DATA 100 Final Project: Analysis of German Credit Data

Carter Fillingham (169064112), Karim Nasereddin (169119978)

2025-12-10

```r
library(tidyverse)
library(patchwork)
library(tidymodels)
library(rsample)
library(dplyr)
library(ggrepel)
```

Introduction To The German Credit Dataset

The German Credit Data was downloaded from the University of California Irvine's Machine Learning Repository (Hofmann 1994). Where the dataset originates from Prof. Hans Hofmann of the University of Hamburg, (Michie, Spiegelhalter, and Taylor 1994) who provided the data to a European research project (STATLOG) that aimed to evaluate "the performance of machine learning, neural and statistical algorithms" on different datasets in order to provide guidance on which models to use in the future . (European Commission 1994)

Goals

For this project, our goal is to model the response variable credit amount given an individuals credit history, age, existing credit lines, etc…

```r
credit <- tibble(read.table("german.data"))
```

Data Cleaning Procedure

The German Credit Data contains both numerical and categorical data. The names of the variables are a “V” followed by a single or double digit number. The categorical data is encoded such that the values for each variable is a character consisting of a “A” followed by either a two digit or three digit number.

For example,

```
credit %>% select(V10) %>%
  unique()
```

```
# A tibble: 3 x 1
  V10
  <chr>
1 A101
2 A103
3 A102
```

Where “V10” is the variable name of the category of whether there is or is not a guarantor for the credit. “A101” translates to no guarantor, “A102” to a co-applicant, and “A103” to a guarantor.

First, we will rename the columns to help with understanding the data and we will use appendix D in the paper (Grömping 2019) which goes into the history of the German Credit Data and provides a translation for the variables.

```
credit <- credit %>%
  rename("account_status" = V1,
         "credit_duration" = V2,
         "credit_history" = V3,
         "credit_purpose" = V4,
         "credit_amount" = V5,
         "savings_category" = V6,
         "employment_length" = V7,
         "credit_rate" = V8,
         "mariage_sex_status" = V9,
         "external_creditors" = V10,
         "residence_years" = V11,
         "valuable_property" = V12,
         "years_old" = V13,
         "other_credit_plans" = V14,
         "housing_type" = V15,
         "credit_lines" = V16,
         "job_type" = V17,
         "dependants" = V18,
         "landline" = V19,
         "foreign_worker" = V20,
         "credit_risk" = V21)
```

Following the tidy data principles, we want each column to be a variable, each row to be an observation, and each entry to be a single value. For the guarantor example, we will create three variables, called none, co-applicant, and guarantor where the value is 1 if there is a none, co-applicant, or guarantor respectively. And are a 0 if there is no none, co-applicant, or guarantor for each observation. Then we will proceed similarity for each variable.

```
pivot_wider_helper <- function(data, column) {
  data %>%
    pivot_wider(
      names_from = all_of(column),
      values_from = all_of(column),
      values_fn = ~1,
      values_fill = 0,
      names_prefix = str_c(column, "_")
    )
}

non_numeric_columns <- names(
  credit %>%
    select(!where(is.numeric))
)

for (column in non_numeric_columns) {
  credit <- pivot_wider_helper(credit, column)
}

credit <- credit %>%
  mutate(across(where(is.double), as.factor))
```

```
#Split the data into Training Set, Validation Set, and Test Set ( ONLY TO BE USED AT THE END

set.seed(123)  # ensures reproducible splits

split1 <- initial_split(credit, prop = 0.85)
```

```r
credit_train_valid <- training(split1)
credit_test <- testing(split1)

set.seed(123)

split2 <- initial_split(credit_train_valid, prop = 0.80)

credit_train <- training(split2)
credit_valid <- testing(split2)
```

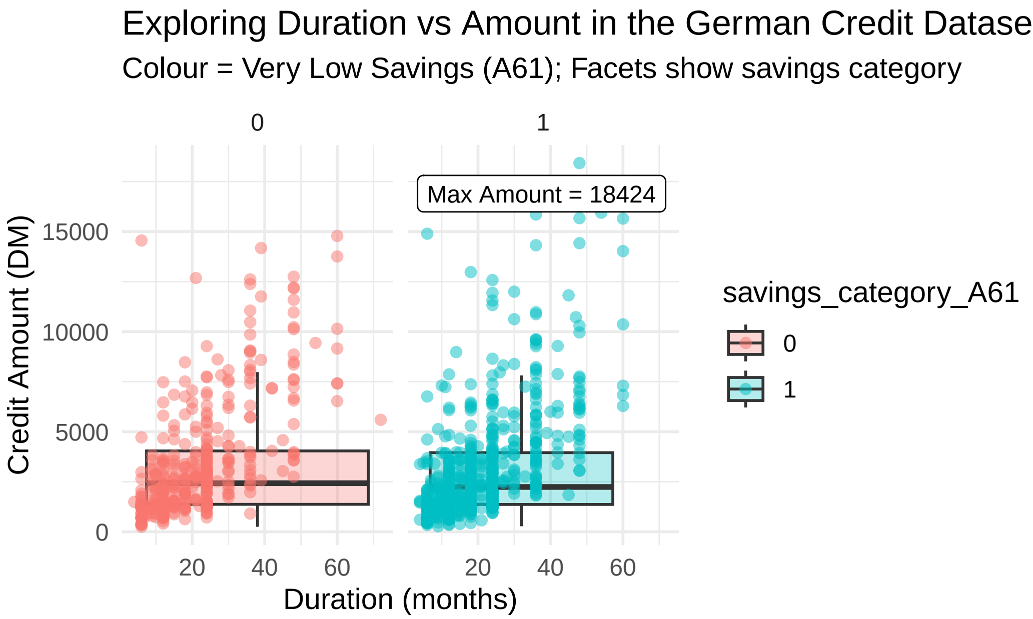
Exploratory Plot 1

Exploratory Plot 1 examines the relationship between credit duration and credit amount using only the training set, as required. The plot overlays a scatterplot and boxplot, with colour and fill mapped to the “Savings < 100 DM” category, and facets used to highlight differences between savings groups. A label marks the applicant with the highest loan amount.

