Project 13: Alyssa Witt, Jumi Lee, Tianchen Guan, Emma Peterson, Yukai Sun

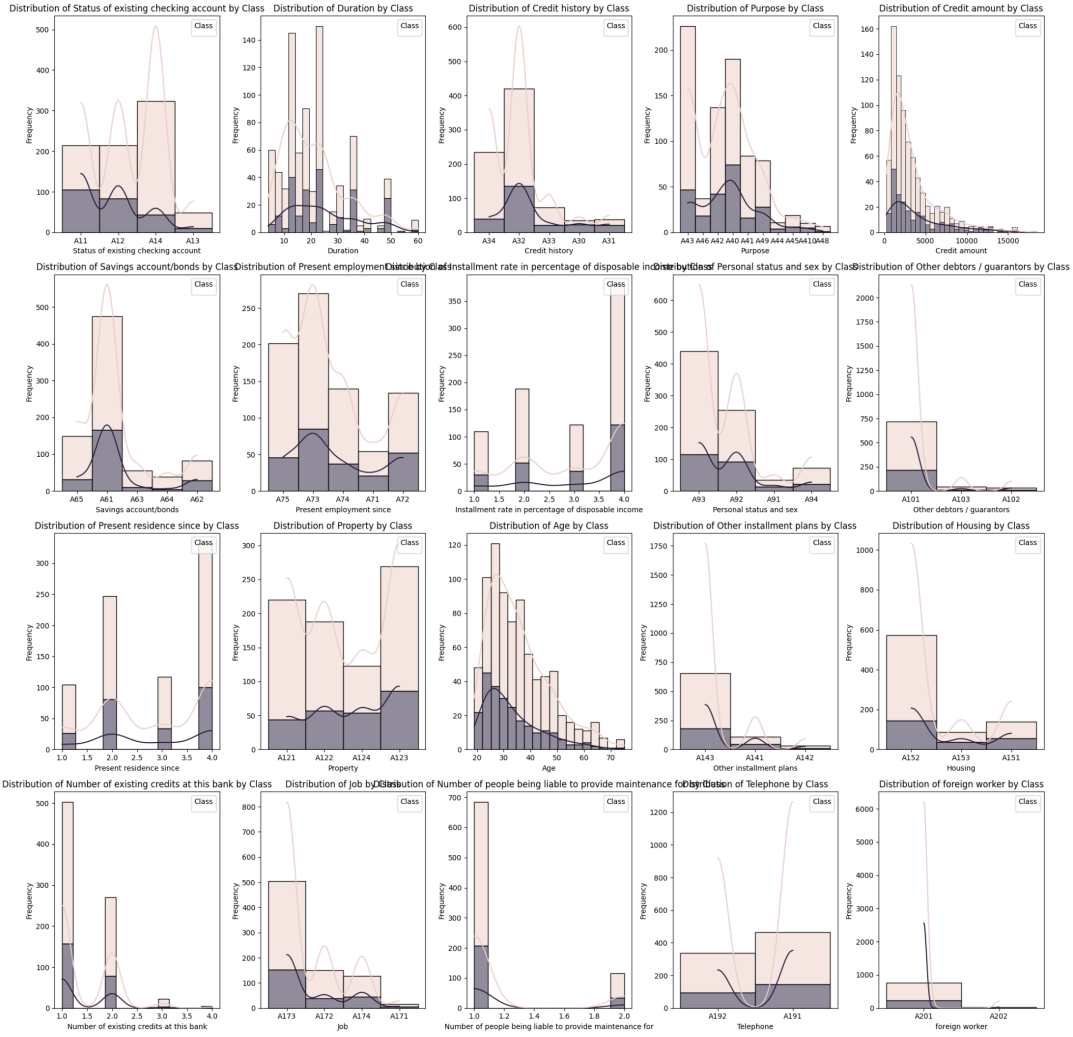
German Credit Data

Our dataset comprises 1,000 instances representing individuals, each characterized by 20 features determining credit risk quality, denoted as '1' for good and '2' for bad. Originating from Germany and collected pre-2002, it's denominated in Deutsch marks. We used a pre-processed version of Hans Hoffman's data, which revealed numerous NaN values. Following our feedback session, we found that the original data was fully encoded, with qualitative values represented by codes beginning with 'A'. This encoding ensured no NaN values were present in the dataset. Our primary question was;'Is it possible to categorize credit risk (good or bad) based on an individual's financial attributes?' we felt we could not accurately categorize credit risk with the size of the dataset used after working with logistic regression and decision trees.

