Data Mining Final Project

Your Name: Anasa Alamgir **Your G Number**: G01300460

Introduction

The National Bank has experienced customers defaulting on their loans and therefore seeks to predict whether an applicant will default on their loan in order to protect the Bank from large financial losses.This project aims to explore the factors that lead to loan default and use machine learning models to predict the chance of an applicant defaulting on their loan in the future.

The loans data set contains information on 3 and 5 year loans that were originated in 2017 from customers residing in the Middle Atlantic and Northeast regions of the United States. Exploratory data analysis can help to find the relationship between whether the applicant defaults and the various factors that affect them defaulting on the loan.

Data

The loans data set shows applicant data with information on the loan amount, installment amount, interest rate, loan purpose, application type, loan term, home ownership, annual income, current job years, debt to income ratio, total credit line, years of credit history, history of bankruptcy and history of tax liens.

In the exploratory data analysis, this report will be using these factors: loan amount, installments, interest rates, loan purpose, application type, term, home ownership, annual income, current job years, debt to income ratio, and years of credit history.

Some questions this report intends to answer:

- What are the factors related to applicants defaulting on their loan?
- Is it possible to predict whether a customer will default on their loan?
  - How accurate are these predictions?
  - How many errors is the model expected to produce?
- Are there any policies that the bank can implement to reduce the risk of loan default?

Summary of Results

The default response variable in this data frame is loan_default, which records whether an applicant has defaulted or not. This variable has also been coded with 'Yes' and 'No' factors. Therefore using visualization techniques this report will show which other factors can explain why some applicants default and others do not.

The data visualization results show that a lower loan amount defaulted more often, and applicants with lower median income defaulted on their loans as well. In terms of home ownership, renters defaulted more often than applicants that owned their home or paid mortgage on their home. Applicants with a higher debt ratio also defaulted since they have more debt that what they earn. Besides that, a higher interest rate also led to applicants defaulting on their loans. The predictive modelling section elaborates on the factors that significantly affect loan default and the steps that the bank might find beneficial to implement to reduce the risk of loan default.

Predictive Modeling

In order to find the most significant factors affecting applicants to default on their loans, this report implements Logistic Regression, Linear Discriminant Analysis and K-Nearest Neighbors algorithms. It also uses ROC and AUC to measure the accuracy of the model. The most accurate model will be used to find the most significant factors affecting loan default.

The ROC AUC of the Logistic Regression model is 99%, which is the highest of all the three models created. Based on the results of the vip function, there are four significant factors affecting loan default:

- interest rate
- term: five year
- installment
- loan amount

Among these four, interest rate and five year term are the most important factors. The model also predicts 8.1% of errors that might be made by the model (negative predictions that are actually positive). Applicant seem to be most affected by interest rates and loan term, therefore in order to reduce the risk of loan default, the bank can take these steps:

- Reduce the loan term so that applicants can pay off their loan earlier
- Reduce the loan interest rate so that applicants can pay off the entire loan before it falls into default.

Confusion Matrix for Logistic regression model

```
##           Truth
## Prediction yes  no
##        yes 352  23
##        no   31 622
```

ROC AUC for regression model

```
## # A tibble: 1 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 roc_auc binary         0.990
```

Exploratory Data Analysis

Question 1

Question: Is there a relationship between loan default and loan amount?

Answer: The distribution of loans that defaulted based on the loan amount shows that a greater number of loans that defaulted had a smaller loan amount. There are a lower number of loans which defaulted at a higher loan amount.

Question 2

Question: Is there a relationship between loan default and history of missed payments in the past 2 years?

Answer: The summary does not show a positive relationship between history of missed payments and whether the applicant defaulted. There is a higher number of loans that defaulted but the applicants did not have history of missed payments in the past 2 years. On the other hand, there is a much higher number of loans that did not default but the applicants had history of missed payments in the past 2 years.

Question 3

Question: Is there a relationship between loan default and annual income?

Answer: Yes, the summary clearly shows that applicants with a lower median income ($60,000) and lower average income ($67,818.80) had defaulted on their loans compared to applicants with higher average and median incomes.

Question 4

Question: How does loan default relate to home ownership?

Answer: The data indicates that applicants renting their home are more likely to default on their loans compared to a far lesser amount of applicants that own their home.

Question 5

Question: Is there a relationship between loan default and debt-to-income ratio?

Answer: The density plot shows that applicants with higher debt-to-income ratio defaulted more often than applicants with a lower debt to income ratio. This is most likely because applicants with higher debt to income ratio already have substantial amount of debt compared to their income to be able to pay on their installments. Applicants with much lower debt to income ratio are less likely to have defaulted on their loans.

