Team: amabri

Predict Creditworthiness with Alternative Data

Yuanshan Zhang, Honglin Jiang, Ruiqi Jiang, Mengxin Zhao

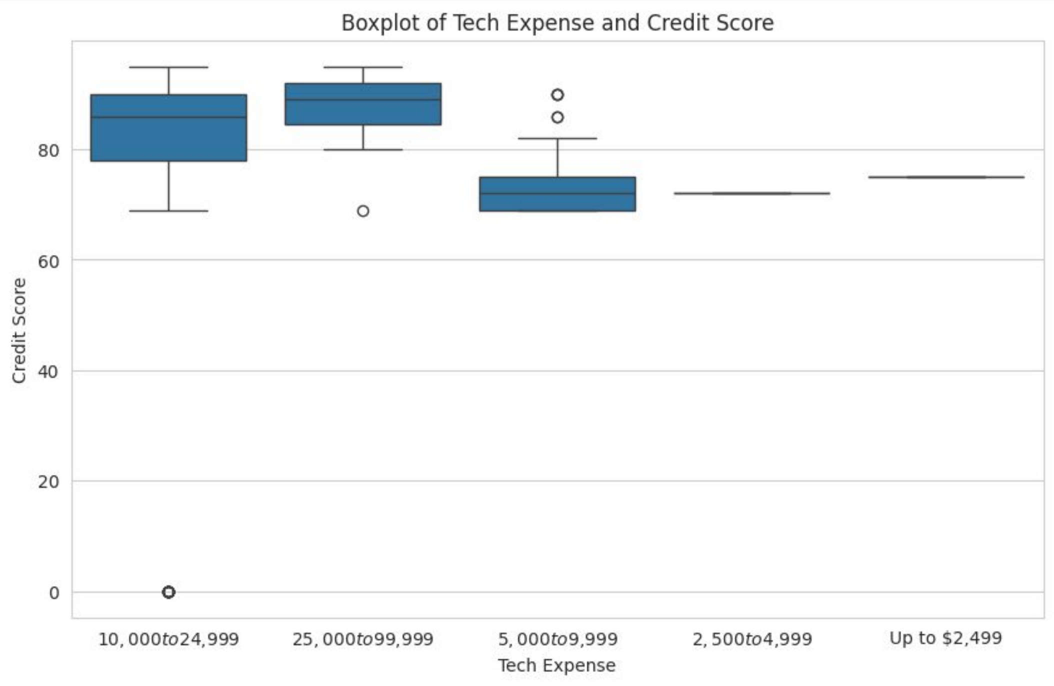
Executive Summary

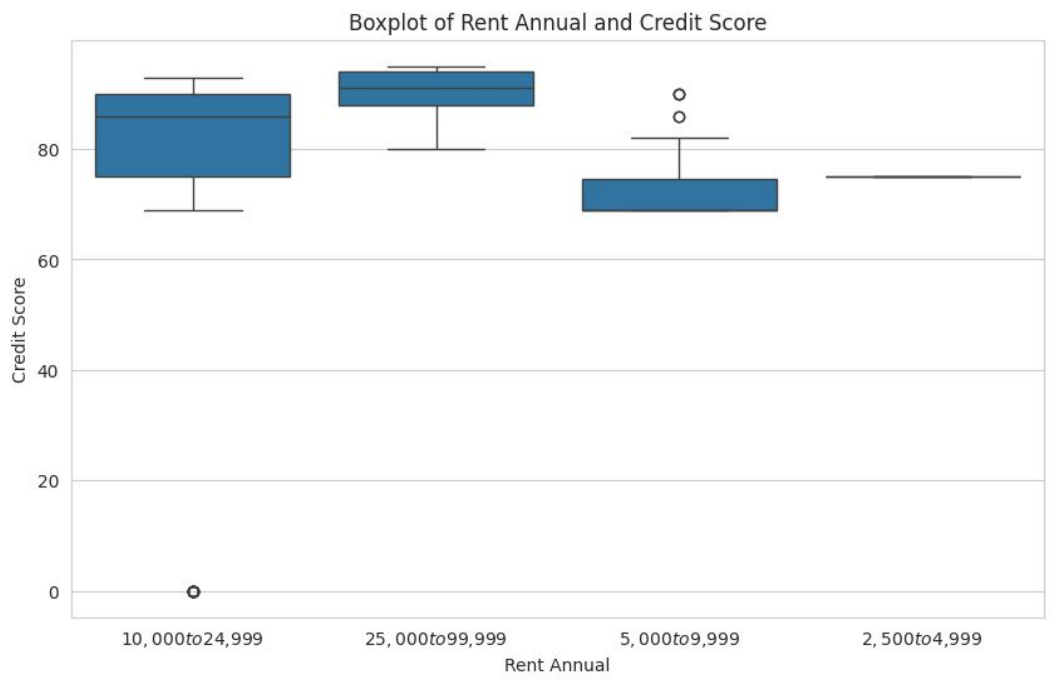
Our model uses financial data (revenue, sales, expenses, employee growth) over 5 years and business characteristics (ownership, location) from 225 SMEs to predict loan repayment capability. This helps banks assess SMEs lacking traditional credit scores. We achieved a highly accurate model (test MAE 2.25, 98% fit) through rigorous feature selection. This promotes financial inclusion by enabling better loan decisions for underserved businesses. Future work will explore additional data sources for further improvement.

