CAPITAL TWO BANK PROPOSAL - ENHANCING CREDIT CARD ASSESSMENT: UNSUPERVISED LEARNING INTEGRATION INTO SUPERVISED MODELS

Ureem James
MSDS 411, Winter 2024, Section 55
Northwestern University, Unsupervised Learning
February 14th, 2024

Abstract

Capital Two Bank has issued a request for a proposal to be submitted on approaches to improve the model performance of its current logistic regression approach and simultaneously reduce modeling efforts. This paper recommends that Capital Two implement a k-means clustering algorithm with five clusters prior to logistic regression modeling and only utilize the labels generated from the clustering as the predictors. The recommended approach leads to an increase in average precision (+32.6%), recall (+7.7%), the consequent f1-score (+29.1%) and a decrease in misclassification costs (-48%); this approach not only increased model performance, but also removed the need for Capital Two to model all explanatory variables as features whilst logistic regression modeling.

Introduction

Capital Two Bank, a mid-Atlantic bank that provides premier credit card services, has issued a request for proposal (RFP) seeking submissions on how it can best improve scoring metrics for its current logistic regression modeling approach. The goal is to predict whether a credit card applicant is a good or bad applicant while simultaneously reducing the time and resources spent on modeling efforts.

In the RFP, the Capital Two has highlighted that given the high inflationary and rate environment of 2024, it is highly concerned about credit card defaults increasing and its impact on revenue, profitability and provisional reserves. In a cost analysis, Capital Two has estimated that the cost of approving a bad applicant is 5 times the cost of approving a good applicant, and has therefore estimated its internal default probability threshold to 8.6% to reduce costs of misclassification (refer to the **Methods** section).

The purpose of this proposal is to explore the pairing of an unsupervised learning technique along with Capital Two's current supervised learning approach to increase model performance and reduce modeling efforts.

Literature Review

The literature review seeks to provide an overview of the diverse unsupervised pre-training methodologies employed to improve the performance of supervised learning within the domain of credit card operations and other applications.

Credit Cards

In the paper, *Combining unsupervised and supervised learning in credit card fraud detection*, researcher Fabrizo Carcillo surveyed various unsupervised techniques integrated with supervised credit card fraud detection classifiers. Carcillo and his colleagues picked the Random Forest model as their baseline supervised learning technique of choice, and surveyed a variety of unsupervised approaches to pair - global, local and clustering.

In the global approach, the researchers assumed that all credit card transactions "..are considered to be samples of a unique global distribution for which outlier scores can be computed. A transaction is considered anomalous if it lies outside the overall multivariate pattern of the entire set of transactions" (Carcillo et al. 2021). In contrast, the local approach focuses on each card, and a transaction is considered anomalous to the card if it differs from past transactions carried out by the card.

K-means clustering was the clustering algorithm of choice as the researchers found it easy to interpret, efficient on large datasets, and allows for the presetting of clusters (Carcillo et al. 2021). The aim of the clustering algorithm was to segment the data and label the customers based on their average spend, which would ultimately be used as a benchmark to detect anomalous spend.

As a challenge to clustering, the researchers specifically outlined the challenge of deciding on which features to cluster on - features referring to cardholder behavior or personal data. In addition to data challenges, researchers also highlighted the inherent problem with using k-means clustering - setting the number of clusters. Consequently, researchers tested a variety of cluster settings with clusters ranging from 10 to 5000.

To measure the performance of the various algorithms, Area under the Curve (AUC), Precision and Recall were utilized. Ultimately, only the clustering algorithm showed promise by outerperforming baseline expectations on anomaly detection, but researchers noted that, in practice, a clustering algorithm would be difficult to sustain given that the hyperparameters could be sensitive to new granular data in addition to the fact that inactive cards could pose a challenge with fewer transactions.

Other Domains

Stepping outside the domain of credit card applications, in the paper *Combining Supervised and Unsupervised Learning Algorithms for Human Activity Recognition*, researchers explored the implementation of clustering data prior to utilizing graph convolutional neural networks to detect human activity (Budisteanu and Mocanu 2021). The primary reason to detect human activity was to assist in decreasing oversight required at elderly care centers and help develop detection software for robots utilized in the assisted living space.

The data utilized for the study was skeleton data and sequences of movement and activities, and the consequential aim of the clustering approach was to partition the data based on similar activities. Researchers noted that there were benefits to this approach; namely that by using clustering, researchers didn’t have to retrain the neural network model as a result of unseen data saving time and resources.

K-means and Gaussian Mixture Models (GMM) were selected as the two primary clustering methods of choice to partition and encode the data by attaching hyper-labels based on the various number of activities. The researchers noted that the advantage of using a Gaussian Mixture Model is that it doesn't use a distance similarity measure, instead it aims to optimize the likelihood of the data points belonging to certain Gaussian distributions; this helps consider both the mean and variance of the data (Budisteanu and Mocanu 2021).

To partition the data, the researchers tested three different values for the number of clusters - 5, 9 and 15 - and utilized t-SNE to visualize the results. The results showed that GMM with 5 clusters was most representative of the data. Following up on that, researchers utilized the encoded data as inputs for the GCN and found that the GMM model with 5 clusters provided any accuracy of 78.33% - a 9% improvement from the baseline GCN model without prior encoding. That said, even though GMM with 5 clusters was best representative of the data, the researchers still tested various configurations of k-means clustering and found that the best k-means encoder utilized 5 clusters and provided an accuracy of 74% - a slight decrease in contrast to the GMM approach, but still an improvement on the baseline un-encoded GCN approach.

