Advanced Data Mining: From Feature Engineering to Cluster Analysis

Project_1: Feature Engineering & Spatiotemporal Data Orchestration

Background

The dataset comprises **21,817,851** GPS points, collected from the driving records of **320** taxi drivers operating in the central area of Rome, Italy, over a **30-day** period, from 00:00:00 on February 1, 2014, to 23:59:59 on March 2, 2014. According to the official data description, GPS locations were recorded **every 7 seconds**, and records with a location accuracy lower than **20 meters** were filtered out prior to release.

1.A 2D plot as ***Fig. 1*** shows the Rome taxi GPS points with full values. The graph clearly shows the dense characteristics of the data, and extreme data can still be observed. The definitions and identification methods of invalid (red), outliers (purple) and noise (orange) are introduced in detail below, as well as the results.

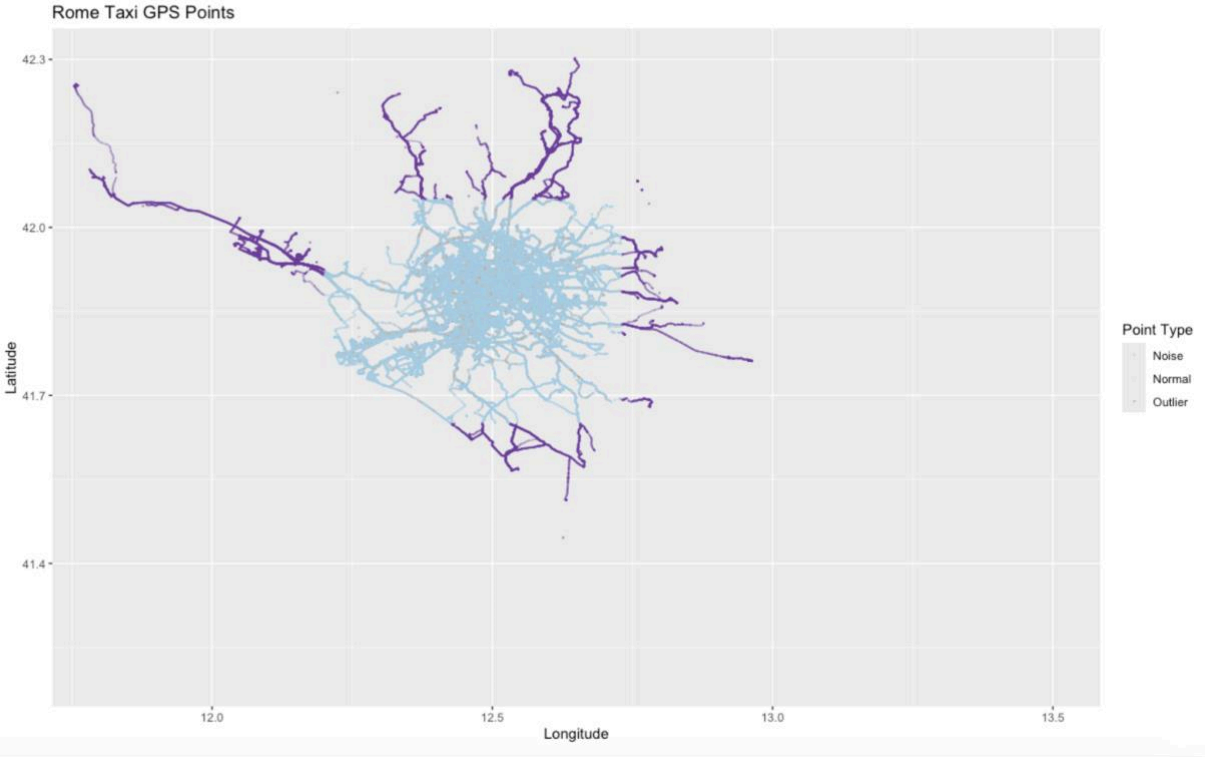


Fig. 1. The Plot of Rome Taxi GPS Points

1.1 Identify and Explanation Invalid

Invalid general indicates records in the data set that have no analytical value, logical errors, or format errors. In this case, it identifies by:

- Missing value (NA), include missing latitude, longitude, time
- Empty value, such as point (0,0)

- Find completely repeated records which the criterion for duplication is whether the first three columns of the data set are exactly the same. Because the first three columns are DriveNo, Latitude, and Longitude, this means that if the same taxi repeatedly records the same location at a certain moment, it may be data collection duplication or system failure. This kind of data is of no value for trace analysis.

1.2 Identify and Explanation Outliers

Outliers refer to values that deviate significantly from the statistical characteristics, usually extreme data points that do not conform to the main distribution. Considering that the data in this case are geographical locations, a rule based on the Roman city boundaries is used to identify them.

- Latitude: 41.65° N - 42.05° N
- Longitude: 12.20° E - 12.73° E

1.3 Identify and Explanation Noise

Noise is defined as impact points that interfere with the analysis of major driving behaviours. Look for unusual driving behaviours as influence points, especially jumps in speed, which are not plausible in the data.

For each taxi, the dataset is first sorted by time. The spherical distance between consecutive GPS points is computed using the Haversine formula based on latitude and longitude. The time difference between each point and its previous observation is also calculated. Speed (in km/h) is then derived by dividing the distance by the time interval, considering only non-missing distances and positive time differences. Finally, a set of reference speed percentiles (e.g., 50%, 75%, 90%, 95%, and 99%) is computed to analyze the distribution and identify abnormal values above 200 km/h. A speed threshold of 200 km/h was set based on two considerations:

- First, while speeds above 100 km/h are rare in urban areas, setting the threshold too low may misclassify some legitimate fast driving.
- Second, speeds above 200 km/h are nearly impossible in the city and are usually caused by GPS jumps or recording errors.

Therefore, values over 200 km/h were treated as clear impact points to balance real behaviour with noise detection.

Percentile	Value	Meaning
50%	1.47 km/h	Half of the taxis have a speed less than or equal to this speed. There may be a large number of stationary or super slow-moving points due to traffic jams and parking.
75%	23.31 km/h	75% (upper quartile) of points at or below this speed, which may represent the normal average speed in the main urban area.
90%	49.52 km/h	This means that the boundaries of normal fast driving in urban areas.

95%	77.19 km/h	Close to the abnormal area, indicating that values above this value may be a jump point.
99%	350.07 km/h	The top extreme values may be influencing points caused by GPS jitter, data jumps.

Count and summarize the values and proportions of invalid points, outliers, and noise:

Type	Amount	Percentage
Invalid	0	0
Outliers	58316	0.267%
Noise	284953	1.306%

After removing invalid, outliers and noise, remaining data cleaned GPS points: **21,474,582**.

