**70-339 Final Project**

Applying Machine Learning to Predicting Loan Defaults

Saloni Dalal, Harshini Malli

**Introduction:**

In class we experienced making small linear regression models on Excel for determining whether an individual will default or pay back their loan. We wanted to expand on that and use Machine Learning to build a model with more variables to determine whether an individual will default on their loans. We wanted to apply our knowledge of training and testing our data and build some of the models illustrated in lecture as well as determining some of the key features in identifying whether an individual will default on their loan. These include logistic regression, support vector machine, decision trees, and neural networks. We also used cross validation in order to evaluate our machine learning models. We believed this project allowed us to apply all the concepts learned in class, as well as dive deeper by learning basic coding in R and evaluating the pros and cons of our model.

We hypothesized that the decision tree model would be the best fitting model for our data set. This is because this model can handle large datasets well. Furthermore, the data we used has very few continuous variables, so this would not pose any problems. Additionally, we hypothesized that the neural networks model would be the worst fitting model as it often overfits the data. Therefore, we thought that this model would be less accurate after performing cross validation.

**Data:**

We used a German dataset on the University of California at Irvine website. This dataset had 20 different attributes that were a mix of quantitative (Q) and qualitative (L):

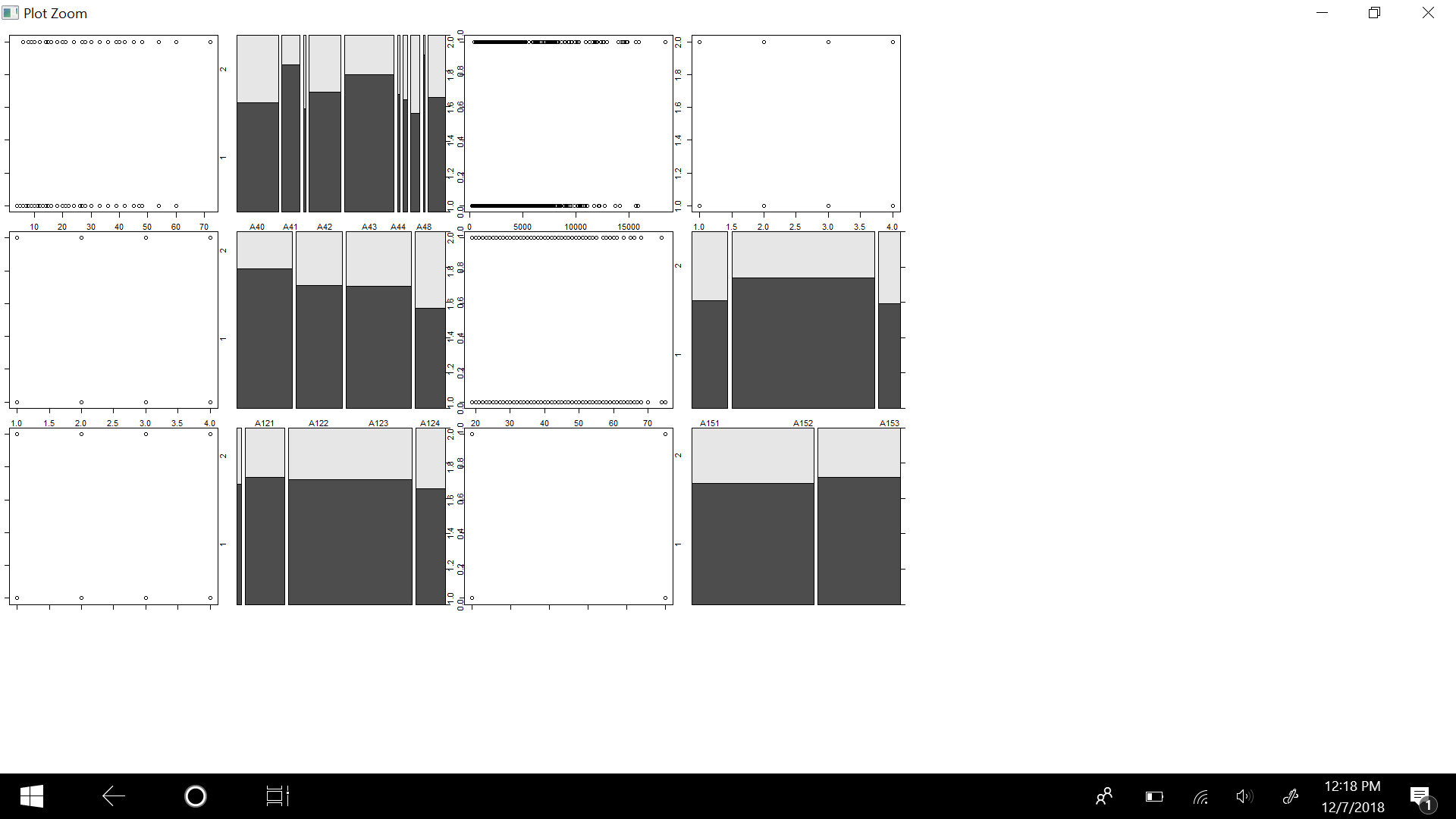
* Status of checking account (L)
* Duration in month (Q)
* Credit history (L)
* Purpose (L)
* Credit amount (Q)
* Savings account/bonds (L)
* Years at present employment (L)
* Installment rate in percentage of disposable income (Q)
* Personal status and sex (L)
* Other debtors/guarantors (L)
* Years at present residence (Q)
* Property (L)
* Age in years (Q)
* Other installment plans (L)
* Housing (L)
* Number of existing credits at bank (Q)
* Job type (L)
* Number of people being liable to provide maintenance for (Q)
* Telephone (L)
* Foreign worker (L)