Based on the literature review and the different applications of encoders utilized to improve supervised learning models, there are **takeaways** to consider:

1) K-means clustering is a popular encoding approach as it was applied for both credit card and non-credit card applications.
2) Both papers noted the benefits of encoding prior to supervised classification - saving of time and resources.
3) Both baseline supervised learning models and clustering approaches combined with supervised learning models must be performed on the same features.

Data

Proprietary data from Capital Two Bank was not provided for the proposal; instead public data regarding German credit cards applicants was provided, which was sourced by Capital Two from OpenML (Hoffman 1994). The raw data provided contained 20 explanatory features with 1000 observations with each observation corresponding to a credit card applicant being deemed good or bad - captured in the separate ‘class’ feature.

In addition to data provisions, the modeling team at Capital Two provided the following guidelines on specific data transformations that were applied in the development of the baseline logistic regression model:

Capital Two Feature Transformations

Feature	Observation	Transformation
Personal Status	● The level “female single" had 0 observational counts.	● “Female single” removed and releveled to a four levels: ○ Male Div/Sep ○ Female Div/Dep/Sep ○ Male Single ○ Male Mar/WID
Purpose	● 11 level categorical variable, but contained low and zero observational counts: ○ Domestic Appliance (12) ○ Repairs (22) ○ Vacation (0) ○ Retraining (9) ○ Other (12)	● The Purpose feature was releveled to a 7 level categorical variable: ○ ‘Retraining’ was relabeled to ‘Education’ (an existing level) ○ ‘Domestic Appliance’, ‘Repairs’ and ‘Vacation’ were relabeled to ‘Other’ (an existing level)
Credit Amount	● Credit Amount was right skewed.	● Natural Log applied to correct for skewness.

Foreign Worker	● Only 3.7% of observations are foreign workers.	● Feature was removed since very few were foreign workers and it was not included in the baseline logistic regression model.

Considering that any alternative encoder model will be compared against the performance of the baseline logistic regression model, the above transformations were applied to the dataset prior to any transformations applied independent of the modeling team at Capital Two.

After applying Capital Two's transformations, the remaining features were explored further and it was observed that 12 of the 19 explanatory features were categorical variables, out of which only the 'own_telephone' feature - indicating whether the applicant owned a telephone or not - was the only binary categorical variable; the rest of the categorical features contained 3 or more levels corresponding to both ordinal and nominal categorical variables (refer to Appendix B for breakdown). The remaining six features were numerical and upon further exploratory data analysis it was observed:

1) The correlation heatmap for numerical features showed strong correlation between Credit Amount and Duration implying that the longer the applicant has had active credit the more credit amount (Appendix A - Fig 1).
2) The median Credit Amount was higher for applicants deemed bad versus applicants deemed good, whereas the median Age was lower for applicants deemed bad versus applicants deemed good (Appendix A - Fig 2 & Fig 3)
3) The Age feature is right-skewed per the histogram and contains potential outliers per the boxplot (Appendix A - Fig 4 & Fig 2).

Given that clustering algorithms - particularly k-means clustering - is better suited for numerical data and that too without skewness, the following feature transformations were performed on both the categorical and numerical data (Kumar et al. 2015):

Feature Transformations (not applied by Capital Two)

Feature	Transformation	Reason
Nominal Categorical Variables (Appendix B)	Encoded as indicator features with binary values (0/1) with 1 indicating an occurrence.	The levels in nominal categorical variables don't have intrinsic order (Gowtham 2022).
Ordinal Categorical Variables	Encoded as integer labels. (Appendix C - Label Mapping)	The levels do imply intrinsic order. For instance, the "checking status" variable has levels that escalate from individuals with no checking accounts to individuals greater than 200 D-Mark in their accounts.
Age	Natural Log	Skewness; K-means is highly impacted by skewness. (Kumar et al. 2015).

After applying the above transformations, the final data set retained included 1000 observations and 45 explanatory features.Lastly, given that clustering is not scale invariant, and can be sensitive to choices of features of varying units of measure, the **data set was scaled with MinMax Scalar**, which scales the features to a range between 0 and 1 by subtracting each observation from the minimum value of the feature, and dividing the result by the difference between the maximum value of the feature and the minimum value of the feature.

The choice for the MixMax Scalar was driven by the fact that only two of the features are normally distributed and given that the data contains ordinal categorical labels that have now been integer encoded, the usage of the minmax scalar will help retain the relative relationship that has been encoded (Kumar 2023).

Methods

Based on the findings from the literature review and the size of the current dataset, **k-means clustering will be an appropriate choice** to encode the dataset and generate labels prior to the logistic regression modeling.

Capital Two's current approach does not take any prior unsupervised learning approach into consideration, and involves logistic regression modeling with the probability of class - whether an applicant is deemed good or not with 1 referring to an applicant deemed bad - being predicted by each of the explanatory variables excluding Foreign Workers.

Capital Two utilized 5-folds cross validation to allow every observation in the dataset to be used for training to increase the model's generalizability. For the purposes of comparability, the combined k-means and logistic regression model will also utilize 5-folds cross validation. Therefore, all scoring metrics related to logistic regression will utilize average scores across the 5 folds. In addition to similar metrics, Capital Two in its current approach is modeling class against all the explanatory variables. Given that a reduction in modeling efforts is a primary objective of Capital Two, an approach **where 'class' is modeled solely against the k-means labels will also be explored**.

K-Means Scoring Metrics

When assessing the fit of the k-mean clustering and the optimal number of clusters, a **scree plot** will be used to validate the initial choice of clusters, contrasting the increase in the number of clusters versus decrease in the **inertia** and increase in the **silhouette score**; the former measuring the distance between the observations and the centroids in each cluster (lower inertia being ideal) and the latter measuring the separation between clusters (higher silhouette score being ideal). A cut off point on the number of clusters will be decided based on the number of clusters that provide decrease in inertia and increase in silhouette scores.