Question (b):
Compute the minimum, maximum, and mean location values as follow below:

Type	Minimum	Maximum	Mean
Latitude	41.65004	42.05	41.89181
Longitude	12.2	12.72998	12.47303

Question (c): Calculating Activity
My approach and explanation for activity performance:

1. Group by DriveNo, which uniquely identifies each driver. Since each driver's GPS data is collected at a 7-second interval, each DriveNo appears multiple times, corresponding to multiple GPS points throughout the observation period.
2. For each driver, calculate the time gap between the first and last recorded GPS point. This time gap is interpreted as the activity duration —— indicating the time span in which the driver was active in the dataset.
3. The activity duration does not reflect continuous driving, but rather the period between the earliest and latest appearance of a taxi in the dataset.
4. Finally, drivers are ranked based on this activity duration to identify the:
   Most activity driver = The longest duration
   Least activity driver = The shortest duration
   Average activity driver = across all drivers

Results and Explanation:

- **Most active Taxi (DriveNo = 352):**
  - 0 Acting for the full period of 720 hours (30 days), suggesting complete participation in the dataset.
- **Least active Taxi (DriveNo = 283):**
  - 0 This taxi had data for only 1.13 hours, which may reflect partial data collection.
- **Average active duration is 631.64 hours:**

0 Taxis are active for 631.64 hours, or approximately 26.2 days, indicating high data coverage and supporting reliable trace analysis.

Question (d):

i. The taxi ID is 261 as My student number is 8374090.

Fig. 2 illustrates the plot of the location points for taxi ID 261. It is centralized, with most of the locations concentrated in the central area of Rome. In addition, the distribution of the locations is radial, radiating from the centre to the periphery, forming a long driving route, indicating that drivers occasionally drive to the periphery. In particular, a long extended track can be seen in the southwest (about 41.80, 12.3).

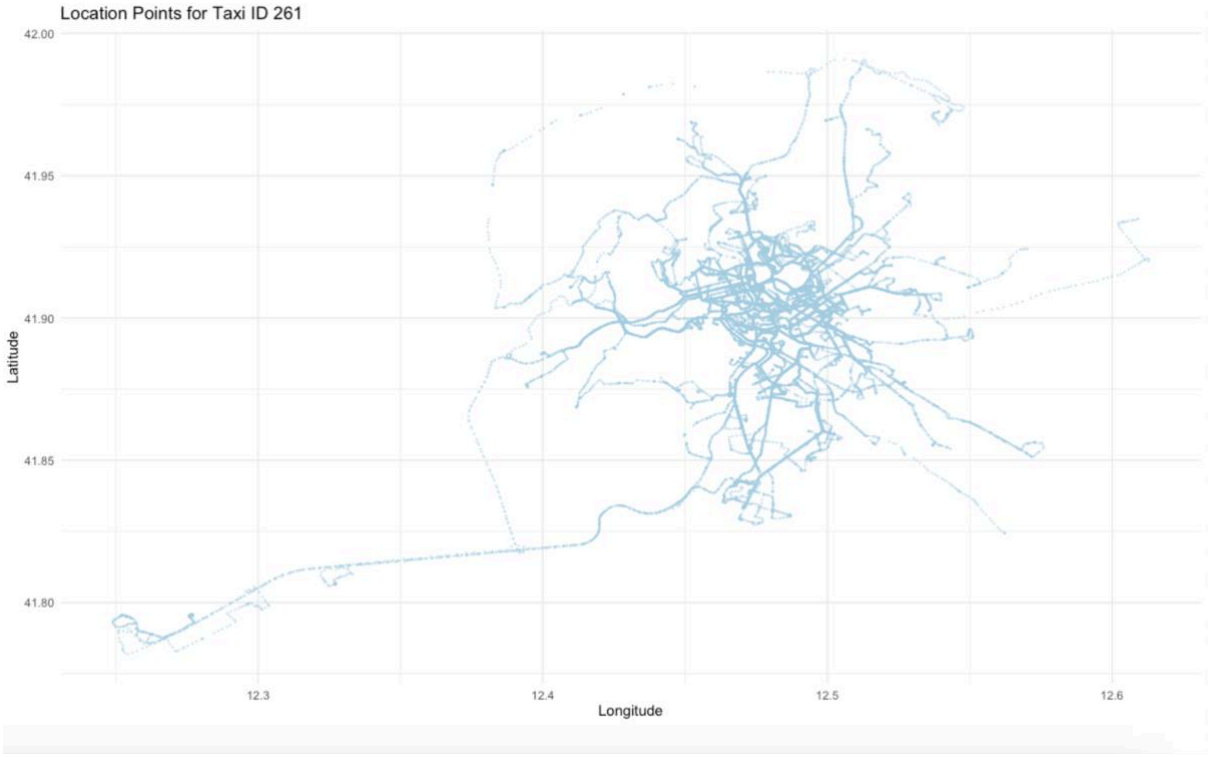


Fig. 2. The Location Points for Taxi ID 261

ii. Compare the location value:

Compared to the global GPS dataset, Taxi ID 261 shows a more localized spatial range:

- The latitude range is from 41.80 to 41.99, with a mean of 41.90, while the global latitude spans from 41.65 to 42.05.
- The longitude range is from 12.25 to 12.61, with a mean of 12.48, compared to the global range of 12.20 to 12.73.

Type	min_lat	max_lat	mean_lat	min_lon	max_lon	mean_lon
Taxi 261	41.78189	41.99083	41.90135	12.24886	12.61292	12.47827
Global	41.65004	42.05	41.89181	12.2	12.72998	12.47303

These statistics support the observation that Taxi 261 primarily operates within central Rome, with relatively limited excursions toward the city's edges.

iii. Compare activity status between taxi 261 and global:

The total time span between the first and last recorded GPS point of **taxi ID 261 is 601.19 hours**.

Compared with **the global mean of 631.64 hours**, the taxi was slightly less active than average, but still operated for over 83% of the 30-day period. The operational pattern of Taxi 261 indicates that it was a **relatively active driver**, although its total driving time of 601.19 hours was slightly below the global average duration.

Index Items	Duration (hours)
Taxi 261	601.19
Global Min Duration	1.13
Global Max Duration	720
Global Mean Duration	632

iv. According to the Haversine-based computation method, the total driving distance of **Taxi 261** during the 30-day observation period was **2,573.96 kilometres**.

This corresponds to an average daily distance of approximately **85.79 kilometres**. The spatial trace of Taxi 261 was primarily concentrated in central urban areas, with most locations falling within the core boundaries of Rome. However, some long-distance excursions were observed, with the farthest trace reaching near the southwestern edge of the city.

Project_2: Exploratory Data Analysis & Unsupervised Cluster Analysis

1.Background

The dataset contains a total of **2500 records** and **46 attributes** for each customer in the past. Among them:

- **1962 records (78.5%)** have been assessed for creditworthiness.
- **538 records (21.5%)** are unassessed (label = 0) and will be used for future prediction.