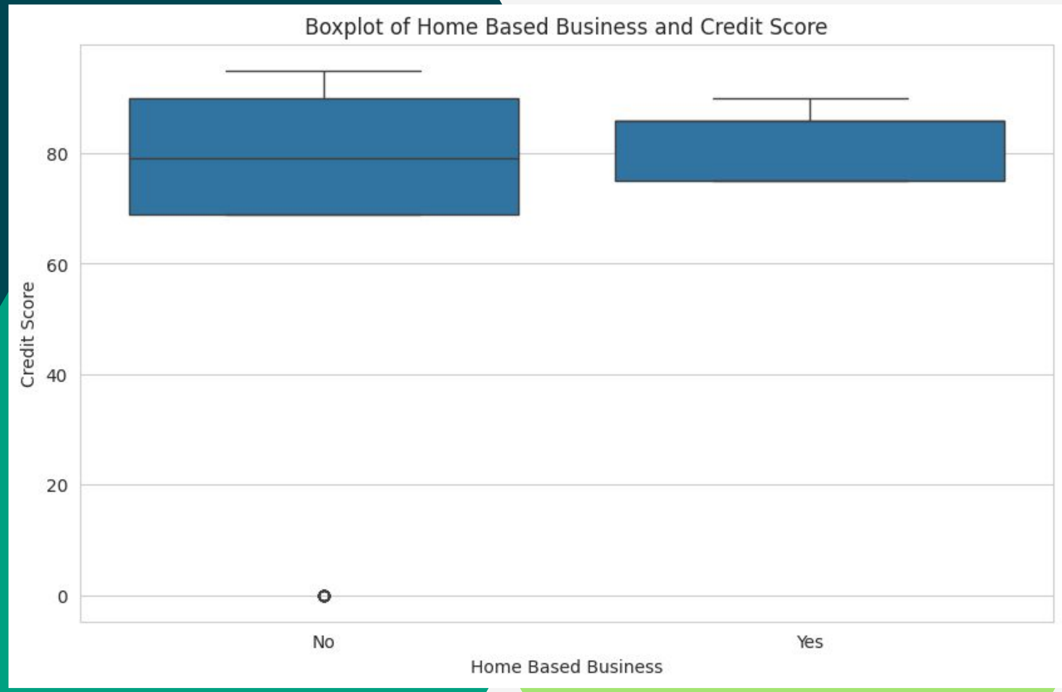
Data

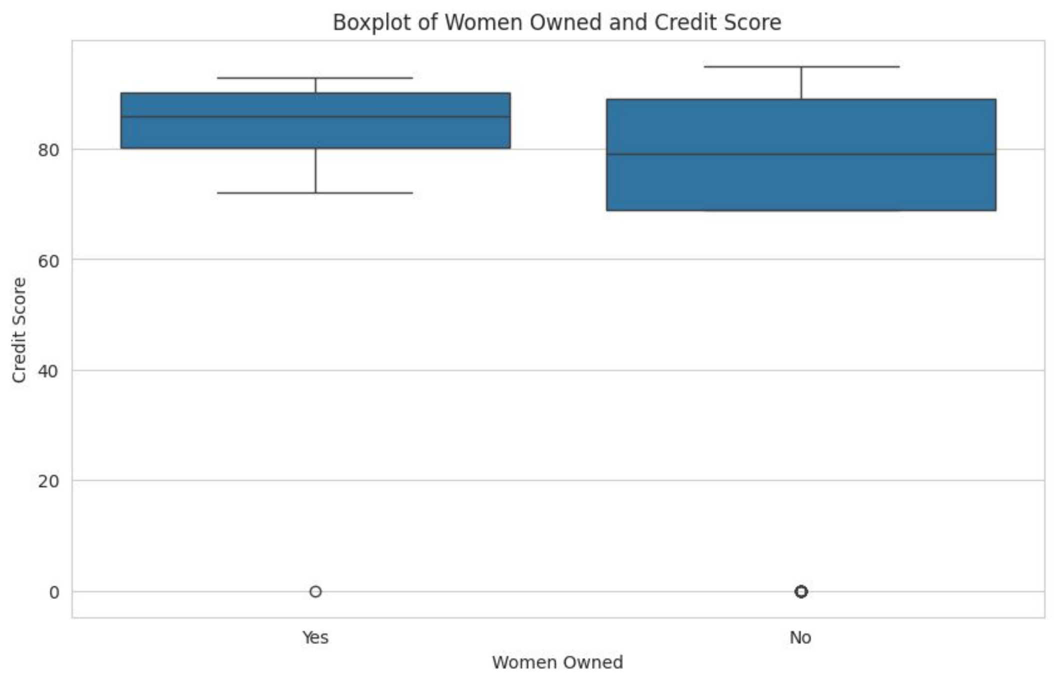
Our model leverages a dataset of 225 small businesses, capturing financial data like revenue, sales, expenses, and employee growth over a 5-year period. Additionally, we included business characteristics such as ownership structure (women-owned, public vs. private) and location (home-based) to investigate potential correlations with creditworthiness and identify any bias towards underserved demographics.

Data









To build the most accurate model, we evaluated various factors including technology expense, annual rent, and business characteristics like being women-owned or home-based. However, through our feature selection process, we determined these factors did not have a statistically significant impact on predicting creditworthiness and were therefore excluded from the final model.

Other Data Considered

Google News Headline: we considered conducting sentiment analysis on news on small businesses. However, due to the limitation of NEWS API, we are restrained in retrieving within 50 data points.

Yelp review sentiment: we considered conducting sentiment analysis on yelp review. However, Yelp doesn’t offer API and the dataset that yelp offers are

Better Business Bureau rating

Industry GDP

Google search trend

Goals and Strategy

- Goal: Our aim is to develop a classification model by integrating diverse datasets from various sources. This model will generate a variable that serves as an indicator of repayment capability for small and medium-sized enterprises (SMEs) lacking a traditional credit score. Banks will utilize this variable to assess loan eligibility for these enterprises.
- Strategy:
  - Data Integration: Gather data from multiple sources including financial records, transaction history, and alternative credit data sources.
  - Feature Engineering: Identify and extract relevant features from the integrated datasets to build a comprehensive set of variables.
  - Model Development: Utilize Supervised machine learning techniques–classification algorithms to develop a predictive model that can classify SMEs based on their repayment capability.
  - Validation and Optimization: Validate the model's performance using historical data and fine-tune it to enhance accuracy and reliability.