Question 6

Question: How is loan default related to application type and years of credit history?

Answer: The summary indicates that individual applications that got defaulted had a smaller average credit history when compared to joint applications with lower credit history that went into default. On the other hand, for both types of applications, the ones with higher average credit history did not go into default.

Question 7

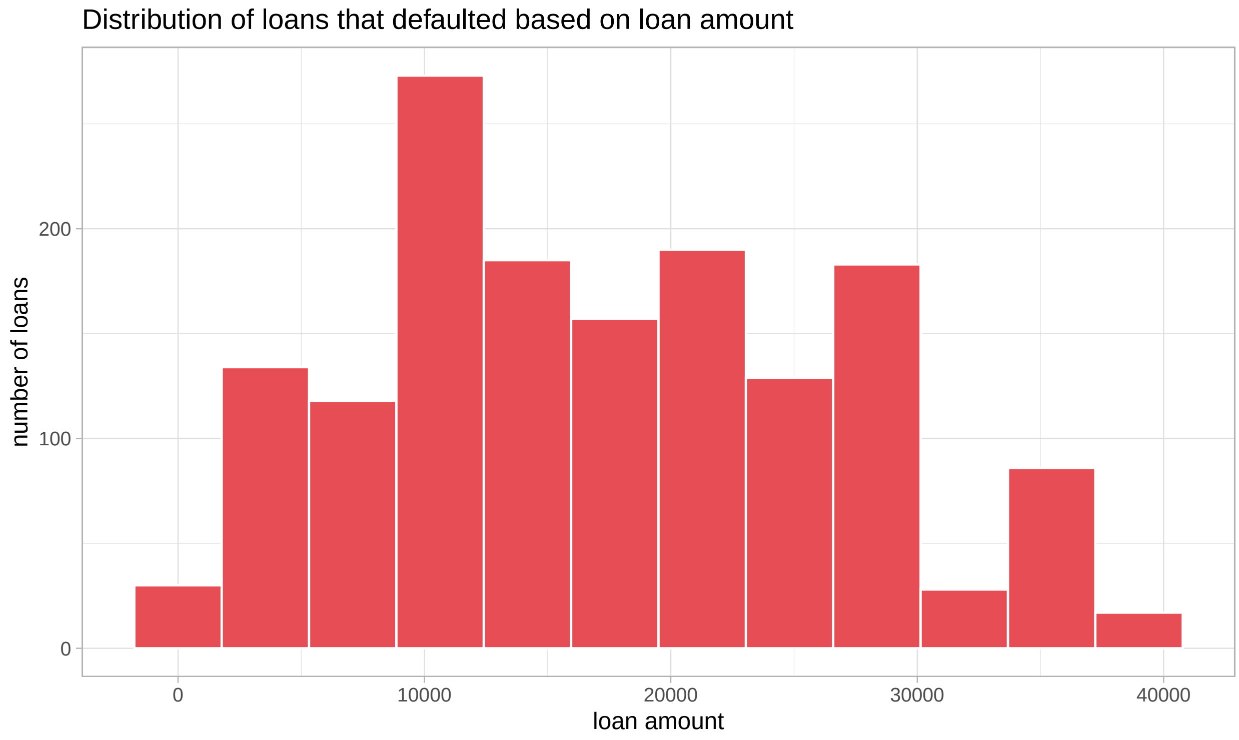
Question: Is there a relationship between loan default rate and interest rates and loan purpose?

Answer: Out of the applicants that defaulted on their loans, small business and credit card loans have the highest median interest rates. Therefore, it can be stated that applicants are more likely to default on their loans with a higher interest rate, especially when it is for a small business or credit card.

Appendix

Tables and visualizations

question 1: Is there a relationship between loan default and loan amount?