The class distribution among the assessed samples of credit rating is:

Rating	Count	Percentage
1	**483**	**24.6%**
2	**970**	**49.4%**
3	**509**	**25.9%**

This distribution indicates a significant imbalance, where Class 2 shows nearly half of the dataset. As a result, making it more difficult to predict the model to equally distinguish all classes, particularly the minority class (Class 1 or 3), and balance accuracy must be considered during model testing between different classes.

2.Attribute Analysis

Utilizing Pearson to calculate top10 correlation with credit rating ***(see Fig. 3)*** most associated with credit rating. Overall, the correlation strength was moderate, with no attribute showing a strong linear relationship ($r \approx 1$). Only the top four attributes had above 0.2, while from the sixth onward, values fell below 0.05 that indicates the linear correlation is weak with credit rating.

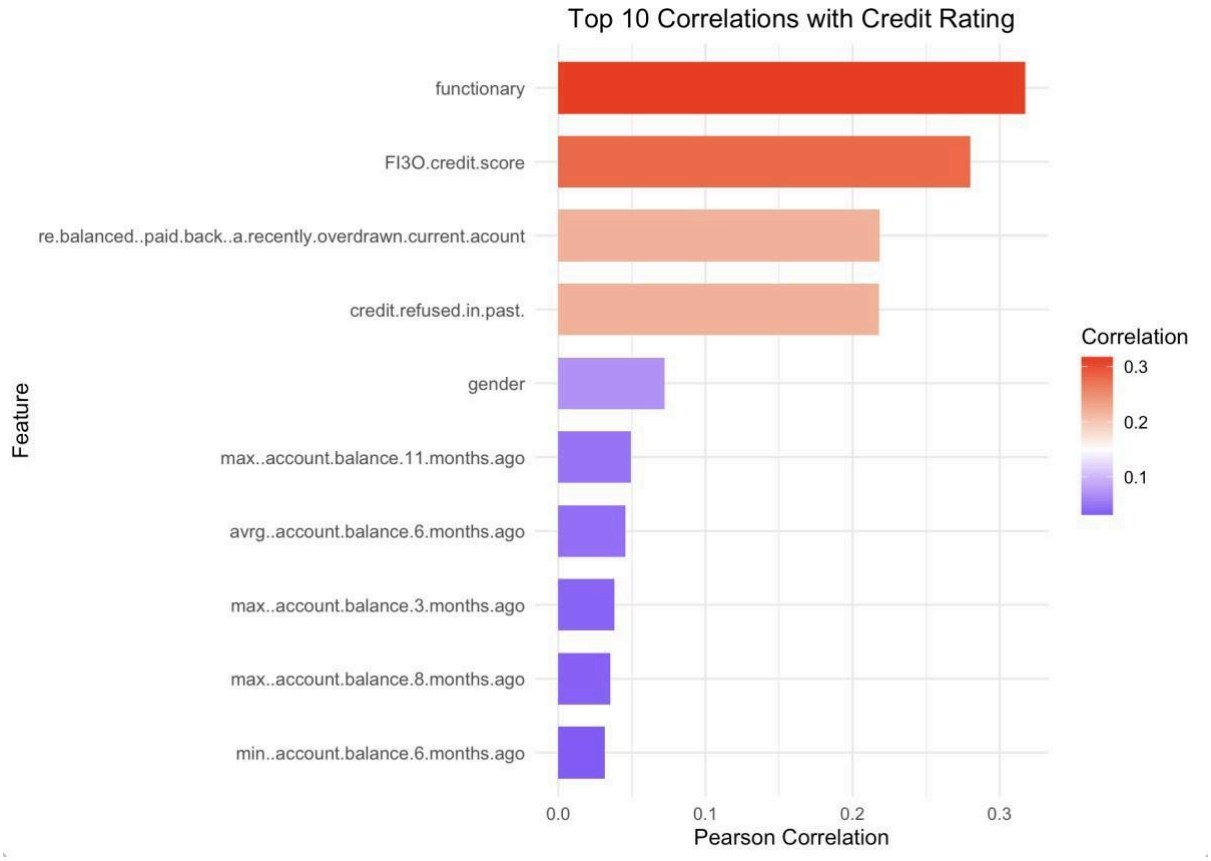


Fig. 3. TOP 10 Attribute Sorting

In summary, the top 5 features — **Functionary, FI3O.credit.score, Repaid Overdrawn, Credit Refused,** and **Gender** — were selected based on their relative correlation strength and visual separability cross credit rating classes.

3. Self-Organizing Map (SOM)

3.1 Inspiration of SOM results for supervised model learning : don't expect 100% classification accuracy. The reasons are summarized in three points:

- Fuzzy class boundaries
- Class overlap
- Sparse or imbalance mapping

3.1.1 Reasons why perfect classification is not achievable Step 1: SOM analysis on all features (*see Fig. 4*)

- Grid size: 20 * 20 hexagonal
- Training steps: 100 (rlen)
- Input: 45 features

(1) U-Matrix analysis:

- The overall map demonstrates partial structural clustering, but with no sharply defined boundaries.
    - 0 The left of bottom region shows several highlighted yellow paths, suggesting strong local dissimilarities between adjacent neurons.
    - ○ Most of the map remains in orange to red tones, indicating relatively low interneuron distances, and thus, high similarity in many regions of the input space.

(2) Class Mapping analysis:

- The feature boundaries between categories are not obvious and difficult to distinguish by simple rules:
    - 0 The three credit ratings (1-red, 2-green, 3-blue) are widely distributed, with no clearly dominant cluster for any specific class.
    - ○ Similar samples are clustered in a few areas, but there is no obvious cluster of similar samples overall.
    - ○ The overlapping black circles indicate that multiple samples are mapped to the same SOM units, suggesting regions of high sample density. While the class labels are not explicitly shown, the possibility of class overlap in these areas cannot be entirely ruled out.
    - ○ A small portion of neurons remain empty, indicating either data sparsity or uneven distribution across the input space.

To further explore the distribution patterns of each feature, we generated feature heatmaps ***(see Appendix A)*** and included them in the Appendix. Overall, these plots confirm that the class discrimination of features is generally poor, except for the top 4 attributes. Interestingly, the attribute **Self.employed** shows a clear clustering pattern in the SOM visualization, indicating its potential in distinguishing the classes. However, it is not among the top 10 features with the highest linear correlation with the target variable. **This suggests that features with low linear correlation may still present strong class-separating patterns in nonlinear models like SOM**.

