OSDA Big Homework

Neural FCA

Gorbunov E.I.

Faculty of Computer Science
Higher School of Economics

December 3, 2024

Dataset

Used Dataset: Credit Score Classification Dataset with 164 objects, 7 attribute columns (2 numerical, 5 categorical) and 1 target column.
The main goal is to determine which credit score a client has (High/Average/Low). But I have decided to determine whether a client has a high score.

Dataset

Attributes and values:

- *Age* (25 - 53 years)
- *Gender* (Female/Male)
- *Income* (25000$ - 162500$)
- *Education* ("Bachelor's Degree", "Master's Degree", 'Doctorate', 'High School Diploma', "Associate's Degree")
- *Martial Status* (Single/Married)
- *Number of Children* (0, 1, 2, 3)
- *Home Ownership*(Rented/Owned)

Target for classification:

- *Credit Score* (High/Average/Low)
- *is Not High* - (False - High, True - Average/Low)

Feature Engineering

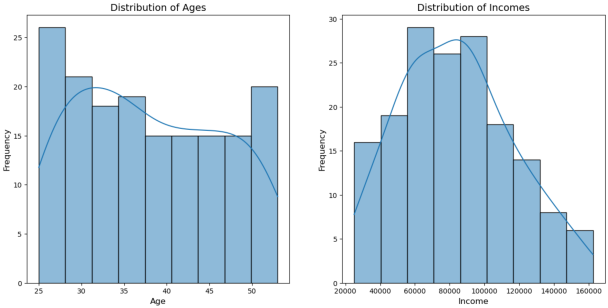


Figure: Distributions of Age/Income

- *Age Group* (”25-33” /”34-43” /”44-53”)
- *Income Group* (”25k-74k” /”75k-124k” /”125k-174k”)
- *Has Children* (False - 0, True - 1/2/3)

Binarization Stategy / Prediction Quality Measure

Scales:

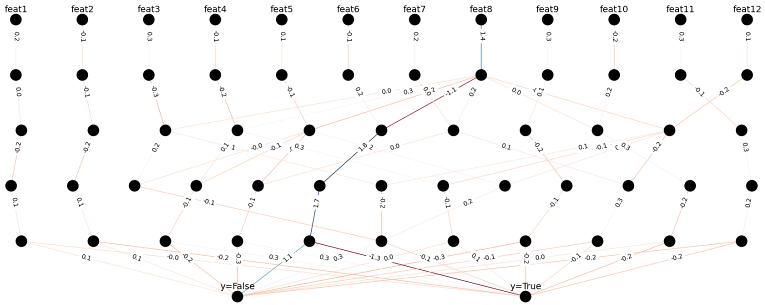
- *Age Group*, *Income Group*, *Education* - nominal scale
- *Gender*, *Martial Status*, *Home Ownership*, *Has Children* - dichotomic scale

After binarization there are 12 binary attributes.

The dataset has 113 samples, which have a high credit score, and 51 samples, which have a low/average credit score. So, the most preferable metric for this date would be **F1 score**.

$$F1 = \frac{2 * precision * recall}{precision + recall}$$

First Results


Neural network with fitted edge weights

16 best concepts give **F1 score** ≈ 0.962 on train set and **F1 score** ≈ 0.88 on test set. This CN was train on **2000 epochs**.

Comparison with Standard Models

Model	DT	RF	KNN	LR	CB	XGB	**CN**
F1 Train	0.987	0.987	0.987	0.987	0.987	0.987	**0.926**
F1 Test	0.917	0.917	0.880	0.917	0.917	0.917	**0.880**