Additionally to visualize the clustering, a **T-SNE plot** will be utilized to interpret the latent descriptions of the clustering groups on a two dimensional space. The descriptions will assist in understanding the similarities and differences between credit card applicants.

Logistic Regression Scoring Metrics

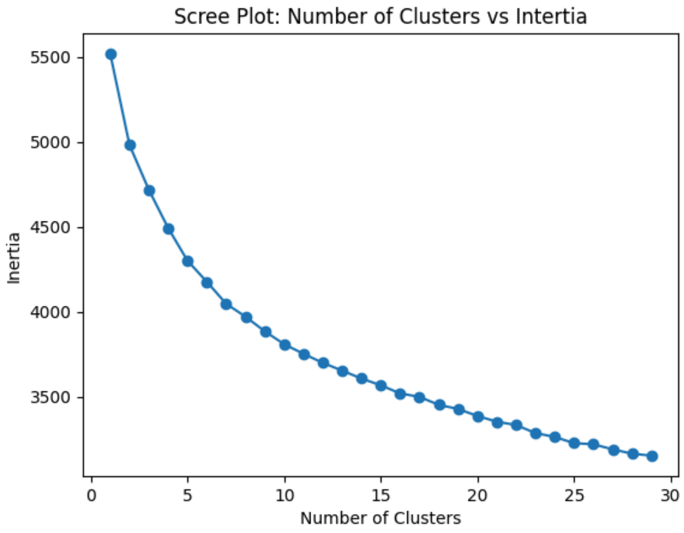
In line with Capital Two's scoring metrics, the combined k-means encoder with logistic regression model will be evaluated on the following classification scoring metrics:

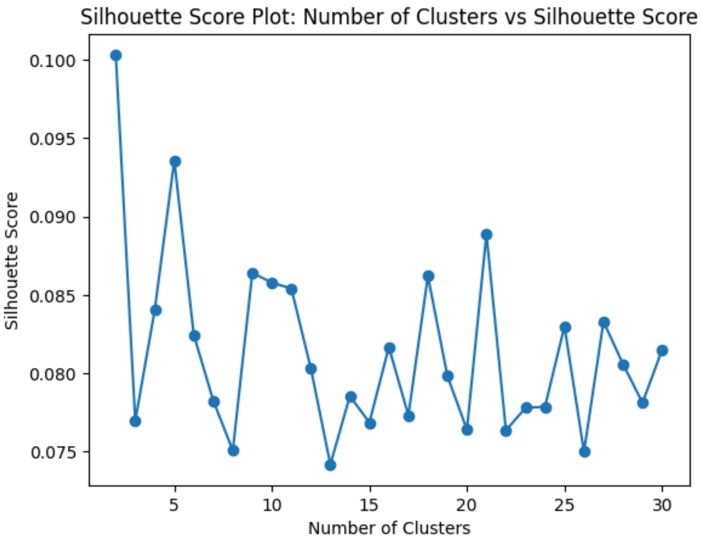
1) **Precision** - measuring the ratio of correctly predicted positive observations of the total predicted positives.
2) **Recall** - measuring the ratio of the correctly predicted positive observation to all the actual positives.
3) **F1-Score** - the harmonic mean of precision and recall ranging between 0 and 1, where 1 indicates perfect precision and recall.

Within the context of the data provided, **the positives represent that an applicant is deemed bad**. Given that logistic regression aims to predict probabilities, a cut-off threshold is required for the probabilities. Capital Two has relayed that a bad applicant being approved is five times as costly as a good applicant being approved. Based on this information and the distribution of the class variable, (700 good applicants vs 300 bad applicants) the cut-off threshold is 0.086 implying that any applicant with the probability that exceeds the threshold will be deemed a bad applicant. As a result, the **fourth scoring metric** utilized will be the **average cost of misclassification across folds**.

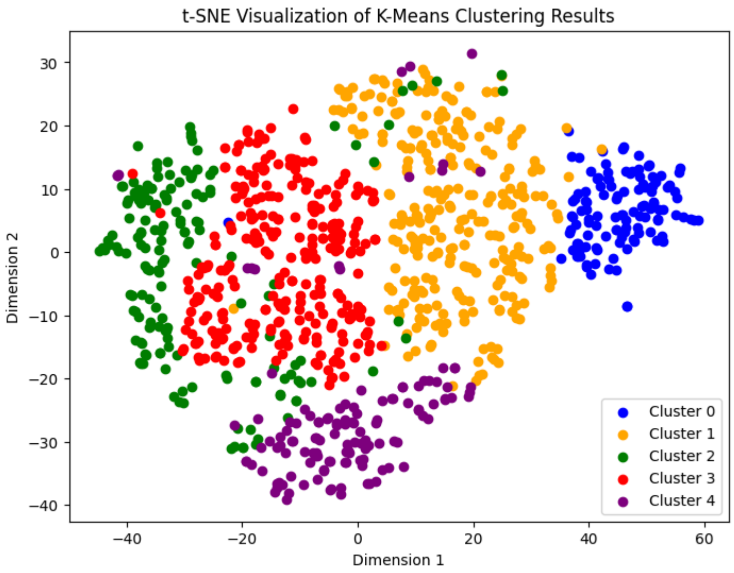
Results

K-Means Clustering





Based on the results - comparing clusters against increase in silhouette score versus decrease in inertia - **five would be the optimal cut-off point** for the number of clusters as five clusters shows the greatest increase in silhouette score paired with a decrease in inertia. However, the inertia and silhouette measurements are only meant to indicate the optimal number of clusters for the k-means algorithm; the T-SNE plot to visualize the groupings is below:



The T-SNE plot shows fine separation between the clusters and little overlap. To interpret the meaning of the two dimensions that the clusters rest on would require subject matter experts. However, as far as what was factored in determining similarities and differences across observations, the table below - comparing **label grouping and median** of certain numerical features - can offer insight:

Labels	Duration	Credit Amount	Installment Commitment	Residence Since	Age
0	24.0	3844.0	4.0	4.0	42.0
1	24.0	2872.0	3.0	3.0	35.0
2	12.0	1527.0	3.0	2.0	37.0
3	18.0	2160.5	4.0	2.0	30.0
4	18.0	2139.0	2.0	4.0	25.0

The label groupings in the table above would suggest that cluster ‘0’ corresponds to applicants oldest in age with longer duration and higher credit amounts. Cluster ‘1’, corresponds to applicants with the same duration profile as cluster ‘0’, but with lower credit amounts and slightly younger in comparison to applicants in cluster ‘0’.

Clusters ‘3' and ‘4’ are similar in both duration and credit amount, but vary on age, residential profile and current installment commitments. Lastly, cluster ‘2’ refers to the applicants with the lowest credit amounts and the duration profiles, but includes the second oldest applicant group with median age at 37.

Logistic Regression

Approach	Avg Precision	Avg Recall	Avg F1-Score	Avg Cost of Misclassification
1) Baseline Logistic	37.4%	92.3%	53.2%	116
2) K-Means (5) with Logistic Regression	70.5%	99.9%	82.7%	59.4
3) K-Means (5) with Logistic Regression - only modeling labels	70%	100%	82.3%	60

The table above shows the scores for the three approaches: the baseline logistic regression approach currently used by CapitalTwo, the first alternative approach that combines k-means clustering with the logistic regression model with the same explanatory variables as in the baseline, and the second alternative approach that combines k-means clustering with the logistic regression model, but only uses the clustering labels as predictors.

In the first alternative approach, the average precision across the 5-folds increased by 33.1% and the average recall increased by 7.6% across the 5-folds, leading to an overall increase in the average F1-score by 29.5%. In contrast, in the second alternative approach the average precision increased by 32.6% and the average recall increased by 7.7%, leading to an overall increase in the average F1-score by 29.1%. Regarding average cost of incorrect classification, both approaches led to an approximate 48% decrease, showcasing the improvement in risk management via the alternative approaches. In conclusion, the results between the two alternative approaches saw similar improvements from the baseline albeit the second alternative approach required just the labels as modeling inputs.

Conclusions

The purpose of this proposal was to assist Capital Two's modeling team in increasing model performance along with reducing modeling team and efforts to predict whether a credit card applicant should be deemed a good or bad candidate for approval. In its current approach, Capital Two applies the logistic regression model where ‘class’ is modeled against all explanatory features. However, this proposal explores the approach of pairing an unsupervised learning technique along with Capital Two’s logistic regression approach to increase performance and reduce modeling efforts.

Based on the literature review and the size of the data provided by Capital Two, k-mean clustering was employed to partition the data into five clusters. Post-clustering, two alternative approaches were explored that paired the k-means model with the stand-alone logistic regression approach:

1) Class modeled against all explanatory features used by Capital Two with the new addition of the k-means labels
2) Class modeled solely against the k-means labels.

Overall, both alternative approaches showed a significant increase in precision, recall and the consequent F1-Score paired with a significant decrease in the average cost of misclassification. However, given that reduction in modeling efforts is a primary objective for Capital Two, it’s recommended that Capital Two employ k-means clustering (5 clusters) along with the logistic regression modeling ‘class’ solely against the labels generated by the k-means clustering.

Appendix A

Fig 1: Correlation between numerical features (1 indicating perfect correlation)