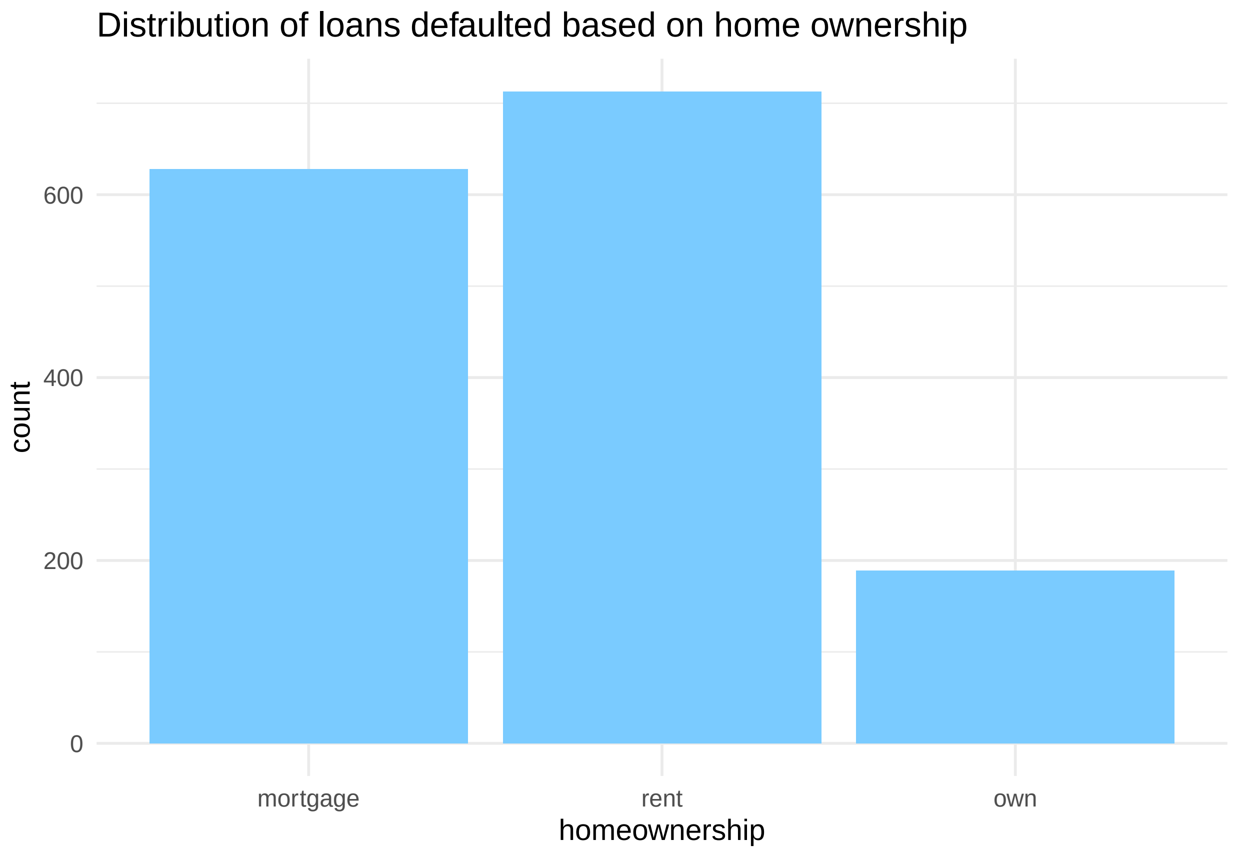
question 2: Is there a relationship between loan default and history of missed payments in the past 2 years?

```
## # A tibble: 4 x 3
## # Groups:   loan_default [2]
##   loan_default missed_payment_2_yr num_loans
##   <fct>        <fct>                   <int>
## 1 yes          yes                       212
## 2 yes          no                       1318
## 3 no           yes                       258
## 4 no           no                       2322
```

question 3: Is there a relationship between loan default and annual income?

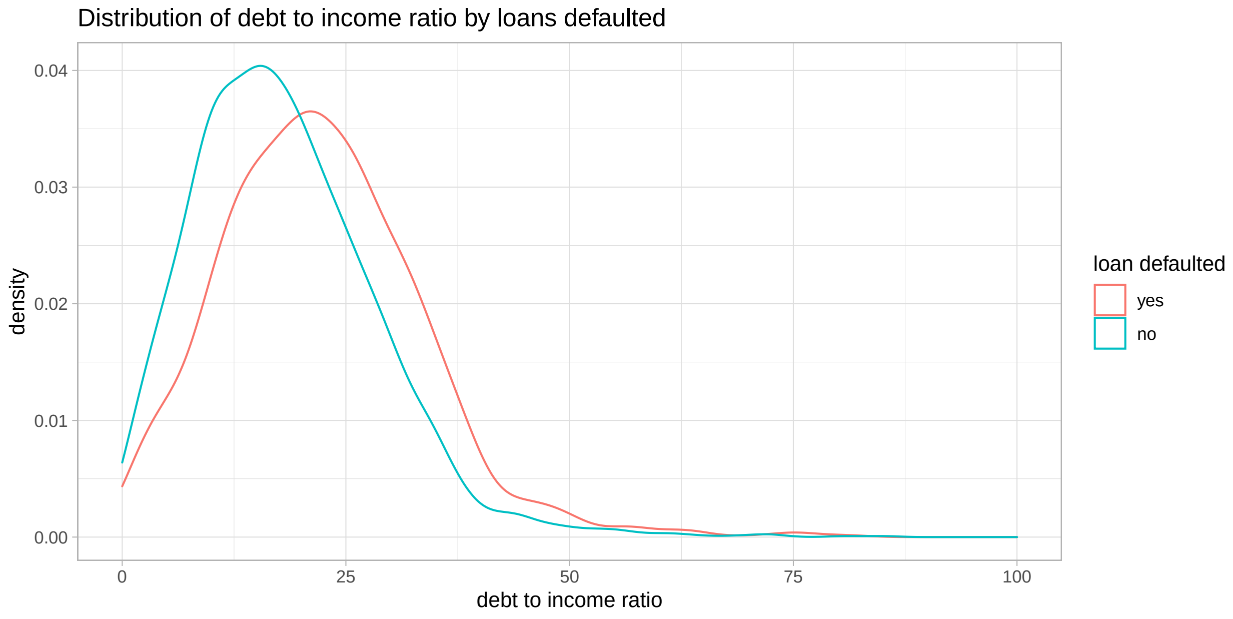
```
## # A tibble: 2 x 5
##   loan_default avg_income median_income min_income sd_income
##   <fct>             <dbl>         <dbl>      <dbl>     <dbl>
## 1 yes              67819.         60000       7500    34930.
## 2 no               76096.         69000       3000    38161.
```

