

STAT 350: Dataset Description

The dataset provided for student use has been modified and sourced primarily from a German credit status prediction dataset hosted at the University of California at Irvine Machine Learning Repository¹. The UCI Machine Learning Repository was created in 1987 by David Aha and his team and serves as repository of data sets, algorithms, and tools for machine learning research.

The objective of this dataset is to apply statistical tools to analyze credit-related information. The dataset has been preprocessed to make it more accessible. To address the inherent variability in the credit amount extended to customers, we have performed a natural logarithm transformation on several variables.

Students can use this dataset to explore the relationship between credit amounts and debtor attributes. The dataset provides a valuable opportunity for students to apply their knowledge of data science and statistical modeling to real-world credit scenarios.

Please note that the dataset has been carefully curated for educational purposes. It serves as a valuable resource for students to enhance their analytical skills and gain practical experience in handling credit-related data. The analysis of this dataset can contribute to a better understanding of credit dynamics and aid in making informed decisions in the realm of credit evaluation and risk assessment.

¹ Hofmann, Hans. (1994). Statlog (German Credit Data). UCI Machine Learning Repository. <https://doi.org/10.24432/C5NC77>.

The following table describes the columns/variables in this dataset:

Variable Name	Dataset Column	Description
age	1	The age in years of the debtor
checkingStatus	2	Status of checking account
creditAmountDollar	3	Credit line of the account in 1994 US dollars
creditHistory	4	Prior credit history
duration	5	The repayment duration of the loan in months.
employment	6	Current employment in years
existingCredits	7	The number of existing lines of credit including the current
foreignWorker	8	Status as domestic or foreign
housing	9	Status of housing
installmentCommitment	10	Installment rate in percentage of disposable income
job	11	Skill level of current job
lamount	12	The log credit line of the account in 1994 US dollars. Note that the first character is a small 'el' to denote log.
lamountPred	13	The predicted log credit line of the account in 1994 US dollars based on an internal statistical model. Note that the first character is a small 'el' to denote log.
numDependents	14	The number of dependent children of the debtor
otherParties	15	Co-applicant or Guarantor
otherPaymentPlans	16	Other existing forms of credit with financial institutions
ownTelephone	17	Has a working telephone number
personalStatus	18	Sex + Marital Status
propertyMagnitude	19	Relative size of the property ownership of the debtor
purpose	20	Purpose for opening line of credit
residenceSince	21	Current residence in number of years
riskStatus	22	The credit risk classification status
savingsStatus	23	Status of savings account and bonds

Disclaimer: the original datasets were modified and changed to suit the purposes of STAT 350. No serious conclusions should be drawn from the analysis.