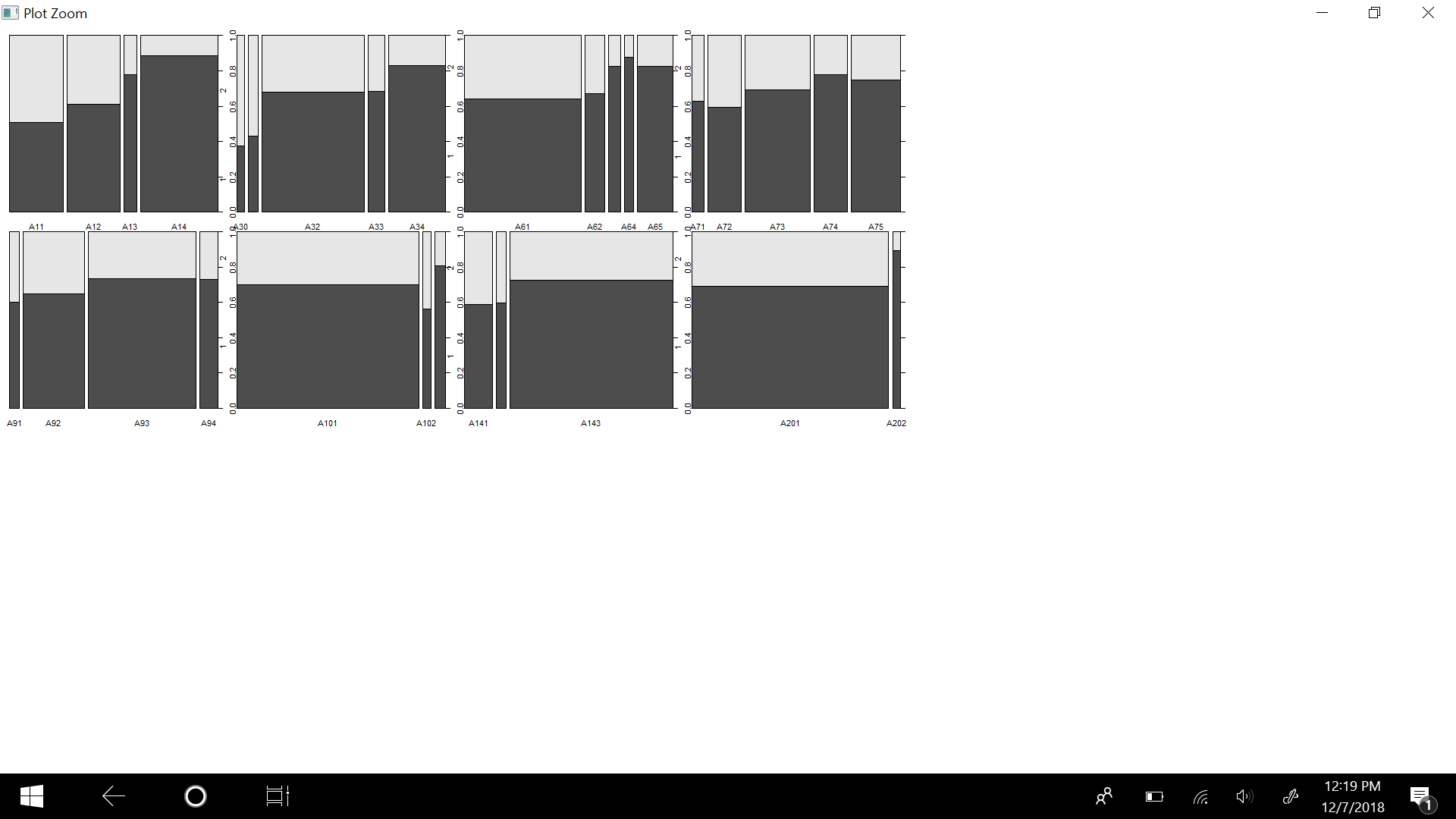
Each of the qualitative variables had different codes to represent those characteristics. These variables are traditionally taken into account in decisions of whether to extend credit to an individual. (Key is located in appendix)