description	Status of existing checking account	Duration	Credit history	Purpose	Credit amount	Savings account/ bonds	Present employment since	Installment rate in percentage of disposable income	Personal status and sex	Other debtors / guarantors	...	Property	Age	Other installment plans	Housing
0	A11	6	A34	A43	1169	A65	A75	4	A93	A101	...	A121	67	A143	A152
1	A12	48	A32	A43	5951	A61	A73	2	A92	A101	...	A121	22	A143	A152
2	A14	12	A34	A46	2096	A61	A74	2	A93	A101	...	A121	49	A143	A152
3	A11	42	A32	A42	7882	A61	A74	2	A93	A103	...	A122	45	A143	A153
4	A11	24	A33	A40	4870	A61	A73	3	A93	A101	...	A124	53	A143	A153

While completing data analysis, we checked for missing values (N/A) to ensure completeness. Then delved into the distribution of each feature concerning the target variable ('Class') by generating a grid of subplots. Each histogram displayed in these subplots represents the distribution of a specific feature in the dataset, with colors distinguishing between different classes.

Project 13: Alyssa Witt, Jumi Lee, Tianchen Guan, Emma Peterson, Yukai Sun

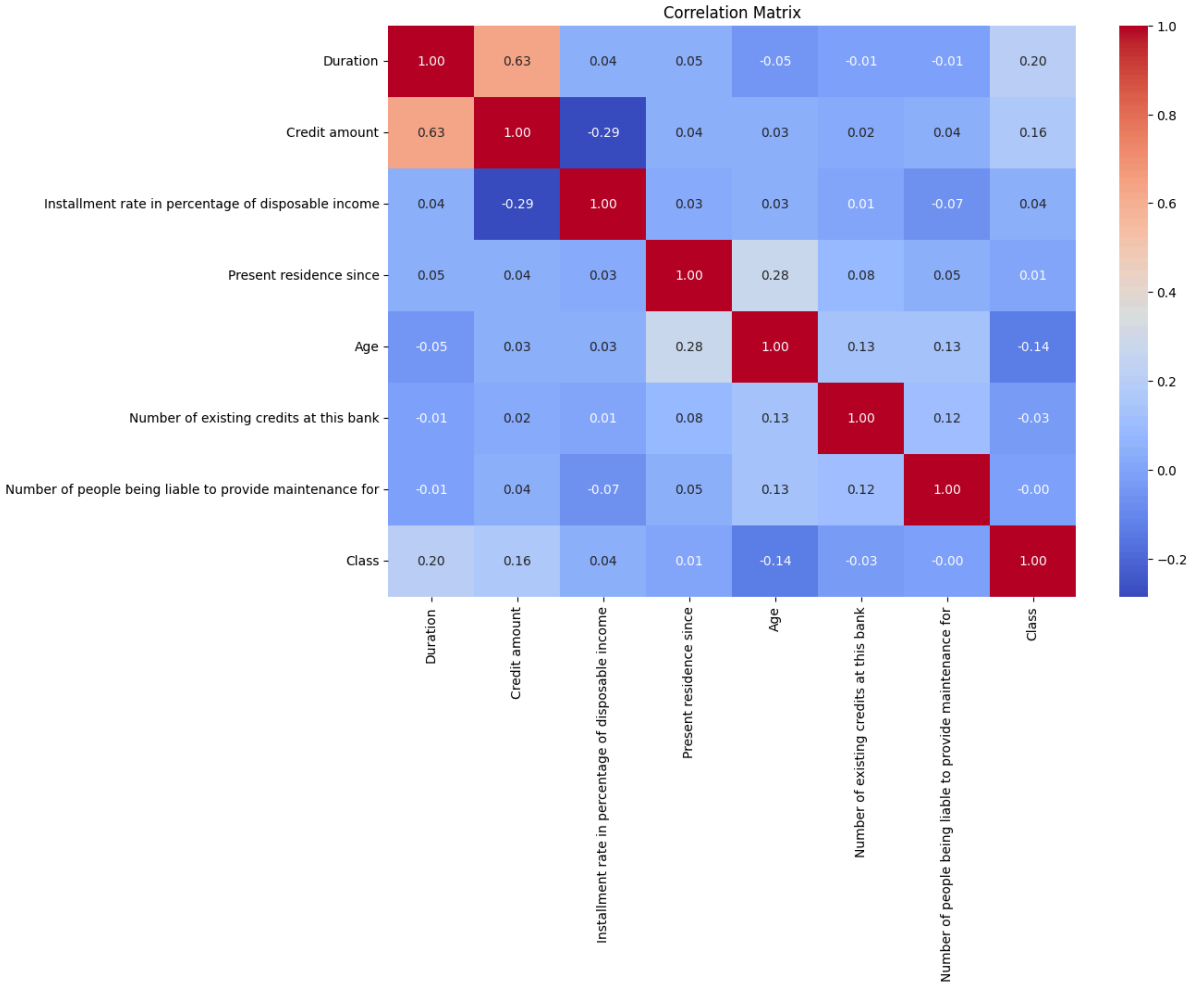