question 4: How does loan default relate to home ownership?



question 5: Is there a relationship between loan default and debt-to-income ratio?

```
## Warning: Removed 9 rows containing non-finite values (stat_density).
```



question 6: How is loan default related to application type and years of credit history?

```
## # A tibble: 4 x 5
## # Groups:   loan_default [2]
##   loan_default application_type avg_credit_history median_cred_history
##   <fct>        <fct>                         <dbl>               <dbl>
## 1 yes          individual                     15.0                  14
## 2 yes          joint                          15.2                  14
## 3 no           individual                     16.1                  15
## 4 no           joint                          16.9                  15
## # ... with 1 more variable: sd_cred_history <dbl>
```

question 7: Is there a relationship between loan default rate and interest rates and loan purpose?

```
## # A tibble: 5 x 3
## # Groups:   loan_default [1]
##   loan_default loan_purpose       med_interestrate
##   <fct>        <fct>                         <dbl>
## 1 yes          debt_consolidation             14.8
## 2 yes          credit_card                    15  
## 3 yes          medical                        14.8
## 4 yes          small_business                 15.2
## 5 yes          home_improvement               14.8
```

Predictive models

Model 1 Logistic Regression

```
#split loans_df into training and test sets
set.seed(172)
loan_split <- initial_split(loans_df, prop = 0.75,
                            strata = loan_default)
loan_training <- loan_split %>% training()
loan_test <- loan_split %>% testing()
```

```
#cross validation folds for hyperparameter tuning
set.seed(172)
loan_folds <- vfold_cv(loan_training, v = 6)
```

```
#feature engineering
loan_recipe <- recipe(loan_default ~ .,data = loan_training) %>%
  step_YeoJohnson(all_numeric(), -all_outcomes()) %>%
  step_normalize(all_numeric(), -all_outcomes()) %>%
  step_dummy(all_nominal(), -all_outcomes())
loan_recipe %>%
  prep(training = loan_training) %>%
  bake(new_data = NULL)
```

```
## # A tibble: 3,082 x 20
##    loan_amount installment interest_rate annual_income current_job_years
##          <dbl>       <dbl>         <dbl>         <dbl>             <dbl>
##  1     -1.21       -1.20         -0.849          0.117            -0.397
##  2      0.0358      0.0759        0.0344        -0.827             1.10 
##  3      1.65        1.01         -1.08           2.22              1.10 
##  4     -0.347      -0.821        -0.924         -1.17             -0.121
##  5     -0.531      -0.527        -1.08          -0.179              1.10
##  6     -1.04       -1.03         -0.158         -0.445             -0.691
##  7     -0.531      -0.498        -0.704          0.503              1.10
##  8     -2.20       -2.38          0.578         -0.0269            -0.691
##  9     -1.40       -1.44          0.692          0.117             -0.121
## 10      1.89        2.16          0.0971         2.16               0.873
## # ... with 3,072 more rows, and 15 more variables: debt_to_income <dbl>,
## #   total_credit_lines <dbl>, years_credit_history <dbl>, loan_default <fct>,
## #   loan_purpose_credit_card <dbl>, loan_purpose_medical <dbl>,
## #   loan_purpose_small_business <dbl>, loan_purpose_home_improvement <dbl>,
## #   application_type_joint <dbl>, term_five_year <dbl>,
## #   homeownership_rent <dbl>, homeownership_own <dbl>,
## #   missed_payment_2_yr_no <dbl>, history_bankruptcy_no <dbl>, ...
```