**Code Explanation:**

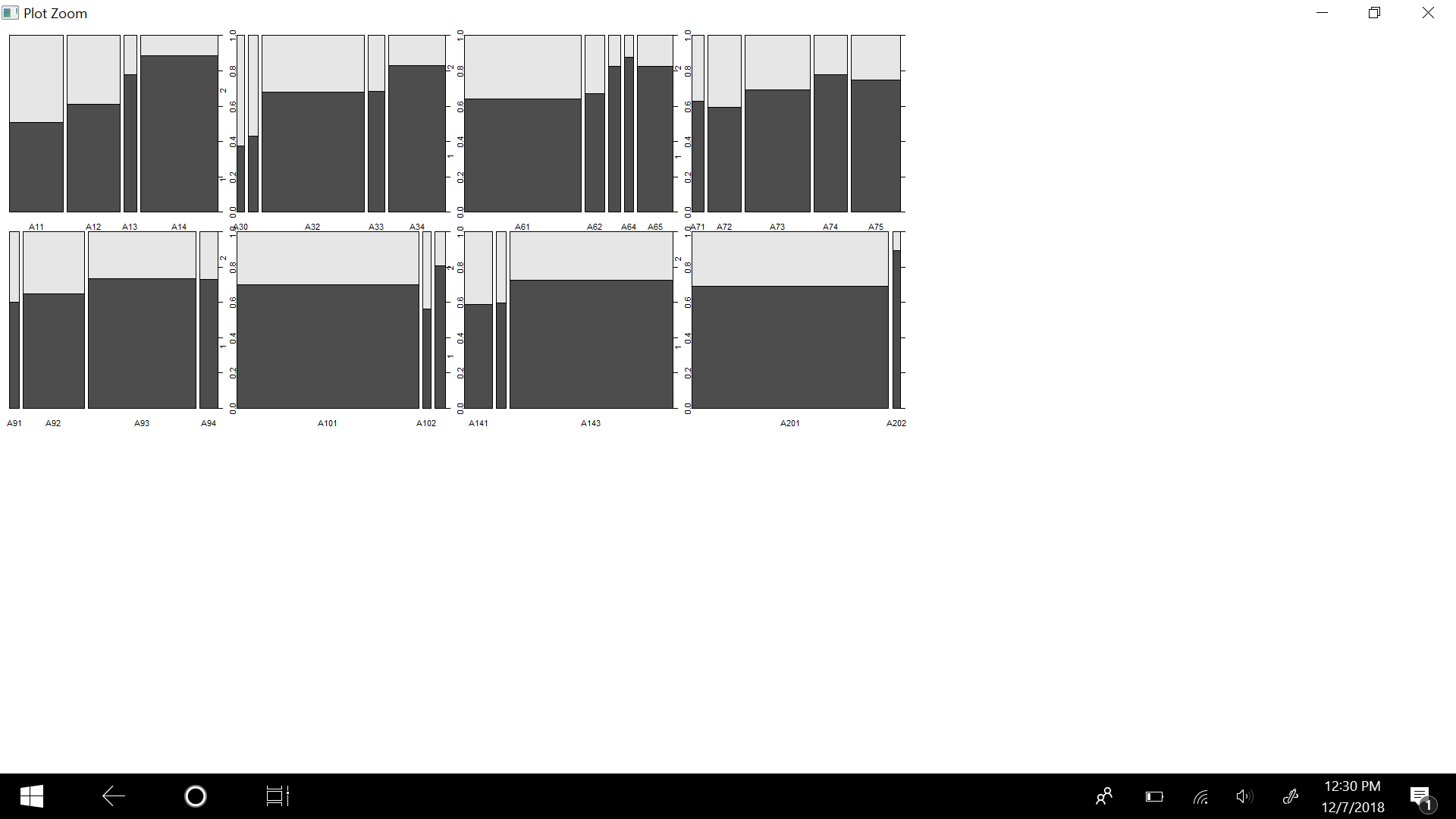
*Selecting the variables:*

As specified above there were 20 columns of data with a total of 48 variables. Most were categorical variables but 7 of them were continuous. In order to first decide which variables to use we used R to build a data frame of the variables and plot all of them against our default. You can see the plots below: those that were continuous showed scatter plots and those that were categorical showed bar graphs.

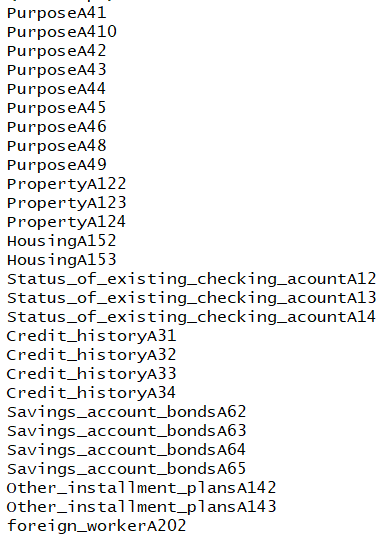
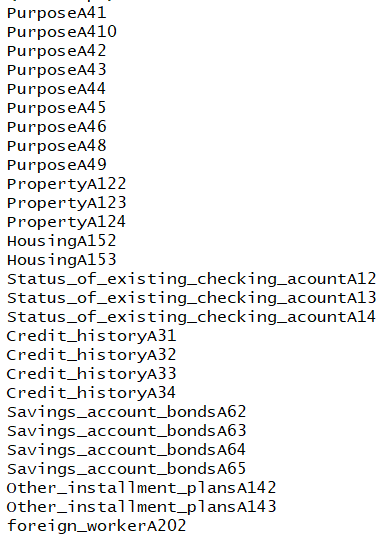




Looking at just the scatter plots we could determine that none of the continuous variables were a good fit. We can see this because at any point as the X values increase there is an equal distribution between whether the plots are categorized as 1 or 2 (pay back or default). We thus eliminated those variables from the data frame when creating our models. Looking at the categorical variables we were looking for variables that had an unequal distribution between each subcategory. We will describe our process by using one plot as an example:



Here is the plot comparing Status of Checking Accounts to Default value. A11 corresponds to a negative debit memo, A12 corresponds to a debit memo of between 0 and 200, A13 corresponds to a debit memo greater than 200, and A14 correlates to no checking account. This seemed like a significant variable because it had an inequal distribution. Those without checking accounts and a debit memo greater than 200 were much more likely to default than those with lower debit memos. Thus we decided to use this variable.