We calculated summary statistics for numerical and categorical features grouped by a target variable ('Class') to understand the dataset's characteristics before modeling.

Class	Duration mean	Credit amount mean	Installment rate in percentage of disposable income mean	Present residence since mean	Age mean	Number of existing credits at this bank mean	Number of people being liable to provide maintenance for mean
1	19.531306	3050.166369	2.932021	2.838998	36.599284	1.423971	1.144902
2	24.721992	4080.713693	3.041494	2.863071	33.145228	1.381743	1.141079

We computed the correlation matrix between numeric features and the target variable (‘Class’). We then visualized the correlation matrix using a heatmap to identify strong correlations.

Project 13: Alyssa Witt, Jumi Lee, Tianchen Guan, Emma Peterson, Yukai Sun



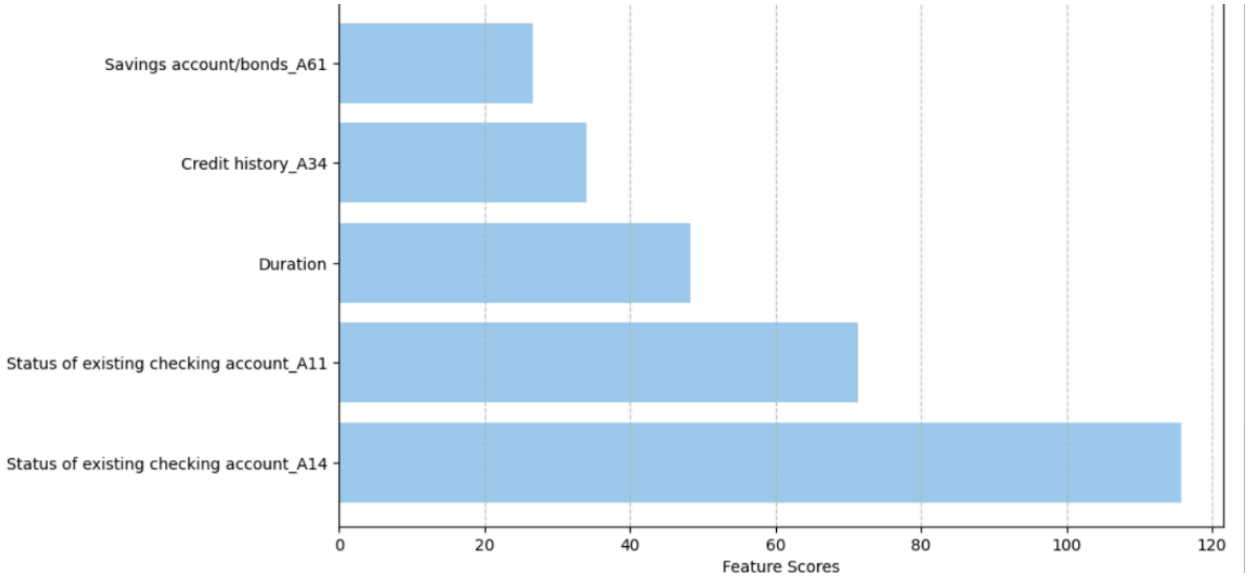
Before using one-hot encoding on our entire dataset, we manually assigned male/female within the sex column. Otherwise, the column would have been incorrectly encoded, with one code for males and four for females. The correct amount was 3 for males and 2 for females.

