Group 10 Experiment Procedure Presentation

Introduction to project

For our project we have chosen a dataset that includes a multitude of variables about a person including Age, Credit, Housing situation etc. We will use these variables as inputs to 3 different models, a one layer logistical regression model, a single hidden layer Neural Network, and a multi hidden layered Deep Neurological Network.

Problem statement

In today’s financial landscape, lending institutions must assess the risk of loan defaults with increasing precision. Traditional credit scoring methods often rely on manually crafted rules or linear models, which may fail to capture complex relationships between applicant attributes and loan outcomes. This project aims to develop a predictive model using deep learning techniques to automatically assess loan applications. By comparing the performance of logistic regression, shallow neural networks, and deep neural networks, we seek to determine the most effective approach for accurately classifying loan applications as accepted or rejected, ultimately aiding institutions in making faster and more reliable lending decisions.

RQs

1) How does the performance of a deep neural network compare to traditional machine learning models?

2) How does the performance of a deep neural network compare to shallow neural networks?

Data set description

Source & Metadata:

Dataset Name: German Credit Data

Source: UCI Machine Learning Repository

Author: Dr. Hans Hofmann

Size: 1,000 Instances, 21 Features

Target Variable: class — Binary outcome: *Good* vs. *Bad* credit risk

Rationale Behind Dataset Choice

✅ Real-World Relevance:

- Mirrors actual lending decisions

✅ Feature Variety:

- Mix of personal & financial info

✅ Binary Classification:

- Ideal for DNN with sigmoid output

✅ Cost-Sensitive:

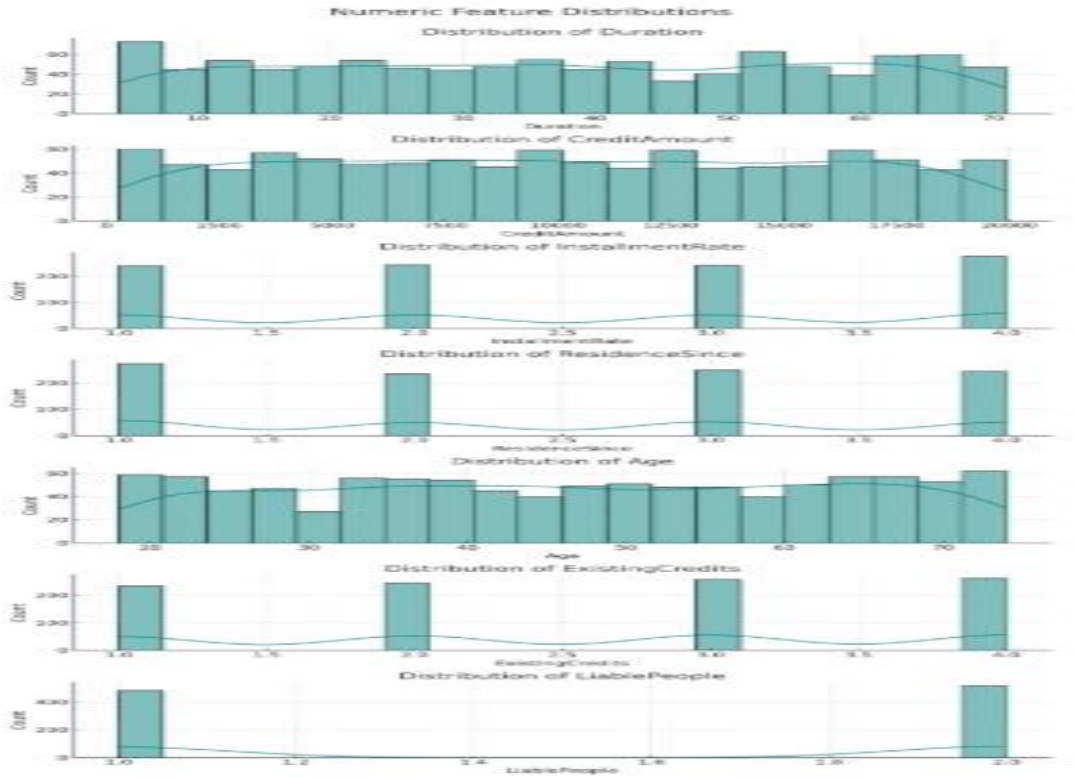
- Built-in cost matrix for evaluation

Dataset Feature Overview Table

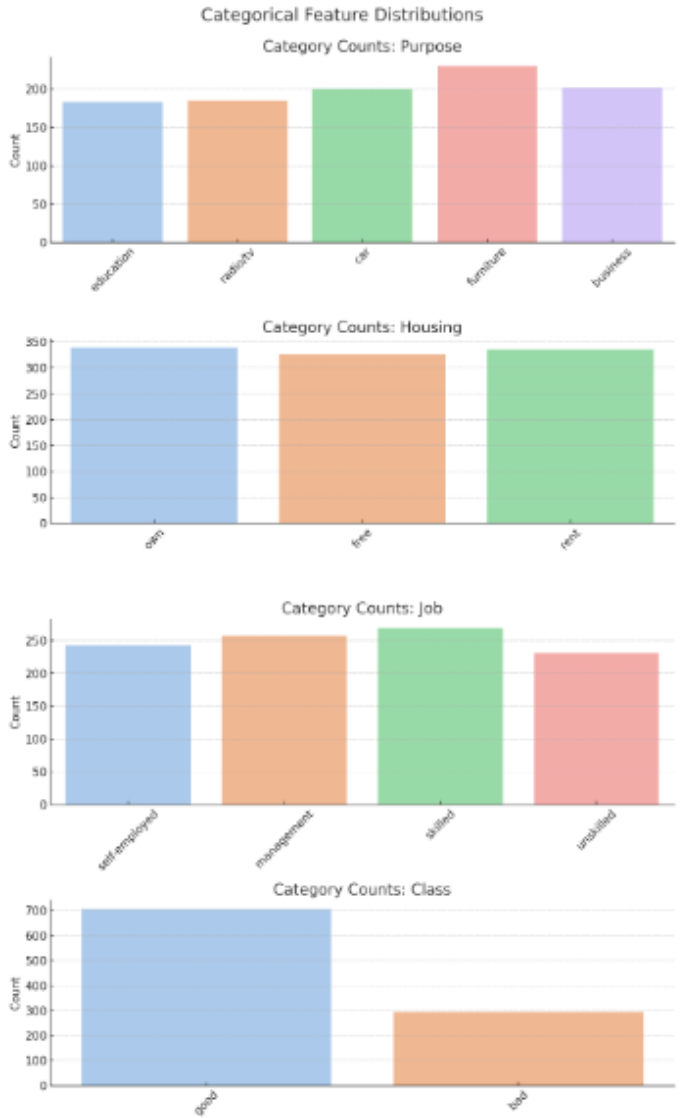
Feature Name	Type	Description
Status	Categorical	Status of existing checking account
Duration	Numeric	Duration in months
CreditHistory	Categorical	Credit history record
Purpose	Categorical	Purpose of the loan (car, furniture, etc.)
CreditAmount	Numeric	Amount of credit in Deutsche Marks
SavingsAccount	Categorical	Savings account/bonds status
EmploymentSince	Categorical	Present employment since (in years)
InstallmentRate	Numeric	Installment rate as a percentage of disposable income
PersonalStatusSex	Categorical	Personal status and sex (e.g., male single, female divorced)
Debtors	Categorical	Other debtors or guarantors
ResidenceSince	Numeric	Present residence since (in years)
Property	Categorical	Property ownership
Age	Numeric	Age in years
OtherInstallmentPlans	Categorical	Other installment plans (bank, stores, none)
Housing	Categorical	Housing status (own, rent, free)
ExistingCredits	Numeric	Number of existing credits at this bank
Job	Categorical	Job classification (skilled, unskilled, etc.)
LiablePeople	Numeric	Number of people being liable to provide maintenance
Telephone	Categorical	Has a telephone (yes/no)
ForeignWorker	Categorical	Is a foreign worker (yes/no)
Class	Categorical	Target: Good or Bad credit risk

Visualizations of Data Distributions

- Numeric feature distributions:



Bar Plots for **categorical feature counts**



Data pre-processing methods

- The dataset is split into train/test sets with a 80:20 ratio
  - Splitting is done with stratified sampling to ensure that each set has an equal ratio of label types
- New ratio columns are calculated to give the model more context
  - monthly_credit_burden = credit_amount / duration
  - installment_per_credit = installment_commitment / credit_amount
  - dependents_per_credit = num_dependents / credit_amount

Data pre-processing methods

- Each numerical column is replaced with its log and standardized
  - This is done to address the right skew of the data
  - log(0) is undefined, so those values are replaced with 0
  - duration, credit_amount, installment_commitment, residence_since, age, existing_credits, num_dependents, monthly_credit_burden, installment_per_credit, dependents_per_credit
- Categorical columns with ordered categories are converted into ordinal columns
  - Categorical columns with only 2 categories are also included
  - checking_status, savings_status, employment, own_telephone, foreign_worker
- Categorical columns with independent categories are one-hot encoded
  - This means that each category is turned into an extra binary column
  - The original categorical columns are deleted
  - job, housing, other_payment_plans, property_magnitude, other_parties, personal_status, credit_history, purpose