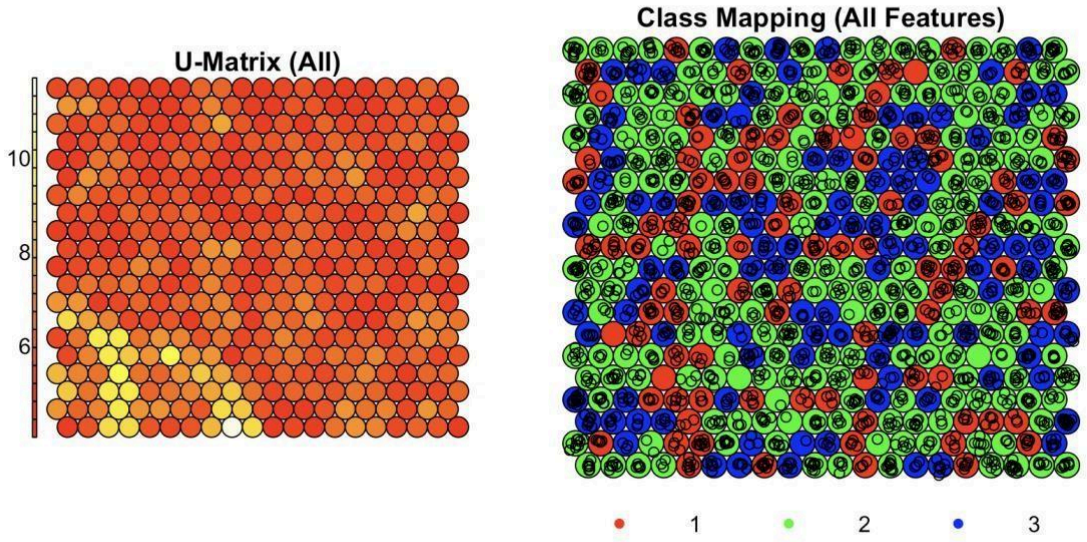


Fig. 4. U-Matrix and Class Mapping of Full Features

Step 2: SOM analysis using top 5 features (*see Fig. 5*)

Compared to the full-feature SOM, the U-matrix trained with the top 5 features is cleaner and more compact, but still shows some important boundary patterns. The bright yellow lines in the figure indicate that some areas are more different from each other. However, most areas are similar in color, which means that the overall difference between neurons is small. This reflects the lower feature variance, which means that the selected features are more consistent. Therefore, the model may have a harder time distinguishing similar classes and may be more sensitive to noise or overfitting (e.g., Class 2).

In the class mapping based on the top 5 features, although dimensionality has been reduced, the samples mapped to SOM neurons show a clear tendency toward specific class. Notably, there is a visible concentration of samples within neurons associated with Class 2, suggesting a re-centralization of the clustering pattern. However, the boundaries between classes remain indistinct, and the overrepresentation of Class 2 may indicate a risk of overfitting. These observations further highlight that, even after feature selection, the classification task remains non-trivial.

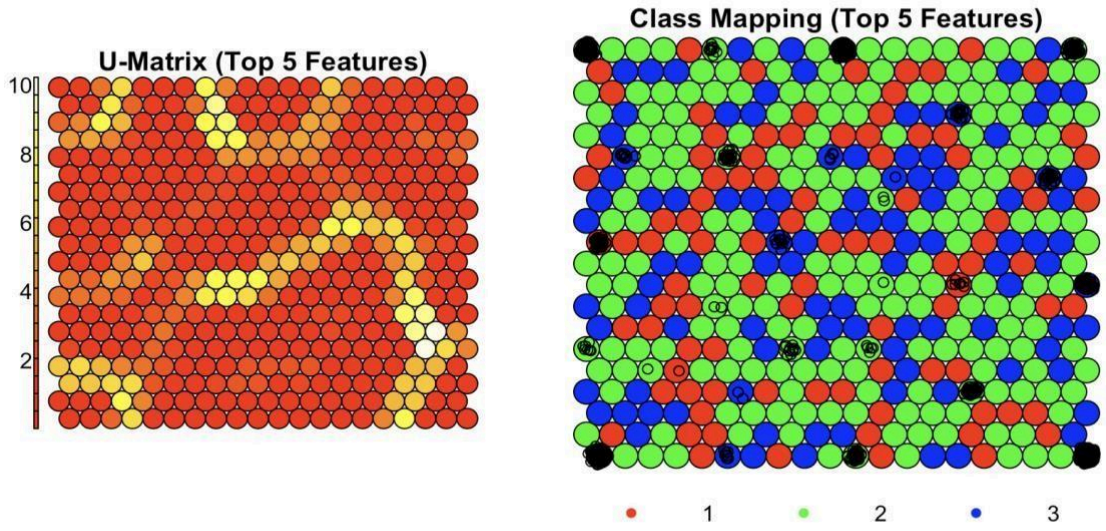


Fig. 5. U-Matrix and Class Mapping of Top 5 Features

3.2 Implications for Classification Model Design

The SOM analysis reveals that the dataset exhibits moderate structural complexity, characterized by overlapping class boundaries and imbalanced class distributions. These findings suggest that it is unrealistic to expect 100% prediction accuracy, especially when using simple linear models. Therefore, classification models must be carefully designed to address these challenges.

Based on the patterns observed in the U-Matrix and Class Mapping, it is evident that the credit rating problem involves nonlinear feature interactions and regionally mixed class separability. To handle these complexities, the chosen classification model should:

- Be nonlinear, such as MLP, decision tree or random forest.
- Incorporate regularization to prevent overfitting due to class overlap.
- Apply cross-validation to ensure robustness.
- Possibly use class weighting or sampling to address the imbalance, particularly for underrepresented classes such as class 3.

In summary, the SOM analysis provides valuable pre-modelling diagnostic insights that inform the design of robust, flexible, and generalizable supervised learning models.

Project_3: Predictive Modeling & Performance Evaluation

a.) To maximize the accuracy of predicting customer credit ratings, a comprehensive strategy was conducted that focused on improving model stability, reducing overfitting and enhancing generalization:

- Data split adjustment: the dataset was split into 70% training and 30% testing, increasing the train set instead of the original 50/50 split used in the lab. This provided the model with more data to learn from, increasing robustness.
- Full feature utilization: based on the SOM analysis in Question 1, the top 5 selected features were found insufficient for accurate classification, especially, Class 2 is almost twice as many as Class 1 and 3.
- Outlier consideration: SOM visualization shows some data with discrete, so outliers will become a strategy to be verified here.
- Model optimization:
    - Increase model depth and number of hidden layers;
    - Reduce the learning rate to achieve smoother convergence;
    - Increase the number of iterations to allow for enough training cycles;
- Prediction strategy: Use the 402040 method and adjust the confidence threshold to balance coverage and accuracy.

b.) The maximum value in the accuracy of predicting the credit rating is **62.63%** and **coverage of 79.96%**, based on the 402040-method applied to the test set.
This result is derived using the strategy outlined in part (a), where an MLP model is trained on the full set of 45 input features using 70% of the assessed dataset. The final model configuration is as follows:

- Hidden layer structure: size = c (15,10).
- Learning rate: 0.005.
- Maximum iterations: 250.
- Classification strategy: 402040 with l = 0.5, h = 0.5.