During model selection for the dataset, we evaluated these classifiers: Support Vector Machine, Logistic Regression, Decision Tree, Random Forest, and K-Nearest Neighbors. Each was subjected to a hyperparameter tuning process using GridSearchCV, which iterates through predefined hyperparameter grids for each model, assesses their performance via cross-validation, and determines optimal settings.

SVM was tested with different regularization parameter values and kernel types, Logistic Regression with variations in C and penalty types (l1 and l2), Decision Tree with different max_depth values, and KNN with varying k values. Each classifier's configuration aimed to balance bias and variance for optimal generalization, assessed against a validation set. These experiments provided a comprehensive comparative analysis, revealing the strengths and weaknesses of each model in credit risk assessment.

Decision Trees achieved a validation score of 0.75, while Logistic Regression attained a score of 0.76 and 73% accuracy on the test data. These models were selected for their distinct advantages and high accuracy. By concentrating on these models, we could intricately optimize their parameters and comprehend their operational dynamics within credit risk evaluation. This ensures that the chosen models perform optimally based on historical data and provide reliable and interpretable predictions for practical use.

For feature selection, we utilized the SelectKBest function from scikit-learn, employing the f_classif scoring function suitable for categorical target variables. Through cross-validation tests, we determined k=5 as the optimal number of features, improving baseline model accuracy from 65% to 75%. These features helped answer these research questions; ‘Which factors play a decisive role in predicting credit risk?’, ‘How do gender, age, and occupation affect an individual's credit rating?’ and ‘How do housing status and savings account balance affect credit risk assessment?’. Notably, gender, age, occupation, and housing status were not selected as significant predictors.



SelectKBest Results:

Selected Features & Their F-value scores

```
[[129  12]
 [ 32  27]]
              precision    recall  f1-score   support

           1       0.80      0.91      0.85       141
           2       0.69      0.46      0.55        59

    accuracy                           0.78       200
   macro avg       0.75      0.69      0.70       200
weighted avg       0.77      0.78      0.76       200
```

Logistic regression results

```
[[124  17]
 [ 34  25]]
              precision    recall  f1-score   support

           1       0.78      0.88      0.83       141
           2       0.60      0.42      0.50        59

    accuracy                           0.74       200
   macro avg       0.69      0.65      0.66       200
weighted avg       0.73      0.74      0.73       200
```

Decision tree results

Finally, accuracy for logistic regression was 0.78 and the Decision tree was 0.74. Both models show strong performance in identifying class 1 with high recall and reasonably good precision. They struggle with identifying class 2, where both precision and recall are significantly lower, suggesting a large number of false negatives when predicting class 2. This may indicate that the model is biased towards class 1, which also has a larger support (141 instances) compared to class 2 (59 instances). We assert here that doing sample balancing or increasing the sample size will help improve class 2 predictions. Without that, we didn't feel we could confidently say yes to our final research question 'Is it possible to categorize credit risk (good or bad) based on an individual's financial attributes?'.

Project 13: Alyssa Witt, Jumi Lee, Tianchen Guan, Emma Peterson, Yukai Sun

[Contributions]

Member	Proposal	Coding	Presentation	Report
Tianchen Guan	1	1	1	1
Jumi Lee	1	1	1	1
Emma Peterson	1	1	1	1
Yukai Sun	1	1	1	1
Alyssa Witt	1	1	1	1

- In the chart above, 1 = full contribution, 0.1-0.9 = partial contribution, 0 = no contribution.
- Everyone put in work together as a team and contributed to every aspect of the project.

Sources:

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