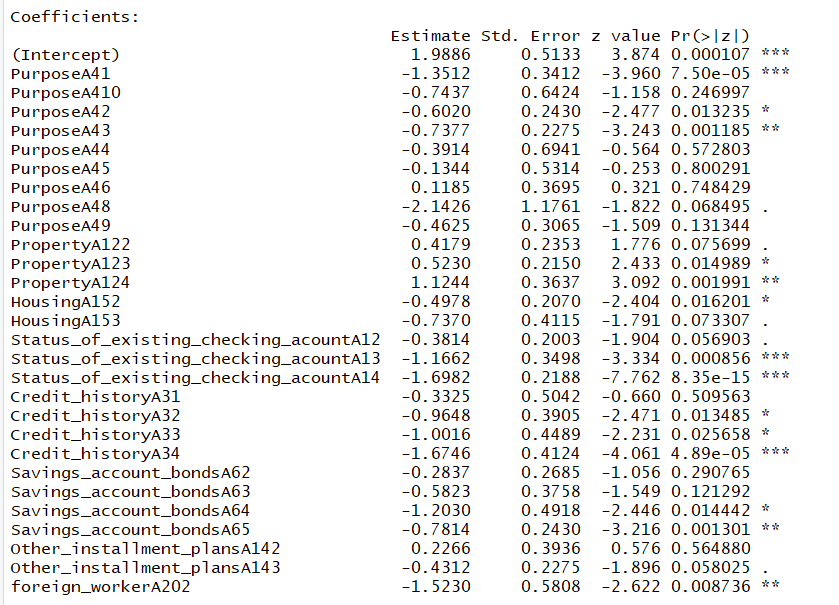
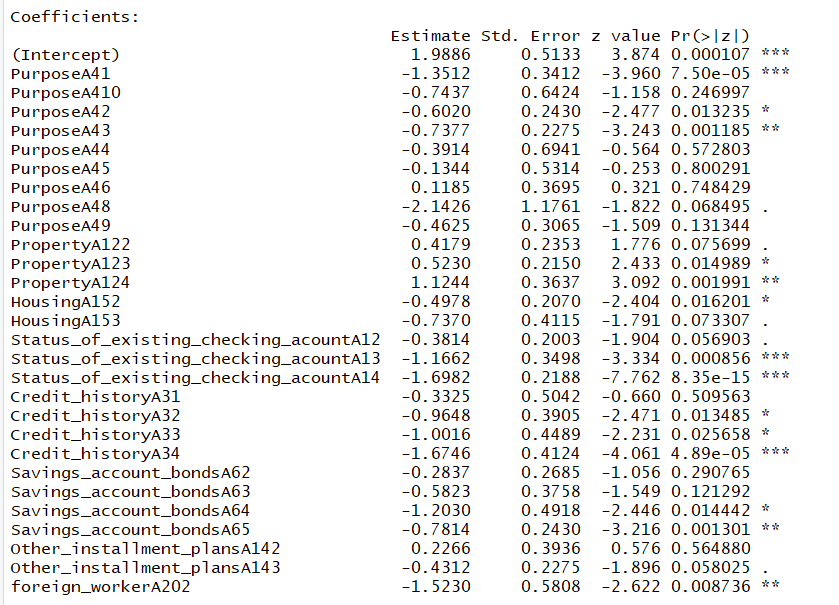
Using this process we decided to eliminate the following variables: personal status and sex (A91-A95), Job (A171-A174), Telephone (A191-A192), and other installment plans (A141-A143). We thus started with 20 categories and 48 total variables and now we’re at 9 categories and 30 total variables. This includes the following variables on the left. We then updated the data-frame to only include those variables for the rest of the models.