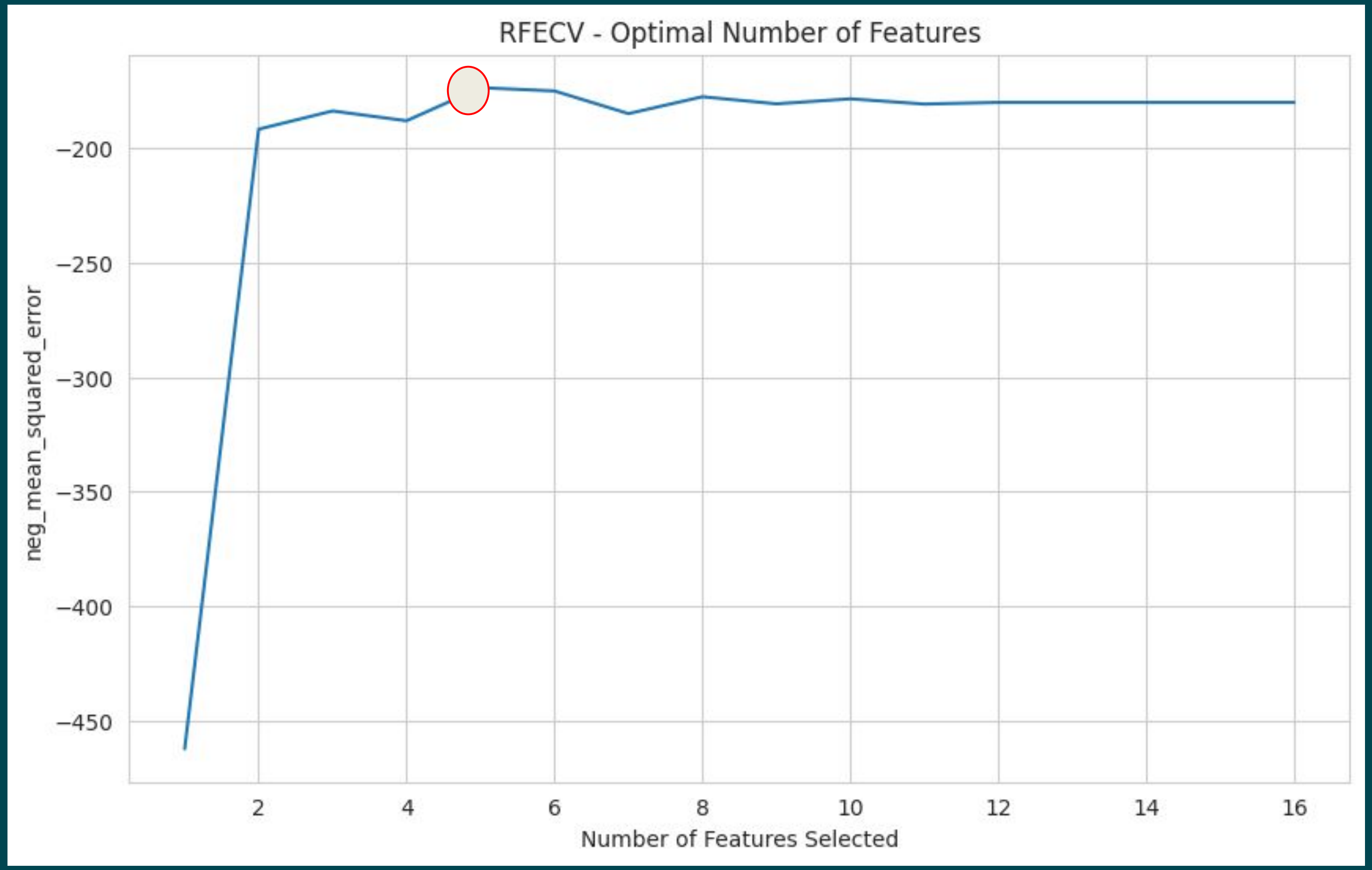
Model

Data Processing Pipeline

Classification Model: Random Forest

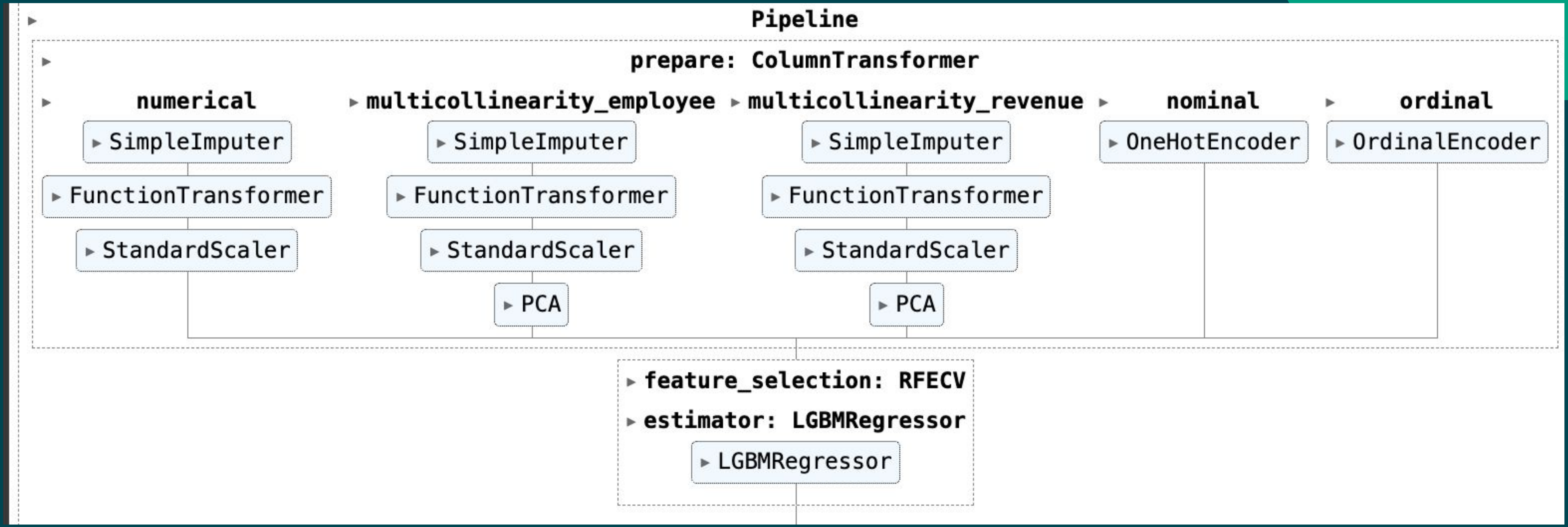
Preprocessing Pipeline

Feature Selection



Model

Pipeline



Model

Classification Model

Classification using Random Forest Classifier → Fine Tuning Using RandomizedSearchCV → Evaluating Performance on Test

Result

- Train Score:
  - MSE: 13.214
  - MAE: 1.174
  - R^2: 0.978
  - RMSE: 3.635
- Cross-Validation Score: R^2: 0.851 ± 0.173
- Best parameters:
  - min_samples_leaf: 5
  - max_depth: 10
- Best R^2 for Random Search is 0.874
- Test Score:
  - MSE: 10.106
  - MAE: 2.252
  - R^2: 0.982
  - RMSE: 3.179

Challenge

Data Collection

1. Hard to capture suitable data
    - Failed web-scraping attempts using APIs
2. Data usability is low
    - Traditional data
    - Hard to be integrated with the model

Data Understanding

1. Confusion over if credit scores can be used as the target variable
2. If credit scores are not desirable, what can we use?