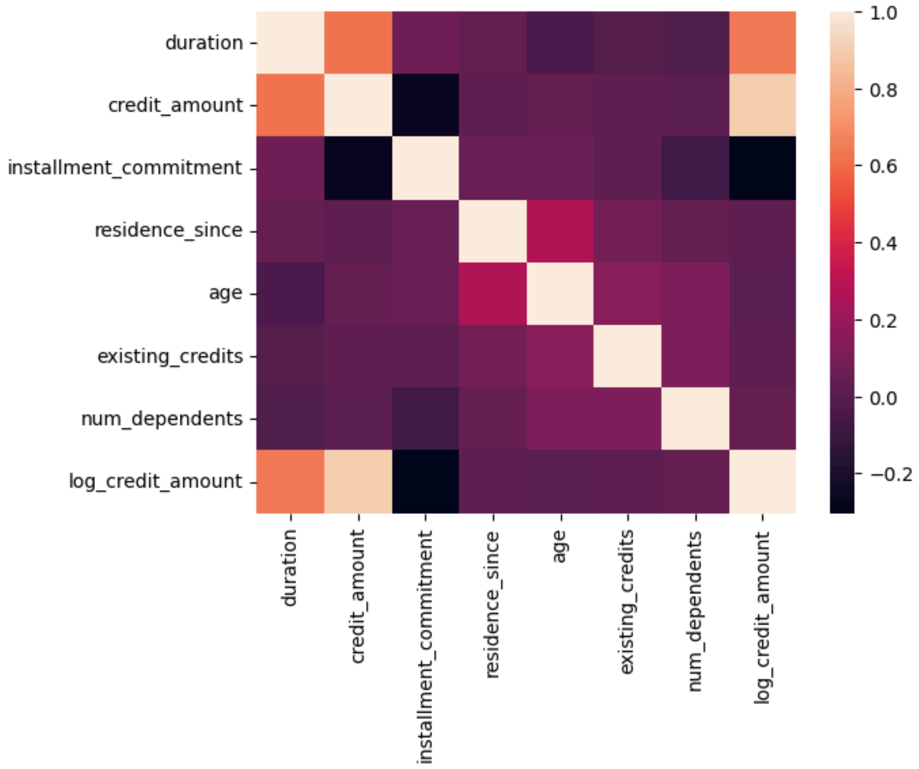


Fig 2: Age and Credit Amount Box Plots contrasted with Class Variable

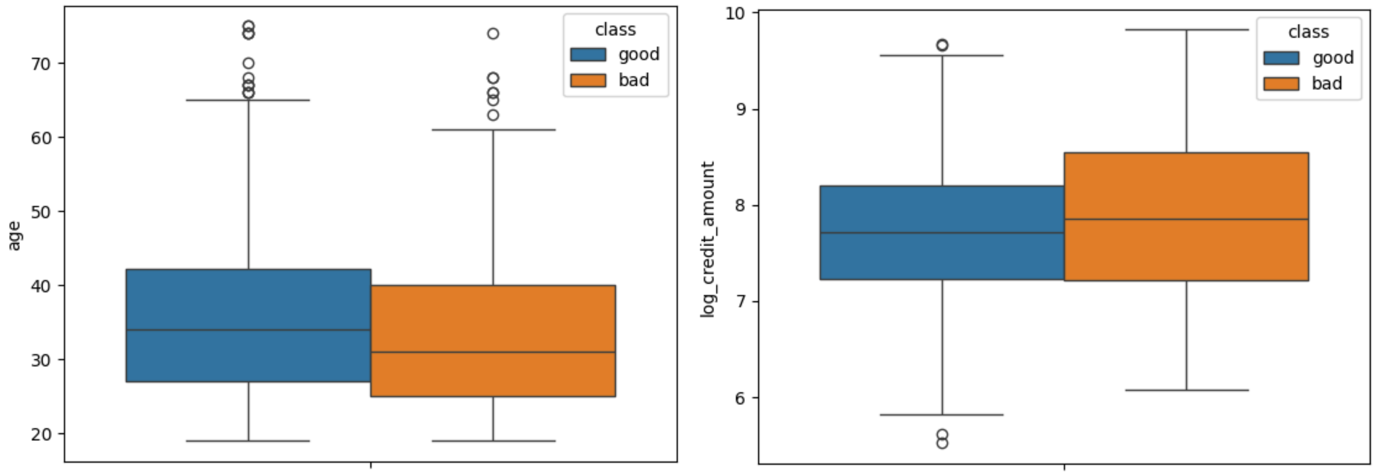
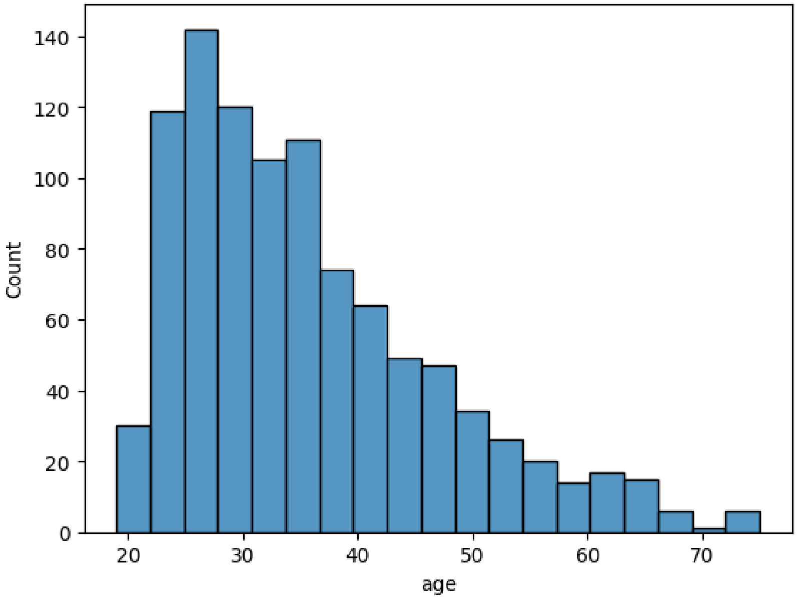


Fig 3: Table comparing median of numerical features across class labels

Class	Duration	Credit Amount	Age	Installment Commitment	Residence Since	Existing Credits	Number of Dependents
Bad	24	2574.5	31	4	3	1	1
Good	18	2244	34	4	3	1	1

Fig 4: Age histogram for right-skewness observation



Appendix B

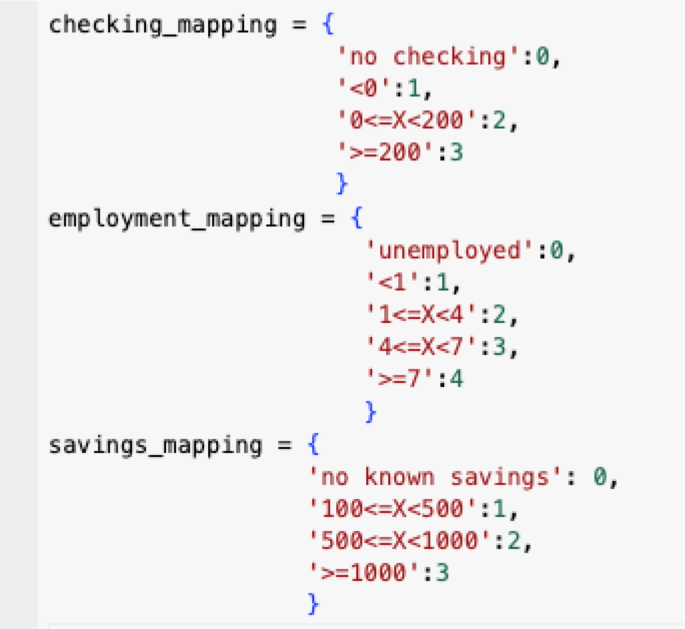
Breakdown of Categorical Variables with 3 or more levels