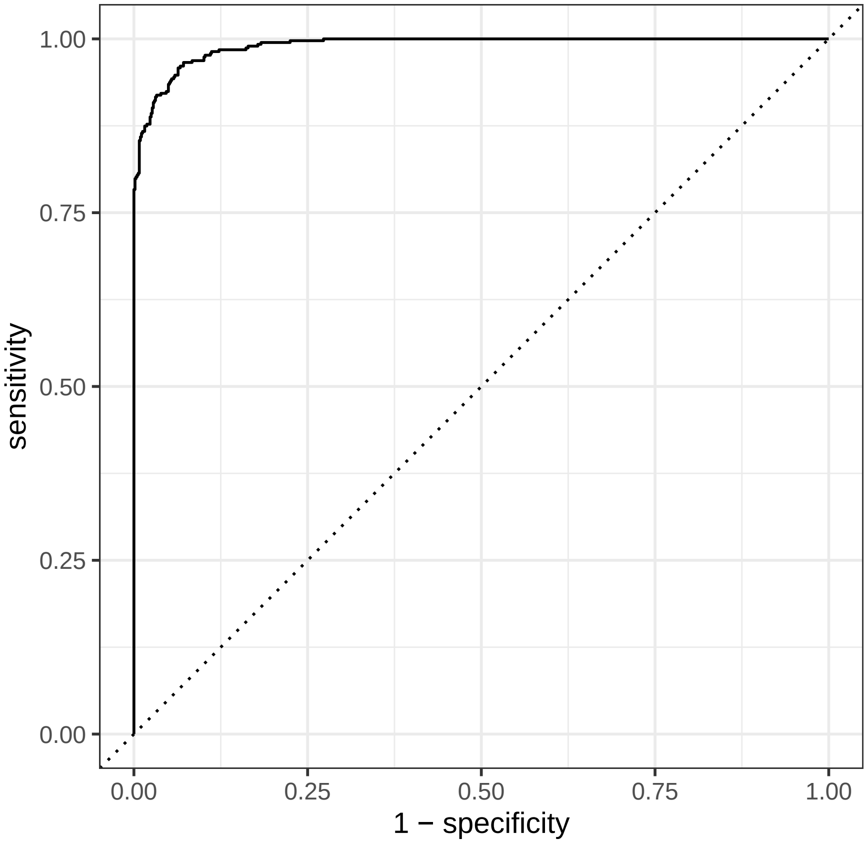
```r
#specify model
loan_logistic <- logistic_reg() %>%
  set_engine('glm') %>%
  set_mode('classification')
```

```r
#create workflow
logistic_wf <- workflow() %>%
  add_model(loan_logistic) %>%
  add_recipe(loan_recipe)
#roc curve and auc
logistic_fit <- logistic_wf %>% last_fit(split = loan_split)
# collect predictions
logistic_predictions <- logistic_fit %>% collect_predictions()
```

```r
#confusion matrix
conf_mat(logistic_predictions,
         truth = loan_default,
         estimate = .pred_class)
```

```
##           Truth
## Prediction yes  no
##        yes 352  23
##        no   31 622
```

```r
#roc curve
roc_curve(logistic_predictions, truth = loan_default, estimate = .pred_yes) %>%
  autoplot()
```



```
#area under curve
roc_auc(logistic_predictions, truth = loan_default, .pred_yes)
```

```
## # A tibble: 1 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 roc_auc binary         0.990
```