Compared to the baseline lab model ***(see Appendix B)*** with **55.86% accuracy** and **56.53% coverage** which evaluated on the same test set. This configuration indicates relatively stable predictions, with significant improvements in prediction accuracy, generalization, and classification ability, and a greater optimization in the overfitting problem. Observation of the training performance plots of the new MLP model ***(see Appendix C)*** shows that SSE presents smooth convergence and stronger generalization ability; the FIT curve is close to the diagonal line and the fit is clear; although the Class 2 classification ability of ROC is average compared with the Lab model, the accuracy is slightly improved.

While the strategy of identifying and removing outliers was also tested, and **a peak accuracy of 69.17%**, it significantly reduced **coverage to 43%** and the number of Class 1 predictions that are accurate is **0**. As such, this strategy was excluded in the final solution.

c.) **A 100% accuracy cannot be achieved** as there is a serious overlap between different classes ***(see Fig. 6),*** especially the distribution between Class 1 (red) and Class 3 (green) is almost cross-mixed, indicating that the classification model is difficult to distinguish perfectly. In addition, feature representation limitations, although the model was trained on all 45 features, some of them have weak correlation with the target variable as shown in ***Appendix A***.

This indicates that some customers with similar attributes may belong to different credit categories, so perfect classification cannot be achieved. In addition, the MLP model, although optimized, still cannot fully learn all the subtle differences and avoid overfitting.

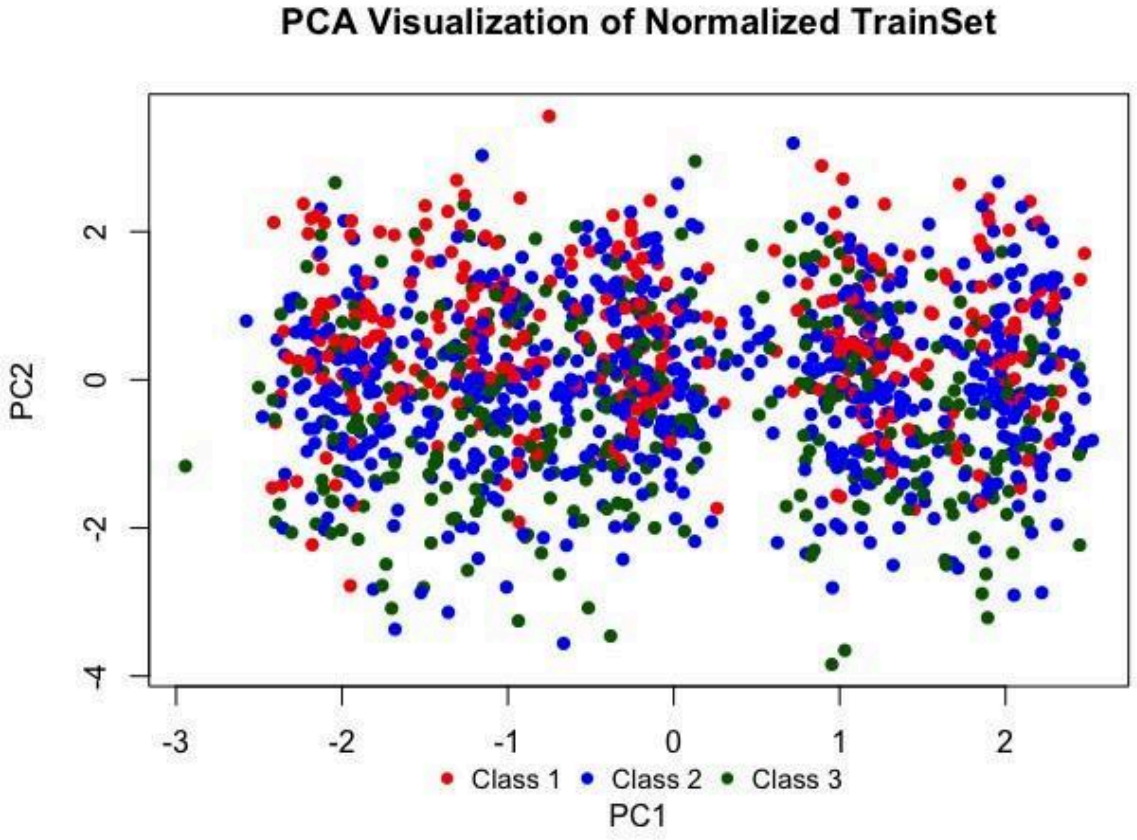


Fig. 6. PCA Visualization of Normalized Trainset

To get closer to 100% accuracy, there are several improvements:

- Larger and more balanced datasets, ensuring the sample size of each class is at the same level and has better representation.
- Integrate more relevant features, such as behavioural data or external credit scores.
- Enhance data quality, including feature normalization and noise reduction.
- Model upgrade, using decision trees or random forests.
- Hyperparameter tuning and cross-validation were performed to further refine model capacity and reduce overfitting.

d.) The Random Forest (RF) configures 700 trees and mtry = 10, and the results were tested using the 402040 strategies (l=0.5, h =0.5), the predicting accuracy is **60.1%** and coverage of **99.15%**. To compare the result with MLP as follows:

Items	MLP	Random Forest
Accuracy	62.63%	60.10%
Coverage	79.96%	99.15%

Class1 Accuracy	71	75
Class2 Accuracy	184	232
Class3 Accuracy	89	44
0 (Rejected)	118	5

Strengths:

- RF provides significantly higher coverage, classifying nearly all test set instances (only 1% rejected).
- RF reduces overfitting and improves robustness by training many decision trees on random subsets of data and features (mtry), then combining their predictions. This averaging process smooths out noisy or biased trees and leads to more stable, reliable predictions.
- RF can get clear feature importance as shown below, which is more transparent than MLP. Additionally, except for the top 3, the other attributes are all below 2 and negative value. This indicates that the features are not representative enough and the data quality is poor.