Feature	**Category Type**	**Levels**
property_magnitude	nominal	4
own_telephone	nominal	2
job	nominal	4
housing	nominal	3
other_payment_plans	nominal	3
checking_status	ordinal	4
other_parties	nominal	3

personal_status	nominal	4
employment	ordinal	5
saving_status	ordinal	5
purpose	nominal	7
credit_history	nominal	5

Appendix C

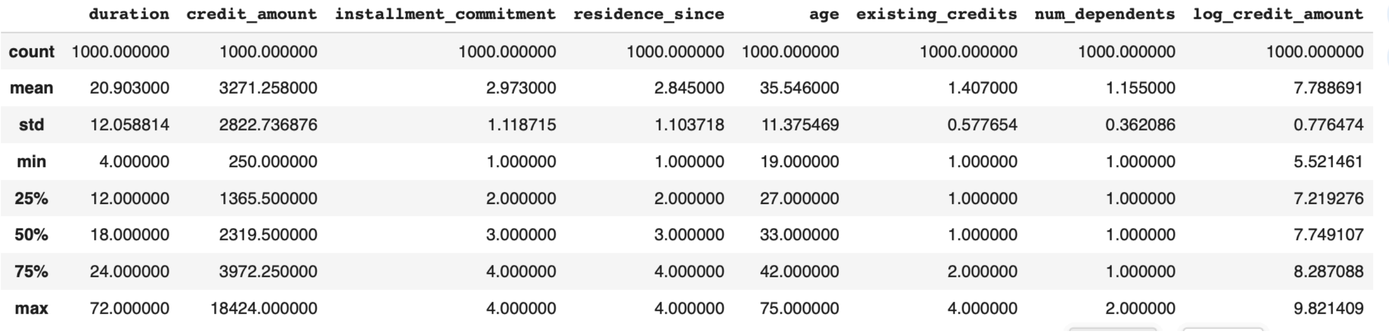
Breakdown of factor to integer mapping of ordinal variables



```
checking_mapping = {
                    'no checking':0,
                    '<0':1,
                    '0<=X<200':2,
                    '>=200':3
                    }
employment_mapping = {
                      'unemployed':0,
                      '<1':1,
                      '1<=X<4':2,
                      '4<=X<7':3,
                      '>=7':4
                      }
savings_mapping = {
                   'no known savings': 0,
                   '100<=X<500':1,
                   '500<=X<1000':2,
                   '>=1000':3
                   }
```

Appendix D

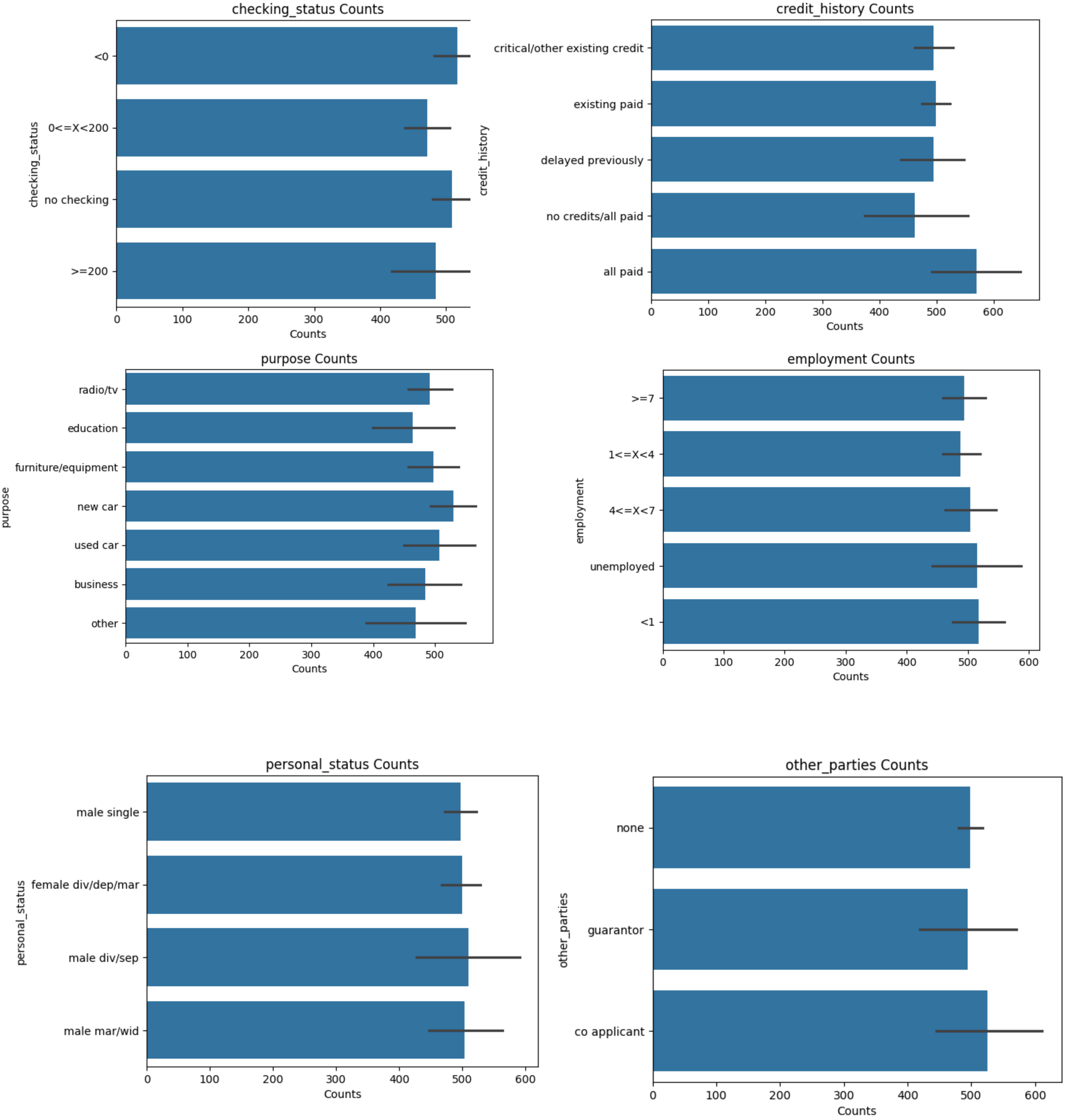
Fig 1: Statistical Summary of Numerical Variables