Table: Comparison

Modifications

Engineering of *Age*/*Income*

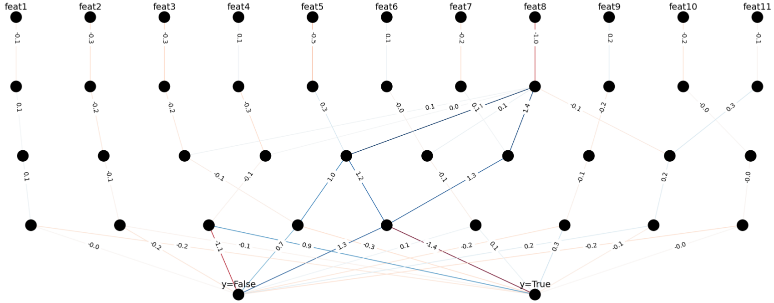
- *Age Group* ("25-40"/"40-53")
- *Income Group* ("25k-59k"/"60k-99k"/"100k-174k")

Scales:

- *Income Group*, *Education* - nominal scale
- *Age Group*, *Gender*, *Martial Status*, *Home Ownership*, *Has Children* - dichotomic scale

After binarization there are 11 binary attributes.

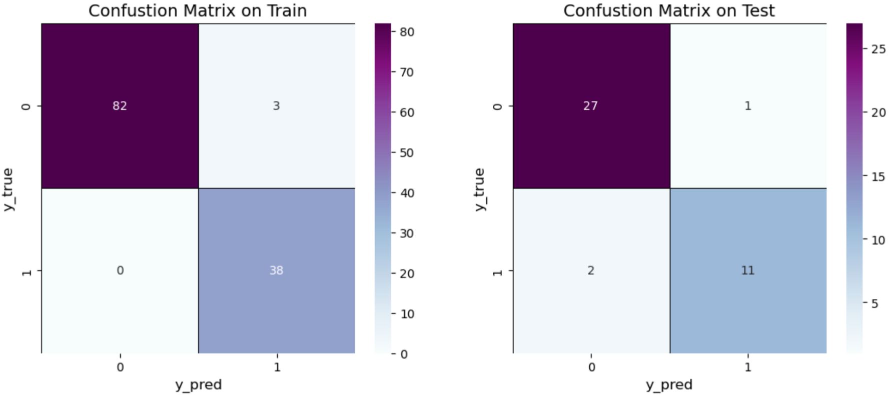
Modifications



Neural network with fitted edge weights

9 best concepts give **F1 score** ≈ 0.962 on train set and **F1 score** ≈ 0.88 on test set. This CN was train on **2000 epochs**.

Confusion Matrix on CN



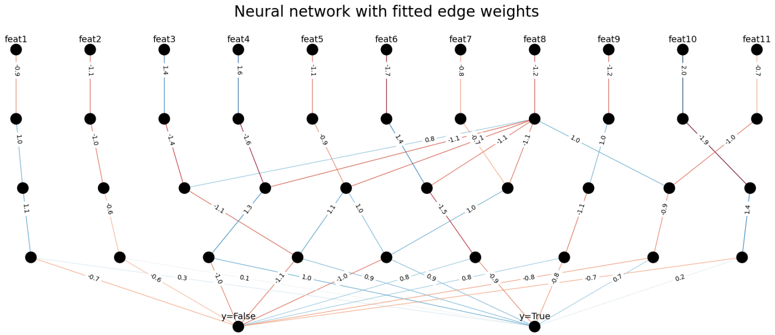
Confustion Matrix on Train

y_true \ y_pred	0	1
0	82	3
1	0	38

Confustion Matrix on Test

y_true \ y_pred	0	1
0	27	1
1	2	11

Modifications with Nonlinearities


Neural network with fitted edge weights

9 best concepts and **tanh** give **F1 score** ≈ 0.987 on train set and **F1 score** ≈ 0.88 on test set. This CN was train on **2000 epochs**.

Conclusion

- **Performance:** Great but Decision Tree, Random Forest, Logistic Regression, CatBoost, XGBoost show better results
- **Interpretability:** Decision Tree provides better results and is easily interpretable.

References

1. Github repository of this project
2. Task formulation
3. Dataset on Kaggle