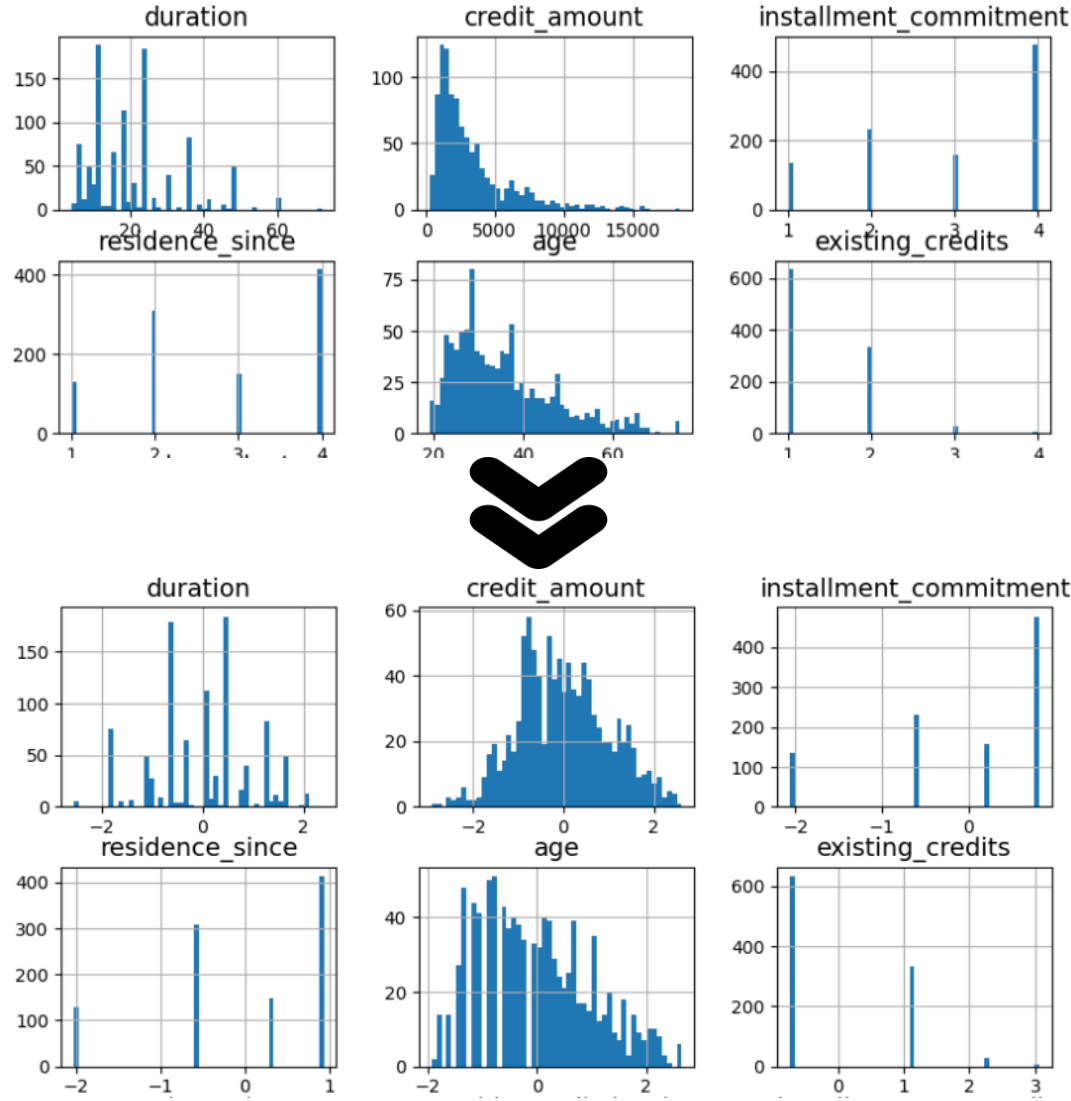
Data pre-processing methods

```
Original class distribution:
class
good    0.7
bad     0.3
Name: proportion, dtype: float64
class
good    700
bad     300
Name: count, dtype: int64

Training set class distribution:
class
good    0.7
bad     0.3
Name: proportion, dtype: float64
class
good    560
bad     240
Name: count, dtype: int64

Test set class distribution:
class
good    0.7
bad     0.3
Name: proportion, dtype: float64
class
good    140
bad      60
Name: count, dtype: int64
```