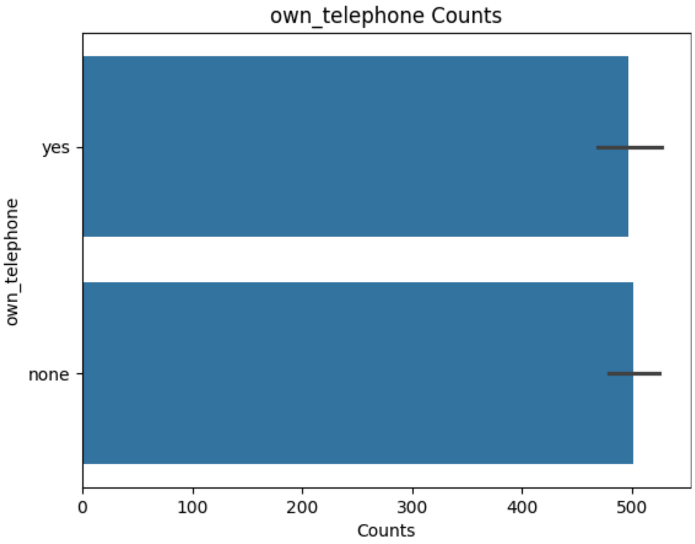


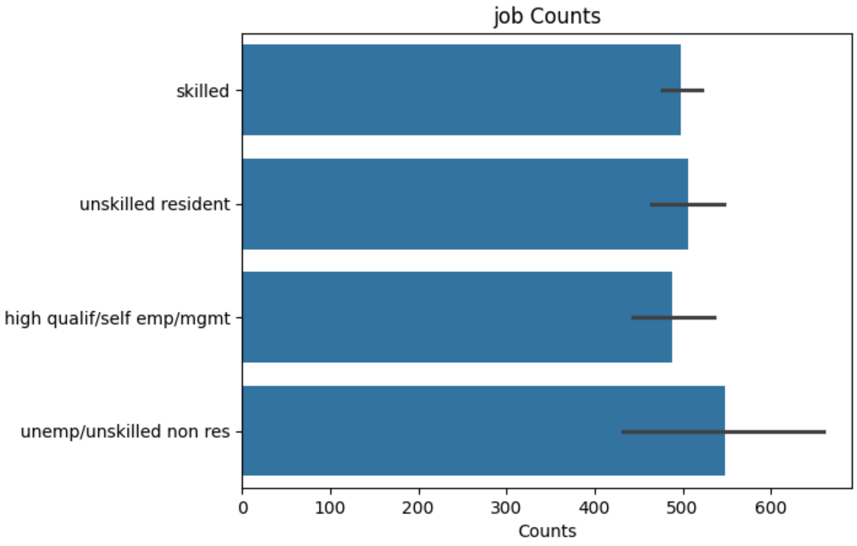
	duration	credit_amount	installment_commitment	residence_since	age	existing_credits	num_dependents	log_credit_amount
count	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000
mean	20.903000	3271.258000	2.973000	2.845000	35.546000	1.407000	1.155000	7.788691
std	12.058814	2822.736876	1.118715	1.103718	11.375469	0.577654	0.362086	0.776474
min	4.000000	250.000000	1.000000	1.000000	19.000000	1.000000	1.000000	5.521461
25%	12.000000	1365.500000	2.000000	2.000000	27.000000	1.000000	1.000000	7.219276
50%	18.000000	2319.500000	3.000000	3.000000	33.000000	1.000000	1.000000	7.749107
75%	24.000000	3972.250000	4.000000	4.000000	42.000000	2.000000	1.000000	8.287088
max	72.000000	18424.000000	4.000000	4.000000	75.000000	4.000000	2.000000	9.821409