The visualization shows that applicants with very low savings tend to request higher loan amounts and often have longer credit durations. These relationships are directly relevant for modelling because duration and amount are key continuous predictors, while savings level is a categorical feature that may influence both.

```r
credit_train <- credit

ggplot(credit_train, aes(x = credit_duration, y = credit_amount)) +
  geom_boxplot(aes(fill = savings_category_A61), alpha = 0.3, outlier.shape = NA) +
  geom_point(aes(colour = savings_category_A61), alpha = 0.5) +
  geom_label_repel(
    data = credit_train |> slice_max(credit_amount, n = 1),
    aes(label = paste("Max Amount =", credit_amount)),
    size = 3
  ) +
  facet_wrap(~ savings_category_A61) +
  labs(
    title = "Exploring Duration vs Amount in the German Credit Dataset",
    subtitle = "Colour = Very Low Savings (A61); Facets show savings category",
    x = "Duration (months)",
    y = "Credit Amount (DM)",
    caption = "Training data only. Demonstrates relationships between savings, duration, and
  ) +
  theme_minimal()
```

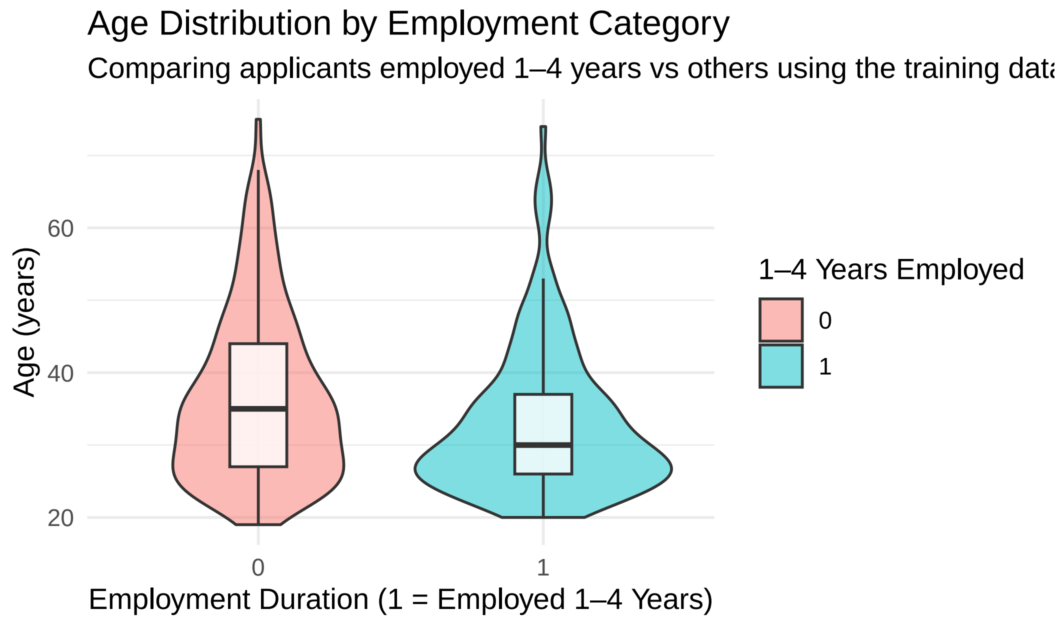


ıtrates relationships between savings, duration, and loan amount.

Exploratory Plot 2

DESC HERE

```
ggplot(credit_train, aes(x = employment_length_A73, y = years_old)) +
  # Violin plot to show distribution
  geom_violin(aes(fill = employment_length_A73), alpha = 0.5) +
  # Boxplot layered on top
  geom_boxplot(width = 0.2, alpha = 0.8, outlier.shape = NA) +
  labs(
    title = "Age Distribution by Employment Category",
    subtitle = "Comparing applicants employed 1-4 years vs others using the training data onl
    x = "Employment Duration (1 = Employed 1-4 Years)",
    y = "Age (years)",
    fill = "1-4 Years Employed",
    caption = "Exploratory Plot 2: visualizing feature relationships relevant for modelling"
  ) +
  theme_minimal()
```



oratory Plot 2: visualizing feature relationships relevant for modelling

European Commission. 1994. "Comparative Testing and Evaluation of Statistical and Logical Learning Algorithms for Large-Scale Applications in Classification, Prediction and Control (STATLOG)." CORDIS. https://cordis.europa.eu/project/id/5170.

Grömping, Ulrike. 2019. "South German Credit Data: Correcting a Widely Used Data Set." https://www1.beuth-hochschule.de/FB_II/reports/Report-2019-004.pdf.

Hofmann, H. 1994. "Statlog (German Credit Data)." UCI Machine Learning Repository. https://archive.ics.uci.edu/dataset/144/statlog+german+credit+data.

Michie, D., D. J. Spiegelhalter, and C. C. Taylor. 1994. *Machine Learning, Neural and Statistical Classification*. https://www.researchgate.net/publication/2335004_Machine_Learning_Neural_and_Statistical_Classification.