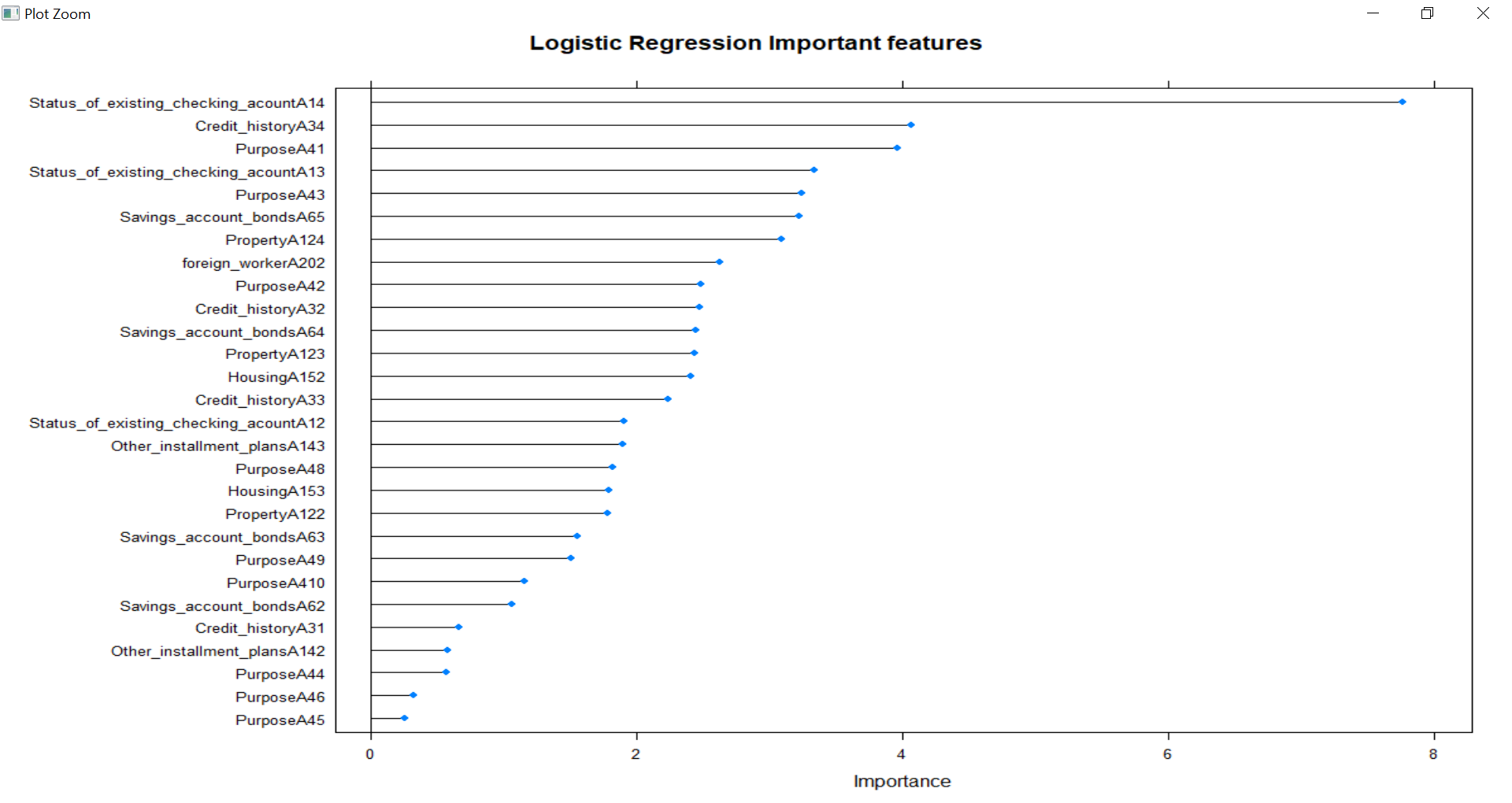
*Building the Models:*

We went through the pros and cons of various models and ultimately decided to use logistic regression, neural networks, support vector machine, and a decision tree model. We chose to do a logistic regression model because it is easy to interpret and explain and relatively fast. However, we also noted that logistic regression cannot handle too many predictor variables. Neural networks are also fast and flexible in terms of handling complex situations, but are rather prone to overfitting. SVMs are good at performing non-linear classifications by using the kernel trick. On the other hand they are very sensitive to noise and are computationally expensive. Decision trees are great for large data sets but become more complicated to understand on larger trees. Additionally, all continuous variables need to be binned into categorical variables. Lastly it easy to under and over fit. Since we removed all the continuous variables using decision trees was more simple.