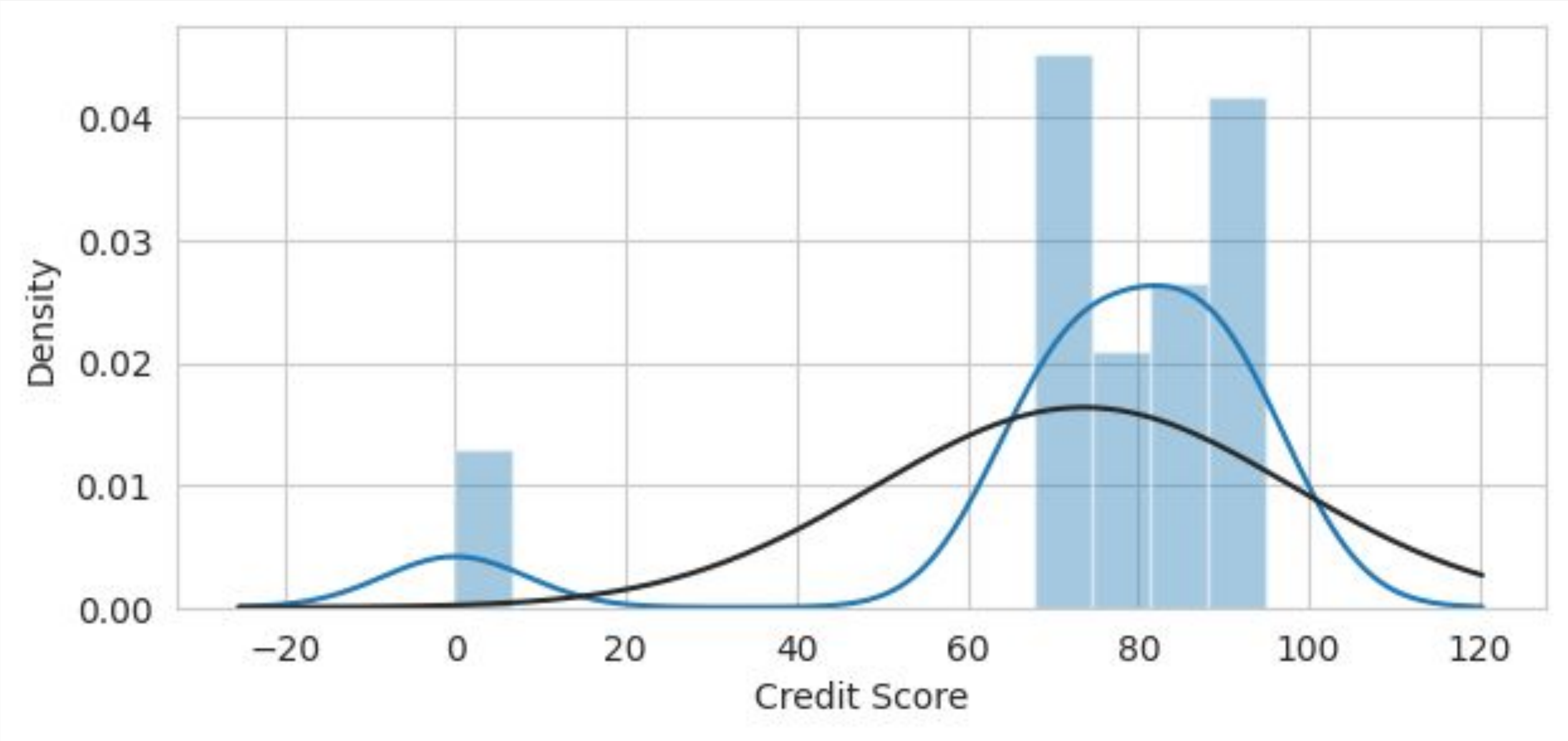
Resources

- Mainly used Dataset: https://www-atozdatabases-com.ezproxy.bpl.org/search

- Data sources intended to use:
    - https://trends.google.com/trends/
    - https://www.bbb.org/overview-of-bbb-ratings
    - https://news.google.com/home?gl=US&hl=en-US&ceid=US:en