Accuracy: 99%

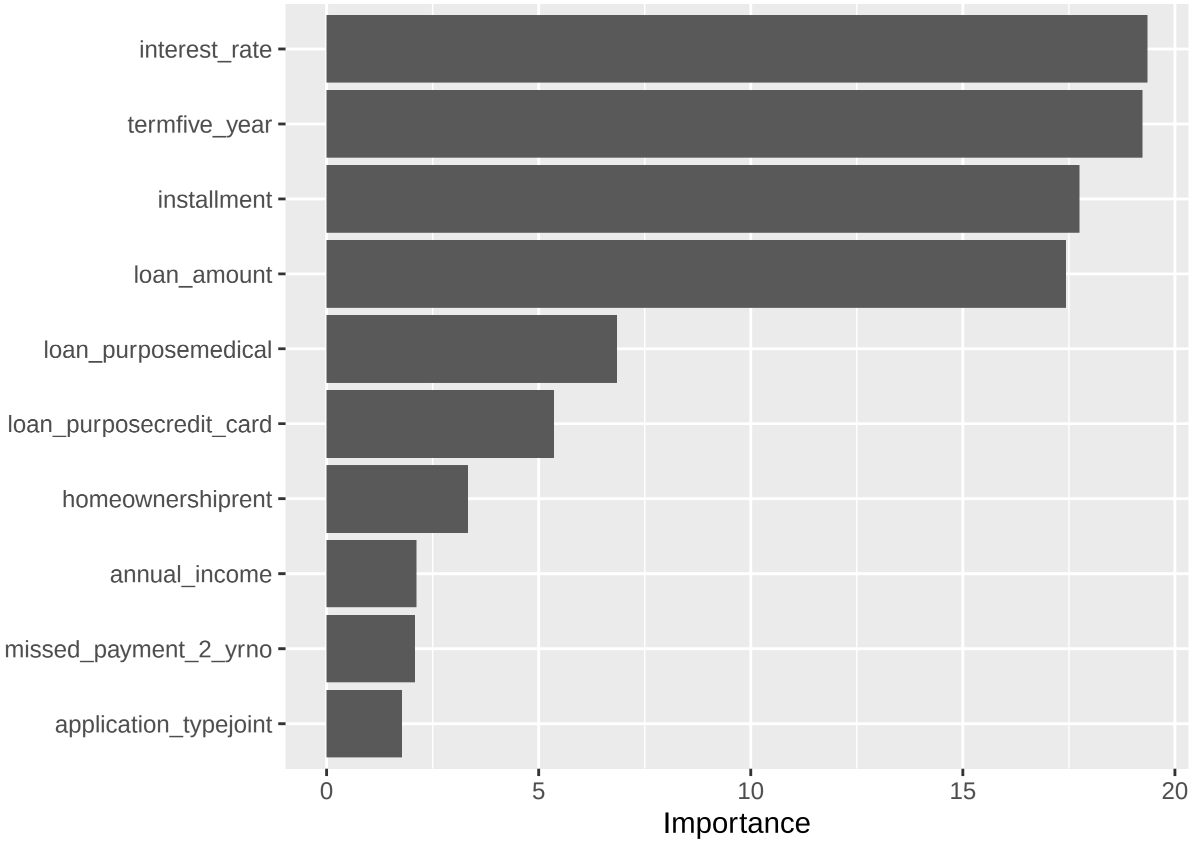
```
#model summary
log_model <- glm(loan_default ~., data = loan_training, family = binomial())
tidy(log_model)
```

```
## # A tibble: 20 x 5
##    term                          estimate  std.error statistic  p.value
##    <chr>                            <dbl>      <dbl>     <dbl>    <dbl>
##  1 (Intercept)                  10.4       0.885         11.7   1.23e-31
##  2 loan_amount                   0.00108   0.0000620     17.4   4.76e-68
##  3 installment                  -0.0342    0.00193      -17.7   1.87e-70
##  4 interest_rate                -0.698     0.0361       -19.3   2.04e-83
##  5 loan_purposecredit_card      -1.12      0.210         -5.36  8.33e- 8
##  6 loan_purposemedical          -1.58      0.231         -6.84  7.80e-12
##  7 loan_purposesmall_business   -0.0414    0.219         -0.189 8.50e- 1
##  8 loan_purposehome_improvement  0.0288    0.248          0.116 9.08e- 1
##  9 application_typejoint        -0.392     0.222         -1.77  7.67e- 2
## 10 termfive_year                -7.32      0.381        -19.2   2.00e-82
## 11 homeownershiprent            -0.560     0.168         -3.32  8.84e- 4
## 12 homeownershipown             -0.282      0.239         -1.18  2.38e- 1
## 13 annual_income                 0.00000528 0.00000250     2.11  3.48e- 2
## 14 current_job_years             0.00343    0.0211         0.162 8.71e- 1
## 15 debt_to_income               -0.00727    0.00453       -1.61  1.08e- 1
## 16 total_credit_lines            0.00624    0.00689        0.906 3.65e- 1
## 17 years_credit_history          0.0178     0.0115         1.55  1.22e- 1
## 18 missed_payment_2_yrno         0.457      0.219          2.09  3.70e- 2
## 19 history_bankruptcyno         -0.0973     0.223         -0.436 6.63e- 1
## 20 history_tax_liensno           0.116      0.631          0.183 8.54e- 1
```

```
#get summary
summary(log_model)
```

```
## 
## Call:
## glm(formula = loan_default ~ ., family = binomial(), data = loan_training)
## 
## Deviance Residuals: 
##      Min        1Q    Median        3Q       Max  
## -3.08458  -0.11368   0.05583   0.27167   3.02825  
## 
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)    
## (Intercept)                   1.035e+01  8.848e-01  11.703  < 2e-16 ***
## loan_amount                   1.081e-03  6.201e-05  17.431  < 2e-16 ***
## installment                  -3.420e-02  1.927e-03 -17.746  < 2e-16 ***
## interest_rate                -6.976e-01  3.605e-02 -19.350  < 2e-16 ***
## loan_purposecredit_card      -1.124e+00  2.097e-01  -5.360 8.33e-08 ***
## loan_purposemedical          -1.578e+00  2.306e-01  -6.842 7.80e-12 ***
## loan_purposesmall_business   -4.135e-02  2.189e-01  -0.189 0.850197    
## loan_purposehome_improvement  2.883e-02  2.483e-01   0.116 0.907570    
## application_typejoint        -3.924e-01  2.217e-01  -1.770 0.076731 .  
## termfive_year                -7.319e+00  3.805e-01 -19.232  < 2e-16 ***
## homeownershiprent            -5.600e-01  1.684e-01  -3.325 0.000884 ***
## homeownershipown             -2.819e-01  2.390e-01  -1.179 0.238259    
## annual_income                 5.277e-06  2.500e-06   2.110 0.034833 *  
## current_job_years             3.427e-03  2.109e-02   0.162 0.870928    
## debt_to_income               -7.275e-03  4.529e-03  -1.606 0.108236    
## total_credit_lines            6.242e-03  6.891e-03   0.906 0.365020    
## years_credit_history          1.778e-02  1.149e-02   1.547 0.121883    
## missed_payment_2_yrno         4.569e-01  2.191e-01   2.086 0.037008 *  
## history_bankruptcyno         -9.730e-02  2.230e-01  -0.436 0.662595    
## history_tax_liensno           1.157e-01  6.310e-01   0.183 0.854456    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 4068.8  on 3081  degrees of freedom
## Residual deviance: 1255.8  on 3062  degrees of freedom
## AIC: 1295.8
## 
## Number of Fisher Scoring iterations: 7
```

```
# get vip
vip(log_model)
```