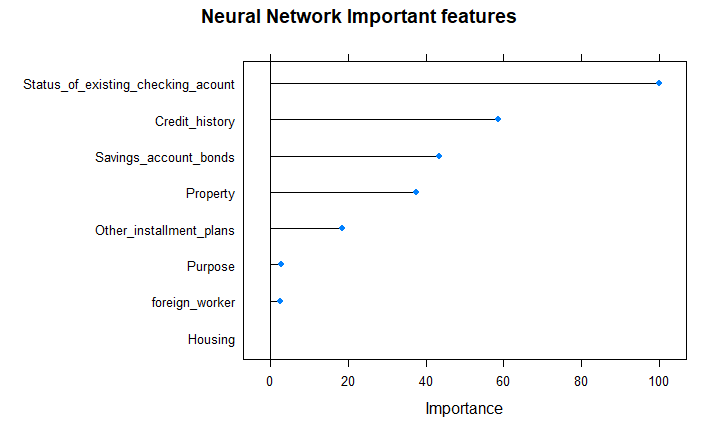
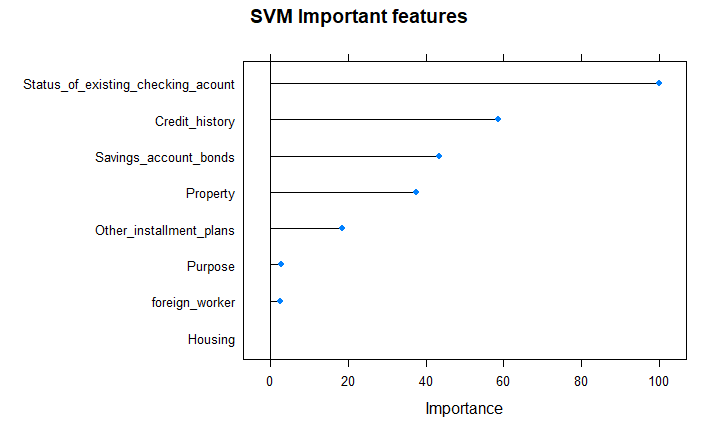
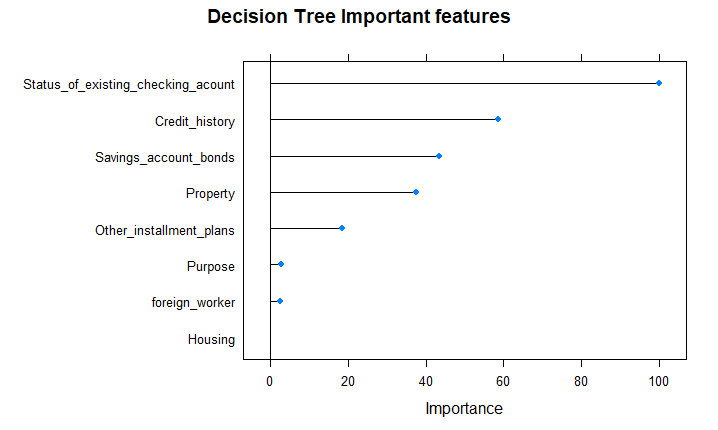
We decided against using a KNN model because we realized the model is very dependent on what K value we use and it doesn’t really produce a model or show how important the predictor variables are to predicting the default outcome.

We have shown each model below as well as the most important variables:

We also used the varimp function to figure out which variables were deemed the most important as shown below. 



It was interesting to see that having no checking account, having a critical account/ other credits existing, using the loan for a used car, having a debt memo of over 200, or using the loan to get a television were top 5 factors in determining whether one would default or not.

Here we can see the importance was the general category rather than the specific variable. The top five included the status of existing checking account, credit history, saving account bonds, property, and whether one has other installment plans. Decision tree and Neural Network has the same rankings as SVM did. 

*Determining the Best Model:*

We used the accuracy to determine the best model. In each of our code, after the models we performed cross validation to determine how well it fit the model. Cross validation basically entails partitioning the data set into subsamples. You test one of the subsample using the rest of the subsamples as the training data. For most models we simply used the number of datapoints - 1. However, since neural networks is computationally intensive we only used 50 to get the average of the 50 models made to determine accuracy. Our results for the accuracy are listed below:

|  |  |
| --- | --- |
| Accuracy | |
| Logistic Regression | 76.57% |
| SVM Model | 74.40% |
| Decision Tree | 75.5% |
| Neural Networks | 70.1% |