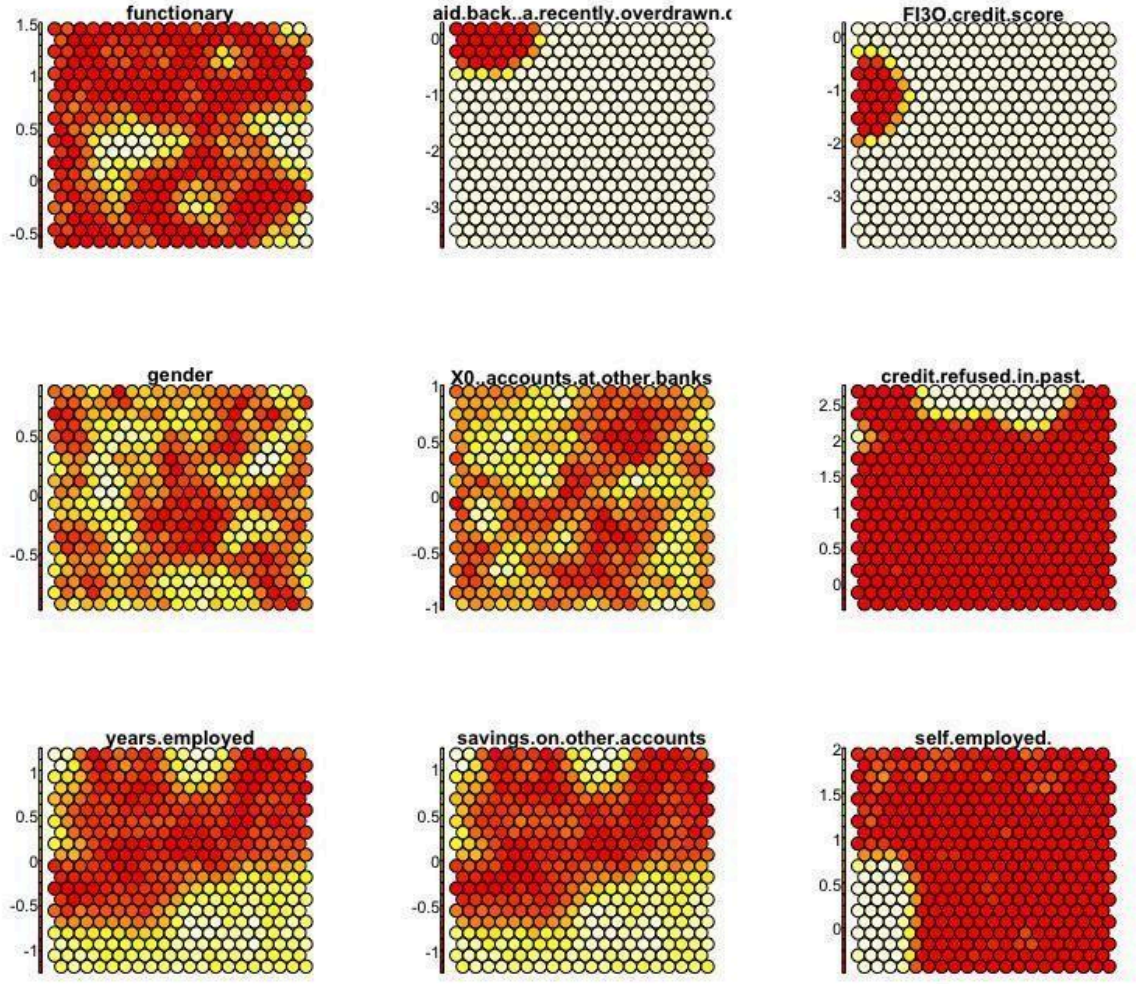
Features	MeanDecreaseAccuracy
FI30.credit.score	43.3
functionary	42.8
re.balanced.paid.back	33.37

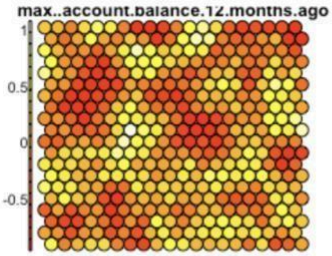
Weakness:

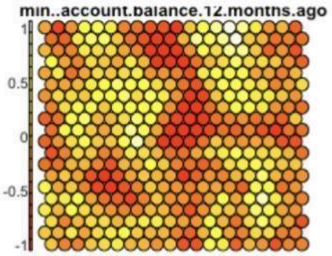
- RF has a slightly lower accuracy compared to the optimized MLP model (62.63%).
- As ntree increases, the training time gets slower for RF.

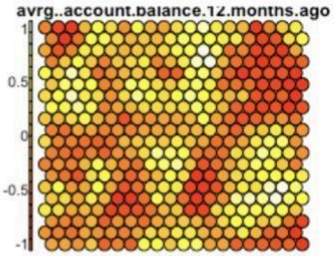
Although the MLP model has a slightly higher accuracy, it rejected **20.03%** of the samples, which impacted its actual usability. In contrast, RF offers a more inclusive prediction strategy, delivering strong performance across most classes without requiring complex structure tuning, and is also better than the “black box” MLP in terms of transparency, making it a more reliable and interpretable option for real-world deployment.

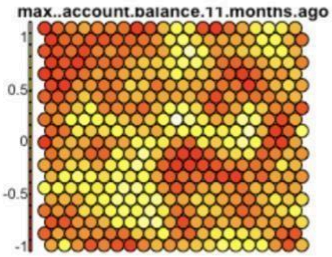
Appendix A - Feature Heatmap of Attributes