Model 2 LDA

```
#specify model
loan_lda <- discrim_regularized(frac_common_cov = 1) %>%
  set_engine("klaR") %>%
  set_mode("classification")
```

```
#workflow
lda_wf <- workflow() %>% add_model(loan_lda) %>% add_recipe(loan_recipe)
```

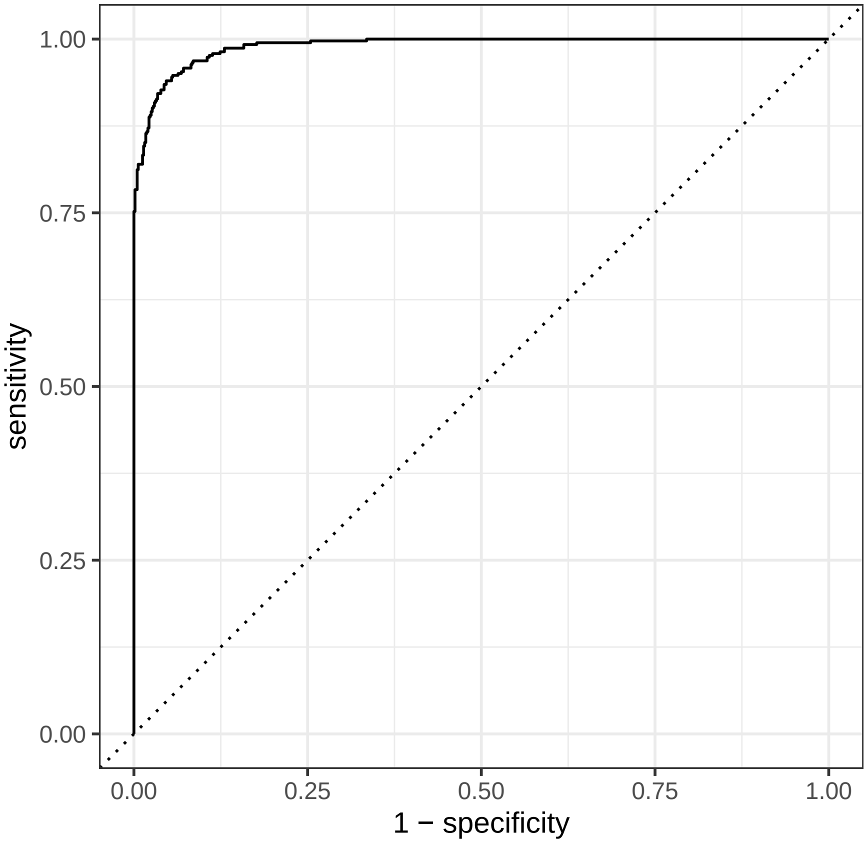
```
#fit workflow
lda_fit <- lda_wf %>% last_fit(split=loan_split)
```

```
#collect predictions
lda_predictions <- lda_fit %>% collect_predictions()
```

```
#confusion matrix
conf_mat(lda_predictions, truth = loan_default, estimate = .pred_class)
```

```
##             Truth
## Prediction yes  no
##        yes 348  19
##        no   35 626
```

```r
#roc curve
roc_curve(lda_predictions, truth=loan_default, estimate= .pred_yes) %>%
  autoplot()
```



```r
#Area under ROC
roc_auc(lda_predictions, truth=loan_default, .pred_yes)
```

```
## # A tibble: 1 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 roc_auc binary         0.990
```