Thus, the two best models with the highest accuracy are Decision Trees and Logistic Regression. We then produced a confusion matrix for both our logistic regression and decision tree models.

|  |  |  |
| --- | --- | --- |
| DT | Pred Will Pay Back | Pred Default |
| Actually Paid Back | 656 | 238 |
| Actually defaulted | 44 | 62 |

|  |  |  |
| --- | --- | --- |
| LR | Pred Will Pay Back | Pred Default |
| Actually Paid Back | 626 | 160 |
| Actually defaulted | 74 | 140 |

We accomplished this by using our model to predict it on the test data and then get the metrics. From the two confusion matrices we can see that by using the Decision Tree we would be rejecting more people to give out a loan even if they would actually pay their loan back than we would by using our logistic regression model. However, with Logistic Regression we would also be giving out more loans to people who will actually default on their loan. Thus the Decision Tree is more conservative but the Logistic Regression is more accurate overall. Depending on the how risk taking or risk averse the bank is as well as the cost-benefit analysis, it may choose to use either.

**Conclusion:**

Overall we believe this project best encapsulated all the material learned about machine learning in class. We were able to look at various machine learning models, apply them to a Finance concept, as well as assess their effectiveness. Additionally, we were close with our hypotheses as our predicted worst fitting model was correct and our predicted best fitting model was close. The most interesting part of the project was analyzing the data as well as learning basic coding in R. We both learned a lot and hope to use this base knowledge and apply it to more complex machine learning problems in finance. The most difficult part of the project was finding a complete dataset and cleaning up the data.

**Appendix**

Code:

<https://github.com/harshinimalli/fintech.project>

Variables:

|  |  |
| --- | --- |
| Attribute 1:  Status of existing checking account  A11 : ... < 0 DM  A12 : 0 <= ... < 200 DM  A13 : ... >= 200 DM / salary assignments for at least 1 year  A14 : no checking account  Attribute 2:  Duration in month  Attribute 3:  Credit history  A30 : no credits taken/ all credits paid back duly  A31 : all credits at this bank paid back duly  A32 : existing credits paid back duly till now  A33 : delay in paying off in the past  A34 : critical account/ other credits existing (not at this bank)  Attribute 4:  Purpose  A40 : car (new)  A41 : car (used)  A42 : furniture/equipment  A43 : radio/television  A44 : domestic appliances  A45 : repairs  A46 : education  A47 : (vacation - does not exist?)  A48 : retraining  A49 : business  A410 : others  Attribute 5:  Credit amount  Attribute 6:  Savings account/bonds  A61 : ... < 100 DM  A62 : 100 <= ... < 500 DM  A63 : 500 <= ... < 1000 DM  A64 : .. >= 1000 DM  A65 : unknown/ no savings account  Attribute 7:  Present employment since  A71 : unemployed  A72 : ... < 1 year  A73 : 1 <= ... < 4 years  A74 : 4 <= ... < 7 years  A75 : .. >= 7 years  Attribute 8:  Installment rate in percentage of disposable income  Attribute 9:  Personal status and sex  A91 : male : divorced/separated  A92 : female : divorced/separated/married  A93 : male : single  A94 : male : married/widowed  A95 : female : single | Attribute 10:  Other debtors / guarantors  A101 : none  A102 : co-applicant  A103 : guarantor    Attribute 11:  Present residence since  Attribute 12:  Property  A121 : real estate  A122 : if not A121 : building society savings agreement/ life insurance  A123 : if not A121/A122 : car or other, not in attribute 6  A124 : unknown / no property  Attribute 13:  Age in years  Attribute 14:  Other installment plans  A141 : bank  A142 : stores  A143 : none  Attribute 15:  Housing  A151 : rent  A152 : own  A153 : for free  Attribute 16:  Number of existing credits at this bank  Attribute 17:  Job  A171 : unemployed/ unskilled - non-resident  A172 : unskilled - resident  A173 : skilled employee / official  A174 : management/ self-employed/  highly qualified employee/ officer  Attribute 18:  Number of people being liable to provide maintenance for  Attribute 19:  Telephone  A191 : none  A192 : yes, registered under the customers name  Attribute 20:  foreign worker  A201 : yes  A202 : no |