- After processing, the resulting dataset has 52 columns
  - Remember that the original has 21 columns
  - All the columns are now 64-bit floats, except for the labels
  - The numerical columns also no longer have a right skew



Data pre-processing methods

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 52 columns):
 #   Column                                        Non-Null Count  Dtype
---  ------                                        --------------  -----
 0   checking_status                               1000 non-null   float64
 1   duration                                      1000 non-null   float64
 2   credit_amount                                 1000 non-null   float64
 3   savings_status                                1000 non-null   float64
 4   employment                                    1000 non-null   float64
 5   installment_commitment                        1000 non-null   float64
 6   residence_since                               1000 non-null   float64
 7   age                                           1000 non-null   float64
 8   existing_credits                              1000 non-null   float64
 9   num_dependents                                1000 non-null   float64
 10  own_telephone                                 1000 non-null   float64
 11  foreign_worker                                1000 non-null   float64
 12  class                                         1000 non-null   object
 13  monthly_credit_burden                         1000 non-null   float64
 14  installment_per_credit                        1000 non-null   float64
 15  dependents_per_credit                         1000 non-null   float64
 16  job_high qualif/self emp/mgmt                 1000 non-null   float64
 17  job_skilled                                   1000 non-null   float64
 18  job_unemp/unskilled non res                   1000 non-null   float64
 19  job_unskilled resident                        1000 non-null   float64
 20  housing_for free                              1000 non-null   float64
 21  housing_own                                   1000 non-null   float64
 22  housing_rent                                  1000 non-null   float64
 23  other_payment_plans_bank                      1000 non-null   float64
 24  other_payment_plans_none                      1000 non-null   float64
 25  other_payment_plans_stores                    1000 non-null   float64
 26  property_magnitude_car                        1000 non-null   float64
 27  property_magnitude_life insurance             1000 non-null   float64
 28  property_magnitude_no known property          1000 non-null   float64
 29  property_magnitude_real estate                1000 non-null   float64
 30  other_parties_co applicant                    1000 non-null   float64
 31  other_parties_guarantor                       1000 non-null   float64
 32  other_parties_none                            1000 non-null   float64
 33  personal_status_female div/dep/mar            1000 non-null   float64
 34  personal_status_male div/sep                  1000 non-null   float64
 35  personal_status_male mar/wid                  1000 non-null   float64
 36  personal_status_male single                   1000 non-null   float64
 37  credit_history_all paid                       1000 non-null   float64
 38  credit_history_critical/other existing credit 1000 non-null   float64
 39  credit_history_delayed previously             1000 non-null   float64
 40  credit_history_existing paid                  1000 non-null   float64
 41  credit_history_no credits/all paid            1000 non-null   float64
 42  purpose_business                              1000 non-null   float64
 43  purpose_domestic appliance                    1000 non-null   float64
 44  purpose_education                             1000 non-null   float64
 45  purpose_furniture/equipment                   1000 non-null   float64
 46  purpose_new car                               1000 non-null   float64
 47  purpose_other                                 1000 non-null   float64
 48  purpose_radio/tv                              1000 non-null   float64
 49  purpose_repairs                               1000 non-null   float64
 50  purpose_retraining                            1000 non-null   float64
 51  purpose_used car                              1000 non-null   float64
```

Hardware/Software descriptions

- Software
  - Python notebook running on the cloud via Google Colab
  - TensorFlow (neural networks)
  - Pandas (dataframes)
  - Numpy (math)
  - Scikit-learn (machine learning models)
  - Matplotlib (graphing and charts)
- Hardware
  - Tesla T4 GPUs from Google Colab
  - 16GB GDDR6 memory
  - 2560 CUDA cores

Task assignments among team members

- **Member 1**: Dataset preprocessing (how the training data will be processed), description of hardware/software
- Member 2: Dataset description, why the dataset was chosen for the project and relation to research questions
- Member 3: Providing evaluation metrics, task assignments between all members, major project functionalities
- Member 4: Introduction, problem statement, and research questions (what is there to learn from this project)

3 Major functionalities of project

- DNN (Deep Neural Network)
  - At least 3 hidden layers
- Shallow Neural Network
  - 1 hidden layer
- Traditional Machine Learning Model
  - No hidden layers
- Determine the performance of each machine learning model on same dataset

Selected Evaluation Metrics

- How will we know our goals are met?

1. *How does the performance of a deep neural network compare to traditional machine learning models?*
2. *How does the performance of a deep neural network compare to shallow neural networks?*
3. *Determining if and how well a DNN can accurately predict loan approvals from applicant data.*

- Metrics
    - How easily it processes complex data
    - Time
    - Accuracy

End