Accuracy: 98.99%

Model 3 K-Nearest Neighbor

```r
#specify model
knn_model <- nearest_neighbor(neighbors = tune()) %>%
  set_engine("kknn") %>%
  set_mode("classification")
```

```r
#workflow
knn_wf <- workflow() %>%
  add_model(knn_model) %>%
  add_recipe(loan_recipe)
```

```
#create grid for hyperparameter testing
k_grid <- tibble(neighbors = c(10,20,30,40,50,75,100,125,150))
```

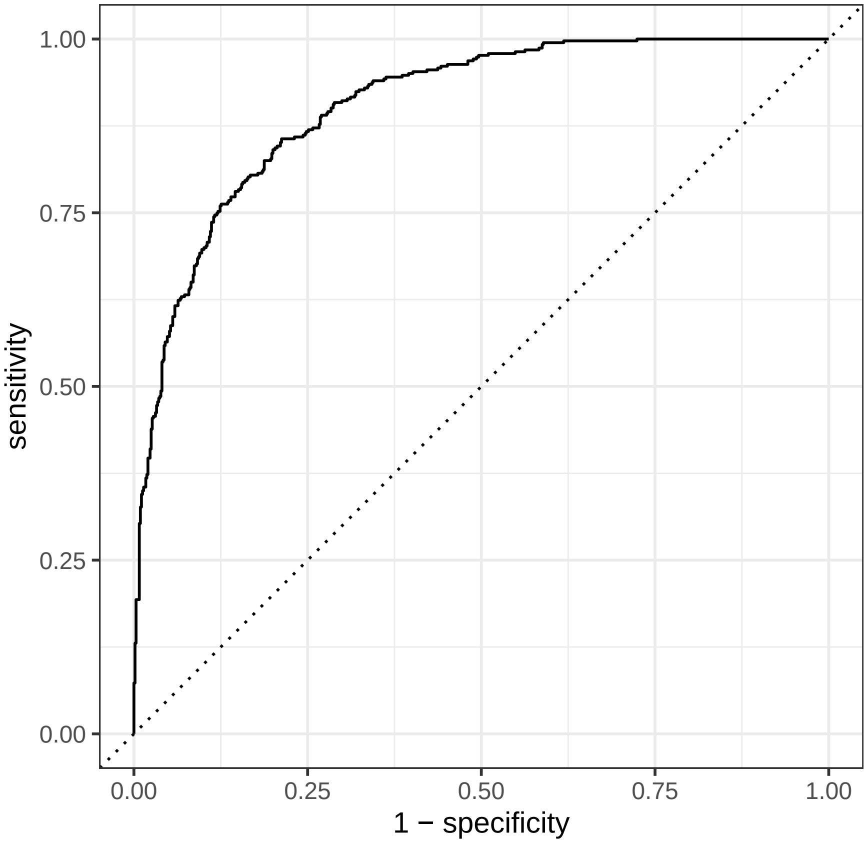
```
#tuning wf
set.seed(271)
knn_tuning <- knn_wf %>% tune_grid(resamples=loan_folds, grid=k_grid)
#select best model from tuning result
best_k <- knn_tuning %>% select_best(metric='roc_auc')
#add optimal model to wf
final_knn_wf <- knn_wf %>% finalize_workflow(best_k)
#fit model
knn_fit <- final_knn_wf %>% last_fit(split=loan_split)
```

```
#get df of test prediction results
knn_predictions <- knn_fit %>% collect_predictions()
```

```
#confusion matrix
conf_mat(knn_predictions, truth= loan_default, estimate = .pred_class)
```

```
##           Truth
## Prediction yes  no
##        yes 213  28
##        no  170 617
```

```
#roc curve
roc_curve(knn_predictions, truth = loan_default, estimate= .pred_yes) %>%
  autoplot()
```



```
#Are under ROC
roc_auc(knn_predictions, truth = loan_default, estimate= .pred_yes)
```

```
## # A tibble: 1 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 roc_auc binary         0.904
```

Accuracy: 90%

— End of the Project —