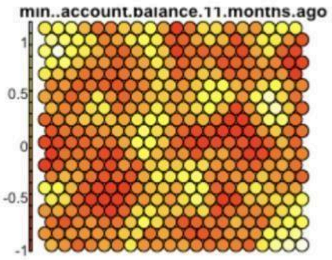


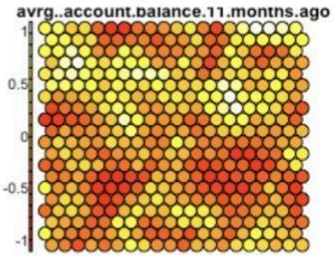

max..account.balance.12.months.ago

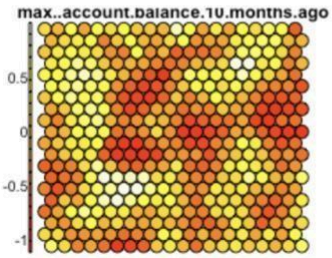

min..account.balance.12.months.ago

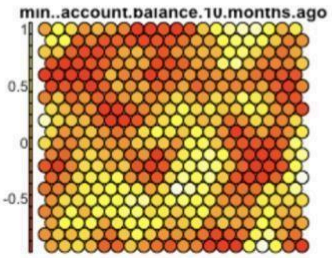

avrg..account.balance.12.months.ago

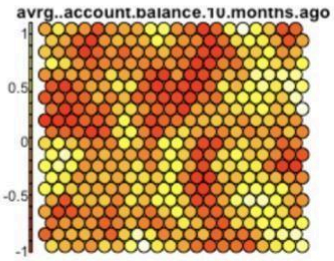

max..account.balance.11.months.ago

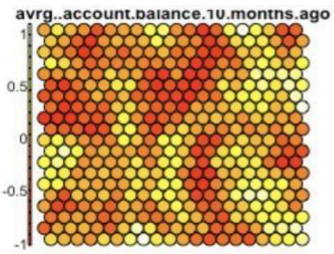

min..account.balance.11.months.ago

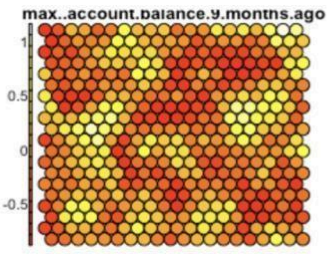

avrg..account.balance.11.months.ago

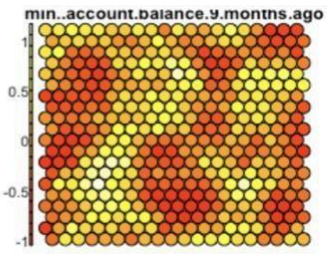

max..account.balance.10.months.ago

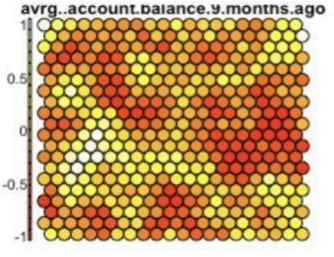

min..account.balance.10.months.ago

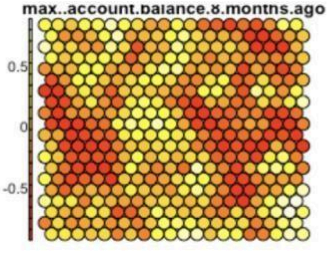

avrg..account.balance.10.months.ago

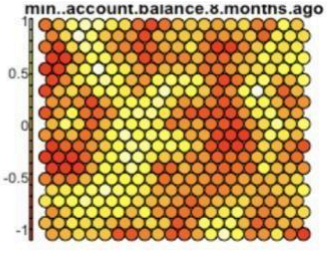

avrg..account.balance.10.months.ago

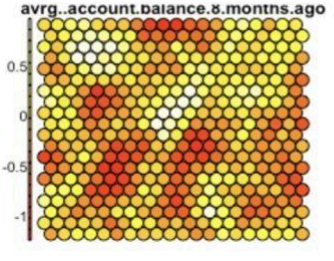

max..account.balance.9.months.ago

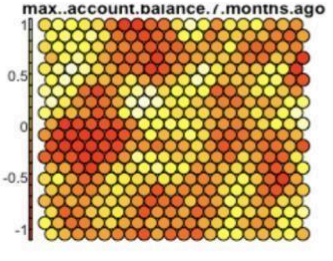

min..account.balance.9.months.ago

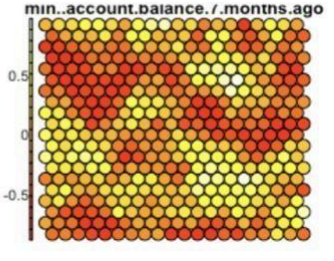

