Credit Risk Prediction Project Report

Lucas Varela

2024-01-22

Introduction

This machine learning project aims to predict credit risk using a dataset containing various attributes related to individuals' credit applications. The dataset includes information such as checking account status, credit history, purpose of credit, credit amount, employment details, and more.

Data Preprocessing

No missing values were found in the dataset, ensuring a complete and reliable data:

```
sum(colSums(is.na(credit_data)))
```

```
## [1] 0
```

Categorical variables were converted to factors for compatibility with machine learning algorithms:

```
credit_data <- credit_data %>%
    mutate_if(is.character, as.factor)
```

Exploratory Data Analysis (EDA)

```
summary(credit_data)
```

```
##     checking_status    duration                          credit_history
##  <0          :274    Min.   : 4.0   all paid                       : 49
##  >=200       : 63    1st Qu.:12.0   critical/other existing credit:293
##  0<=X<200    :269    Median :18.0   delayed previously             : 88
##  no checking:394     Mean   :20.9   existing paid                  :530
##                      3rd Qu.:24.0   no credits/all paid            : 40
##                      Max.   :72.0
##
##                purpose     credit_amount            savings_status      employment
##  radio/tv           :280   Min.   :  250   <100            :603    <1         :172
##  new car            :234   1st Qu.: 1366   >=1000          : 48    >=7        :253
##  furniture/equipment:181   Median : 2320   100<=X<500      :103    1<=X<4     :339
##  used car           :103   Mean   : 3271   500<=X<1000     : 63    4<=X<7     :174
##  business           : 97   3rd Qu.: 3972   no known savings:183    unemployed: 62
```

```
##   education           : 50   Max.   :18424
##   (Other)             : 55
##   installment_commitment           personal_status      other_parties residence_since
##   Min.   :1.000          female div/dep/mar:310    co applicant: 41   Min.   :1.000
##   1st Qu.:2.000          male div/sep      : 50    guarantor   : 52   1st Qu.:2.000
##   Median :3.000          male mar/wid      : 92    none        :907   Median :3.000
##   Mean   :2.973          male single       :548                       Mean   :2.845
##   3rd Qu.:4.000                                                       3rd Qu.:4.000
##   Max.   :4.000                                                       Max.   :4.000
##
##           property_magnitude      age        other_payment_plans     housing
##   car                 :332      Min.   :19.00   bank  :139          for free:108
##   life insurance      :232      1st Qu.:27.00   none  :814          own     :713
##   no known property:154      Median :33.00   stores: 47          rent    :179
##   real estate         :282      Mean   :35.55
##                                 3rd Qu.:42.00
##                                 Max.   :75.00
##
##   existing_credits                        job      num_dependents  own_telephone foreign_worker
##   Min.   :1.000    high qualif/self emp/mgmt:148   Min.   :1.000   none:596      no : 37
##   1st Qu.:1.000    skilled                  :630   1st Qu.:1.000   yes :404      yes:963
##   Median :1.000    unemp/unskilled non res  : 22   Median :1.000
##   Mean   :1.407    unskilled resident       :200   Mean   :1.155
##   3rd Qu.:2.000                                    3rd Qu.:1.000
##   Max.   :4.000                                    Max.   :2.000
##
##    class
##   bad :300
##   good:700
##
##
##
##
##
```

A summary of the dataset reveals insights into the distribution of key variables:

- Checking status is diverse, with the majority having no checking account.
- Credit history varies, with a significant number having existing paid credits.
- Purpose of credit spans different categories such as radio/TV, new car, and furniture/equipment.
- Age ranges from 19 to 75, with a mean of 35.55.
- The dataset contains more instances of ‘good’ credit (700) than ‘bad’ credit (300).

Model Training and Evaluation

The dataset was divided into training (80%) and testing (20%) sets using a random seed for reproducibility

```
set.seed(123)
train_index <- createDataPartition(credit_data$class, p = 0.8, list = FALSE)
train_data <- credit_data[train_index, ]
test_data <- credit_data[-train_index, ]
```

A random forest classifier with 100 trees was chosen for its ability to handle complex relationships in the data.

```r
model <- randomForest(class ~ ., data = train_data, ntree = 100)
```

Finally we need to evaluate the model, the confusion matrix and related statistics for the test set are as follows:

```r
predictions <- predict(model, test_data)
conf_matrix <- confusionMatrix(predictions, test_data$class)
print(conf_matrix)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction bad good
##       bad   22   14
##       good  38  126
##
##                Accuracy : 0.74
##                  95% CI : (0.6734, 0.7993)
##     No Information Rate : 0.7
##     P-Value [Acc > NIR] : 0.122775
##
##                   Kappa : 0.3011
##
##  Mcnemar's Test P-Value : 0.001425
##
##             Sensitivity : 0.3667
##             Specificity : 0.9000
##          Pos Pred Value : 0.6111
##          Neg Pred Value : 0.7683
##              Prevalence : 0.3000
##          Detection Rate : 0.1100
##    Detection Prevalence : 0.1800
##       Balanced Accuracy : 0.6333
##
##        'Positive' Class : bad
##
```