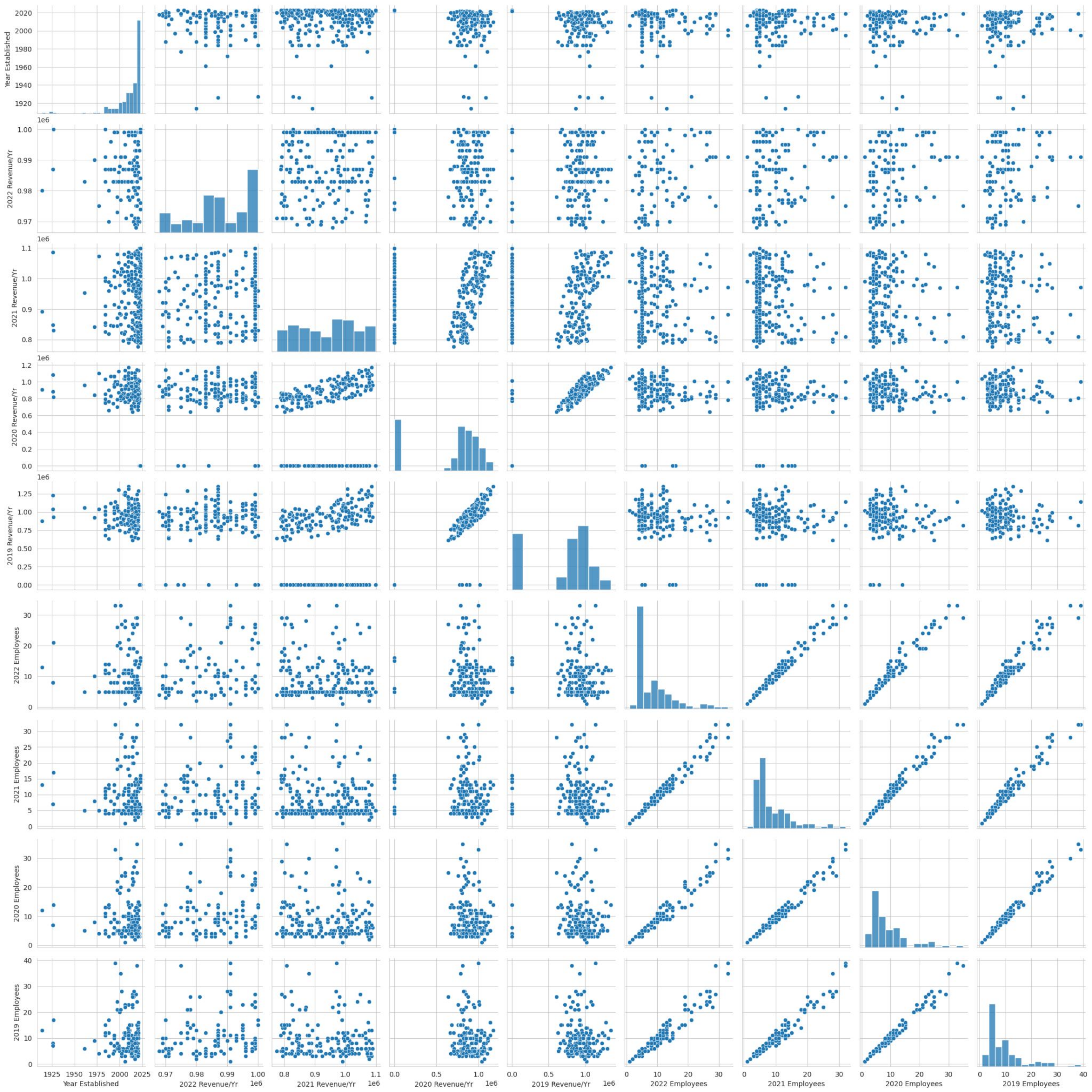
Appendix

Distribution of credit scores



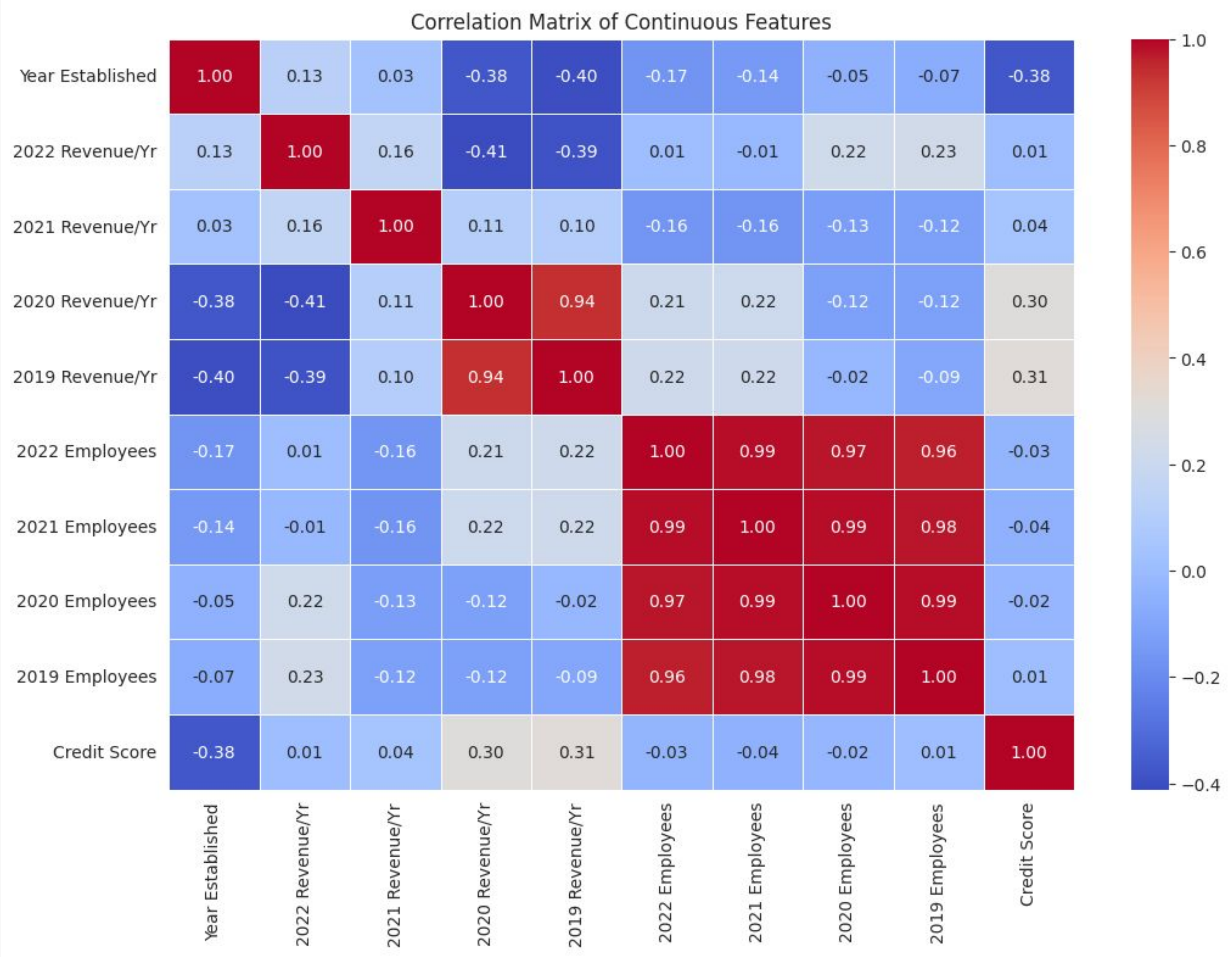
Appendix

Pairplot for all numerical values



Appendix

Heatmap for all numerical values



Appendix

Visualizing coefficients of most important features