience)
bank.df$Income <- as.numeric(bank.df$Income)
bank.df$Age <- as.numeric(bank.df$Age)
bank.df$Mortgage <- as.numeric(bank.df$Mortgage)
str(bank.df)
```

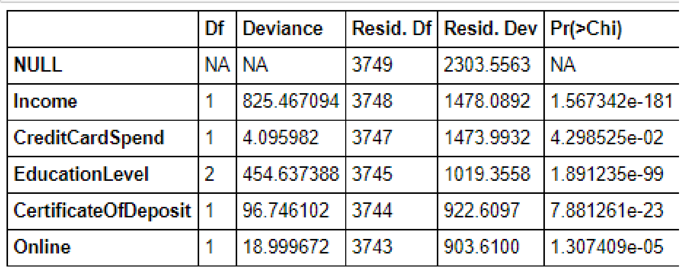
Data Modelling:

```
library(rpart)
library(rpart.plot)
treeAll = rpart(as.factor(PersonalLoan) ~ ID+Age+ Experience + Income + ZIPCode + FamilySize + CreditCardSpend+ EducationLevel+  Mortgage+
            SecuritiesAccount+ CertificateOfDeposit+  Online+ CreditCard, data = banktrain, method =
"class")
prp(treeAll, cex = 0.8 , extra = 1)
   text(treeAll,cex=.5)
predict = predict(treeAll,newdata=banktest,type="class")
accuracy <- table(predict, banktest[,10])
print(accuracy)
sum(diag(accuracy))/sum(accuracy)
Tree1 <- rpart(as.factor(PersonalLoan) ~Income + FamilySize +  CreditCardSpend+ EducationLevel+
           CertificateOfDeposit+  Online, data = banktrain, method = "class")
prp(Tree1, cex = 0.8, extra = 1)
   text(Tree1)
predict = predict(Tree1,newdata=banktest,type="class")
accuracy <- table(predict, banktest[,10])
print(accuracy)
sum(diag(accuracy))/sum(accuracy)


Glm2 <- glm(as.factor(PersonalLoan) ~  Income +  CreditCardSpend+
EducationLevel+CertificateOfDeposit+
          Online, data = banktrain, family = "binomial")
summary(Glm2)
```



```
Call:
glm(formula = as.factor(PersonalLoan) ~ Income + CreditCardSpend +
    EducationLevel + CertificateOfDeposit + Online, family = "binomia
    data = banktrain)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2188  -0.2078  -0.0814  -0.0232   3.5664

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             -11.714031   0.549316 -21.325  < 2e-16 ***
Income                    0.058000   0.003266  17.757  < 2e-16 ***
CreditCardSpend           0.174936   0.049336   3.546 0.000391 ***
EducationLevelGrad        4.285273   0.309286  13.855  < 2e-16 ***
EducationLevelAdvanced    4.300266   0.301588  14.259  < 2e-16 ***
CertificateOfDepositYes   3.004122   0.290713  10.334  < 2e-16 ***
OnlineYes                -0.819753   0.190302  -4.308 1.65e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2303.56  on 3749  degrees of freedom
Residual deviance:  903.61  on 3743  degrees of freedom
AIC: 917.61

Number of Fisher Scoring iterations: 8
```

```
pred = predict(Glm2, newdata=banktest, type = "response")

pred = ifelse(pred>.5, 1, 0)

accuracy <- table(pred, banktest[,10])

sum(diag(accuracy))/sum(accuracy)

anova(Glm2, test="Chisq")
```



	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL	NA	NA	3749	2303.5563	NA
Income	1	825.467094	3748	1478.0892	1.567342e-181
CreditCardSpend	1	4.095982	3747	1473.9932	4.298525e-02
EducationLevel	2	454.637388	3745	1019.3558	1.891235e-99
CertificateOfDeposit	1	96.746102	3744	922.6097	7.881261e-23
Online	1	18.999672	3743	903.6100	1.307409e-05

```
install.packages("e1071")

library(e1071)


set.seed(123)

mod.svm <- train(as.factor(PersonalLoan) ~Income + FamilySize +  CreditCardSpend+ EducationLevel+

            CertificateOfDeposit+  Online, method = "svmRadial",data = banktrain)

pred.svm <- predict(mod.svm, banktest)

confusionMatrix(pred.svm, banktest$PersonalLoan)$overall[1]
```

```
###Ensemble learning method
bank.df1 <- select(bank.df,PersonalLoan, Income, FamilySize, CreditCardSpend, EducationLevel,
            CertificateOfDeposit,  Online )


dim(trainingData)
dim(validationData)
dim(testingData)


labelName <- 'PersonalLoan'


predictors <- names(trainingData)[names(trainingData) != labelName]


myControl <- trainControl(method='cv', number=3, returnResamp='none')



test_model1 <- train(validationData[,predictors], validationData[,labelName], method='glm',
trControl=myControl)
test_model2 <- train(validationData[,predictors], validationData[,labelName], method='svmRadial',
trControl=myControl)
test_model <- train(validationData[,predictors], validationData[,labelName], method='rpart',
trControl=myControl)


###Accuracy
preds <- predict(object=test_model, testingData[,predictors])
library(pROC)
auc <- roc(testingData[,labelName], preds)
print(auc$auc)
model_tree <- train(trainingData[,predictors], trainingData[,labelName], method='rpart',
trControl=myControl)


model_glm <- train(trainingData[,predictors], trainingData[,as.factor(labelName)], method='glm',
trControl=myControl)
```

```
model_svm <- train(trainingData[,predictors], trainingData[,labelName], method='svmRadial',
trControl=myControl)


validationData$tree_PROB <- predict(object=model_tree, validationData[,predictors])
validationData$glm_PROB <- predict(object=model_glm, validationData[,predictors])
validationData$svm_PROB <- predict(object=model_svm, validationData[,predictors])

testingData$tree_PROB <- predict(object=model_tree, testingData[,predictors])
testingData$glm_PROB <- predict(object=model_glm, testingData[,predictors])
testingData$svm_PROB <- predict(object=model_svm, testingData[,predictors])

predictors <- names(validationData)[names(validationData) != labelName]
final_blender_model <- train(validationData[,predictors], validationData[,labelName], method='rf',
trControl=myControl)

preds <- predict(object=final_blender_model, testingData[,predictors])
auc <- roc(testingData[,labelName], preds)
print(auc$auc)
```