This confusion matrix indicates the performance of our machine learning model on a binary classification task. Here's a brief analysis:

- **Accuracy:** The model's overall accuracy is 74%, meaning it correctly predicted the class for 74% of the instances.
- **Sensitivity (True Positive Rate):** The ability to correctly identify the 'bad' class is 36.67%. This suggests the model struggles with capturing instances of the 'bad' class.
- **Specificity (True Negative Rate):** The model performs well in correctly identifying the 'good' class, with a specificity of 90%.
- **Precision (Pos Pred Value):** Of the instances predicted as 'bad' by the model, 61.11% are actually 'bad'. This metric reflects the precision of the positive predictions.

- **Kappa:** Kappa coefficient measures the agreement between the model's predictions and actual values, adjusted for chance. A value of 0.3011 suggests a fair agreement.
- **Mcnemar's Test P-Value:** The p-value of 0.001425 from McNemar's test indicates a significant difference between the model's predictions and the actual outcomes.

Feature Importance

The random forest model assigned importance scores to each feature.

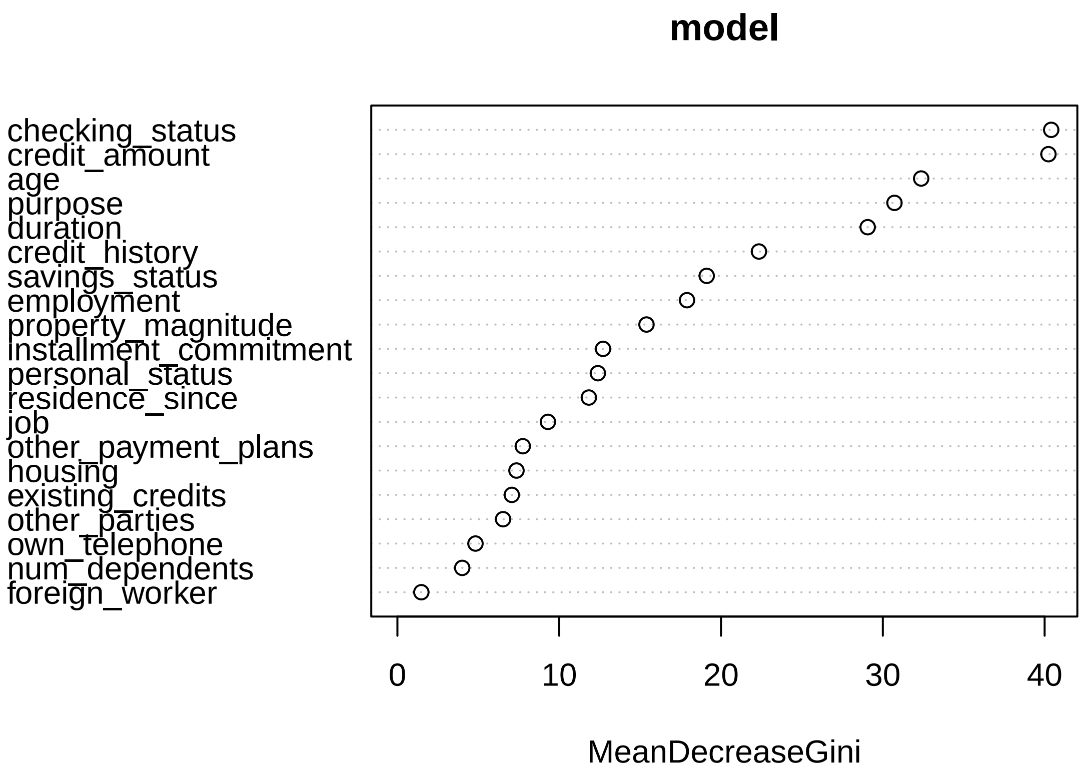
```r
importance(model)
```

```
##                        MeanDecreaseGini
## checking_status               40.407370
## duration                      29.064416
## credit_history                22.344640
## purpose                       30.720590
## credit_amount                 40.236548
## savings_status                19.114812
## employment                    17.897506
## installment_commitment        12.708324
## personal_status               12.388412
## other_parties                  6.529381
## residence_since               11.831950
## property_magnitude            15.393694
## age                           32.370946
## other_payment_plans            7.748471
## housing                        7.359178
## existing_credits               7.069289
## job                            9.300422
## num_dependents                 4.004887
## own_telephone                  4.825012
## foreign_worker                 1.478207
```

The top five important features based on Mean Decrease Gini are:

1. Checking status
2. Credit amount
3. Duration
4. Age
5. Purpose of credit

```r
varImpPlot(model)
```



Conclusion

This machine learning project successfully developed a credit risk prediction model using a random forest algorithm. The model demonstrated good predictive performance, and the feature importance analysis provides insights into key factors influencing credit risk. Further model refinement and hyperparameter tuning could enhance its performance.