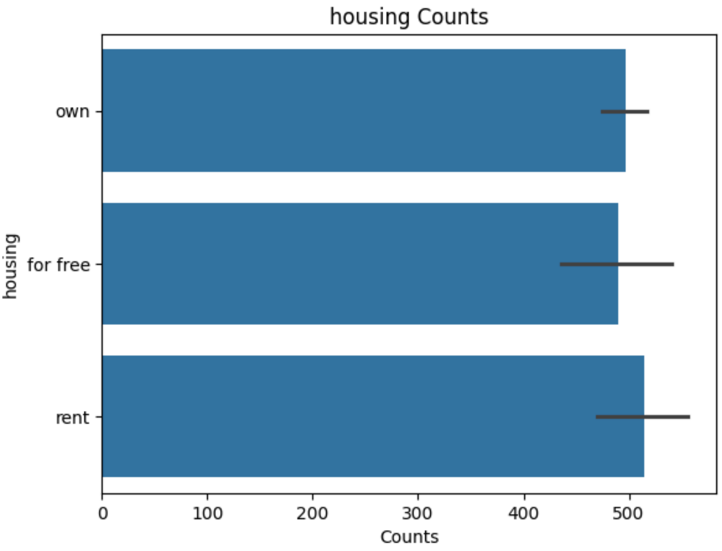
Appendix E

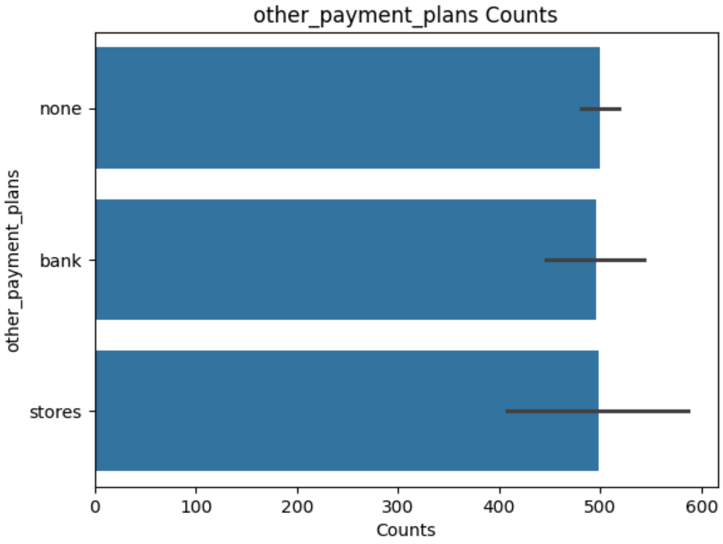
Figures: Barplots of Categorical Variables

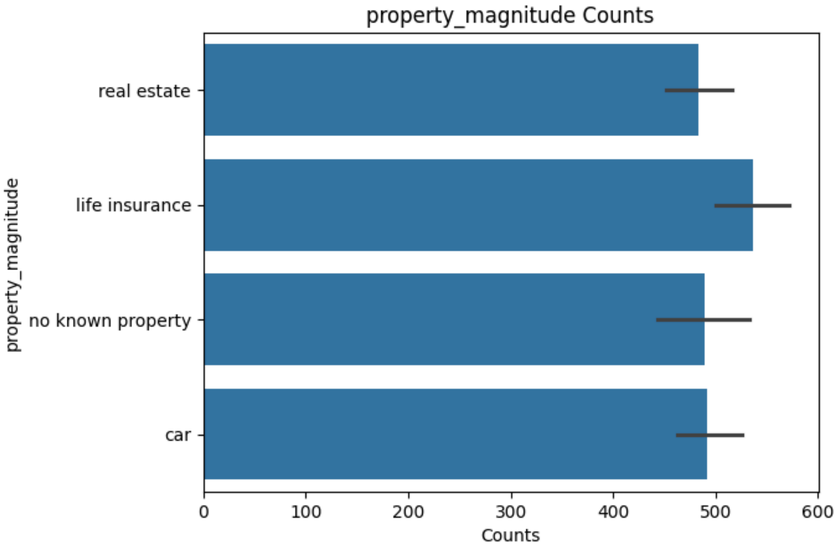



own_telephone Counts
own_telephone
yes
none
0
100
200
300
400
500
Counts


job Counts
job
skilled
unskilled resident
high qualif/self emp/mgmt
unemp/unskilled non res
0
100
200
300
400
500
600
Counts


housing Counts
housing
own
for free
rent
0
100
200
300
400
500
Counts


other_payment_plans Counts
other_payment_plans
none
bank
stores
0
100
200
300
400
500
600
Counts


property_magnitude Counts
property_magnitude
real estate
life insurance
no known property
car
0
100
200
300
400
500
600
Counts

References

Brus, Patrick. “Clustering: How to Find Hyperparameters Using Inertia.” Medium, July 29, 2021.https://towardsdatascience.com/clustering-how-to-find-hyperparameters-using-inertia-b0343c6fe819.

Budisteanu , Elena-Alexandra, and Irina Georgiana Mocanu. “Combining Supervised and Unsupervised Learning Algorithms for Human Activity Recognition.” National Library of Medicine - National Center for Biotechnology Information, September 2021. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8473063/#B23-sensors-21-06309.

Carcillo, Fabrizio, Yann-Aël Le Borgne, Olivier Caelen, Yacine Kessaci, Frédéric Oblé, and Gianluca Bontempi. “Combining Unsupervised and Supervised Learning in Credit Card Fraud Detection.” *Information Sciences* 557 (May 2021): 317–31. https://doi.org/10.1016/j.ins.2019.05.042.

Jarapala, Krishnakanth Naik. “Categorical Data Encoding Techniques.” Medium, March 27, 2023. https://medium.com/@jkkn.iitkgp/categorical-data-encoding-techniques-d6296697a40f.

Kumar, C.N.S., Rao, K.N., Govardhan, A., Sandhya, N. (2015). Subset K-Means Approach for Handling Imbalanced-Distributed Data. In: Satapathy, S., Govardhan, A., Raju, K., Mandal, J. (eds) Emerging ICT for Bridging the Future - Proceedings of the 49th Annual Convention of the Computer Society of India CSI Volume 2. Advances in Intelligent Systems and Computing, vol 338. Springer, Cham. https://doi.org/10.1007/978-3-319-13731-5_54

Kumar, Ajitesh. “MinMaxScaler vs Standardscaler - Python Examples.” Analytics Yogi, December 7, 2023. https://vitalflux.com/minmaxscaler-standardscaler-python-examples/#Why_is_Feature_Scaling_needed.