avrg..account.balance.9.months.ago

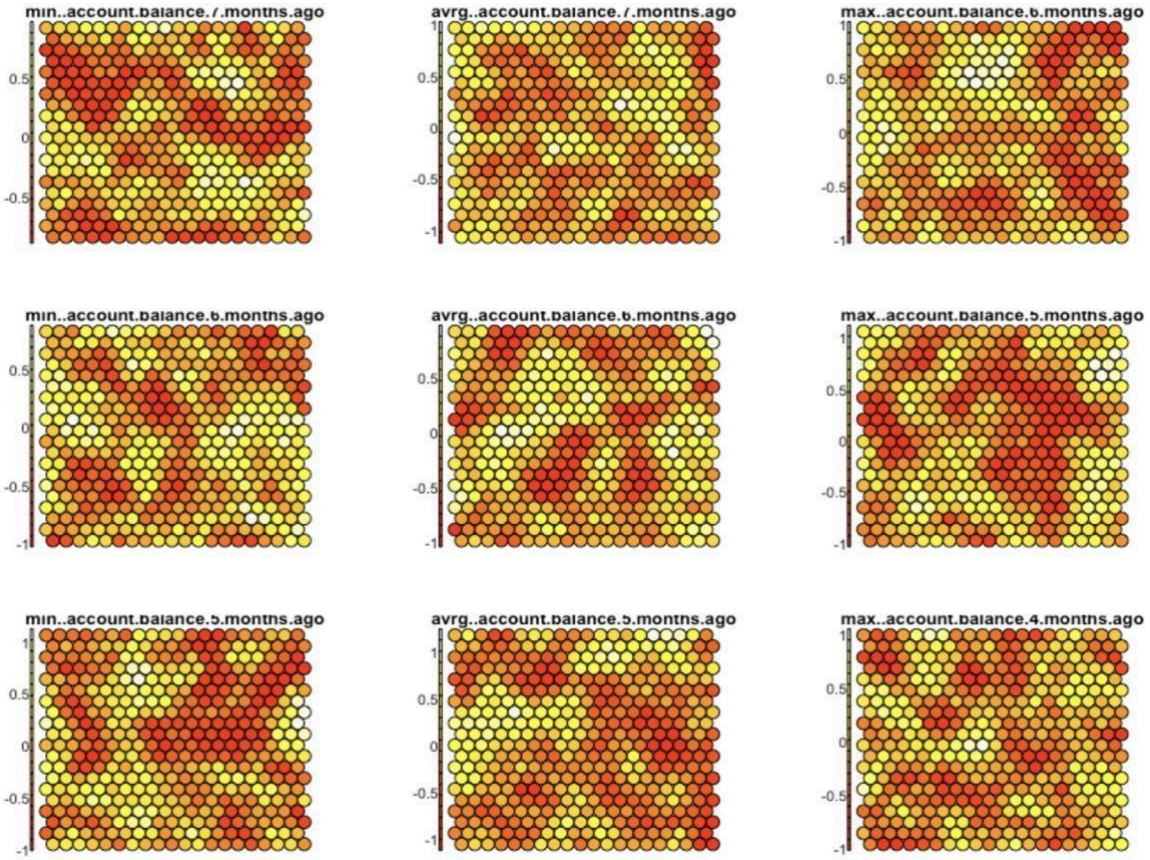

max..account.balance.8.months.ago

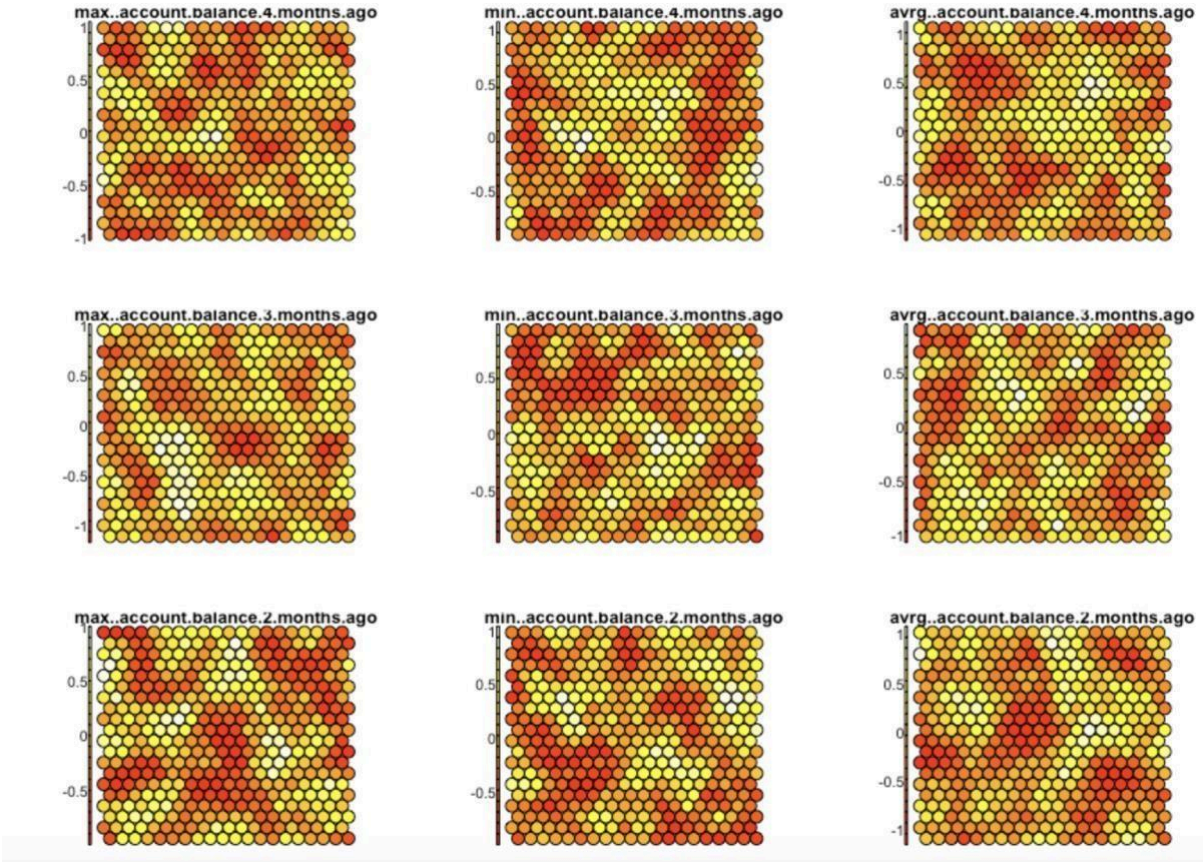

min..account.balance.8.months.ago


avrg..account.balance.8.months.ago

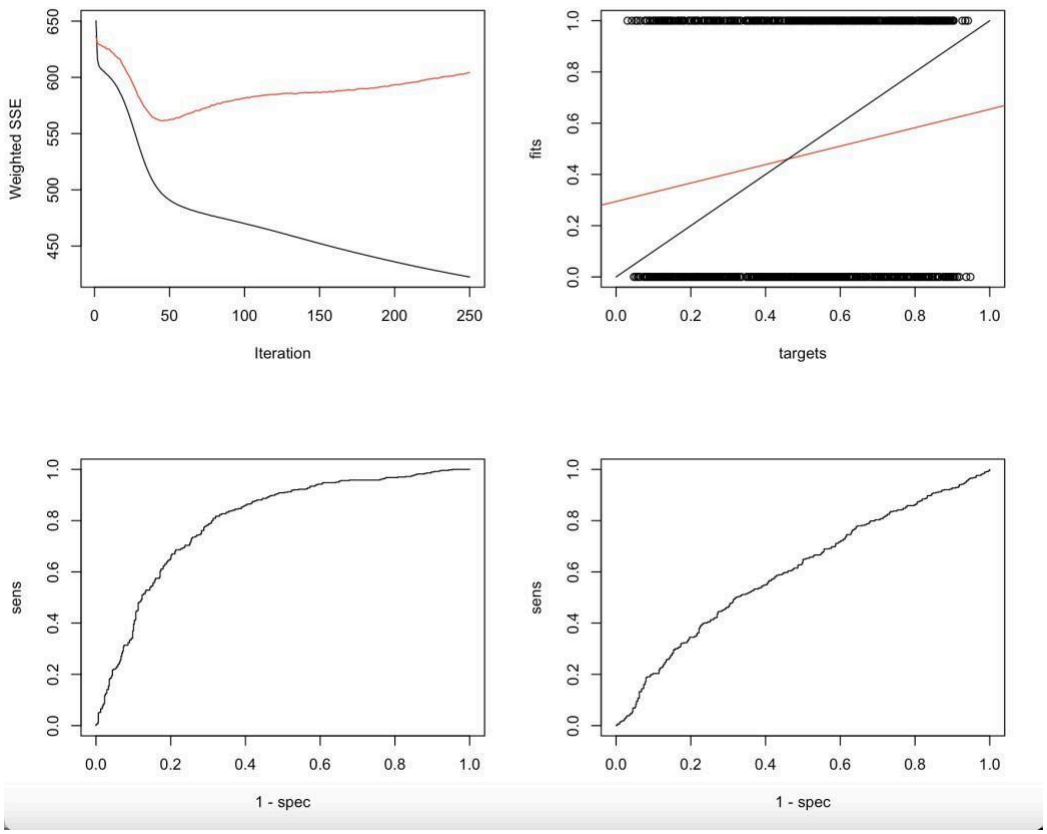

max..account.balance.7.months.ago


min..account.balance.7.months.ago


min..account.balance.7.months.ago
avrg..account.balance.7.months.ago
max..account.balance.6.months.ago
min..account.balance.6.months.ago
avrg..account.balance.6.months.ago
max..account.balance.5.months.ago
min..account.balance.5.months.ago
avrg..account.balance.5.months.ago
max..account.balance.4.months.ago


max..account.balance.4.months.ago
min..account.balance.4.months.ago
avrg..account.balance.4.months.ago
max..account.balance.3.months.ago
min..account.balance.3.months.ago
avrg..account.balance.3.months.ago
max..account.balance.2.months.ago
min..account.balance.2.months.ago
avrg..account.balance.2.months.ago

Appendix B - MLP Model Trained in Lab



Appendix C - MLP Model Trained in Assignment 1