Explainability study of Feed Forward NN models– CS221, Fall 2019

Ksenia Ponomareva, kp260@stanford.edu

Problem Statement

Neural networks are gaining popularity in variety of applications, including the quantitative finance, [1]. However, due to these models' complexity, their predictions are often difficult to explain and validate. Simpler techniques such as linear or logistic regressions are well understood and controlled by model developers and analysts. Neural networks, in contrast, have many hidden layers and neurons, the exact roles of which are not easily understood by humans.

The primary focus of this project is to examine the feasibility of several methods (relevance and sensitivity analyses) in interpreting and explaining decisions made by credit/default risk neural-network-based models in the context of a credit card portfolio.

Social Impact

The social impact of human interpretability in credit risk estimation is huge.

Customers have a fundamental right to be fairly treated. This means that any lender's decision on a credit application must be explainable. This applies to rejection and acceptance decisions alike: in the first case, the customer has the right to know if any discrimination (e.g. decisions based on ethnicity, gender, etc.) took place, whilst in the second case the customer should be reassured that the bank did not blindly accept a credit application that is likely to end up in heavy financial distress for the applicant – the so-called concept of credit affordability.

Dataset and Features

A publicly available dataset from UCI machine learning repository, [2], is used. This data provides information on default payments, demographic factors, credit data, history of payment and bill statements of credit card clients in Taiwan from April 2005 to September 2005.

The input used for the project has 23 features. There are four demographic features covering gender, education, marital status and age of each client. These are followed by 18 features providing history of payment and bill statements, i.e. repayment status as well as amount of bill statements and previous payments for the six consecutive months. The last feature is the amount of credit given to the client.

The ground truth, or the true label, output takes two values, zero for no default and one if the client defaults on the next month's payment.

Models

Feed-forward neural network (FFNN) model has five hidden layers with 100, 50 and 10 nodes respectively. All activations are ReLU apart from the final one, which is sigmoid. Dropout has been used for each layer to prevent overfitting.

Logistic regression (LR) model has a single node with a sigmoid activation function.

Random forest (RFC) model generates many individual decision trees. For each tree, number of variables considered at each split is limited to a random subset. This helps reduce model variance and de-correlate the trees.

Conclusions

The goal was to show feasibility of these techniques/tools to interpret the outcomes of a neural network. These interpretability techniques could be invaluable in all other banking sectors that may benefit from the use of neural network models.

The hope is that, after being largely restricted based on interpretability grounds, banks and financial institutions may finally be able to consider the development of neural network models on a much broader scale than before.

Relevance Analysis

The most cited relevance analysis technique applied to deep learning models is the layer-wise relevance propagation (LRP) method, introduced in [3]. LRP makes use of deep Taylor decomposition. The aim of this approach is to assign a relevance measure to each input feature of the neural network. This, in turn, represents the direct contribution of each input feature to the final outcome of the neural network.

LRP framework, has been modified and used to obtain relevance for FFNN and LG. Feature importance for these models is based on the percentage contributions to the sigmoid score.
RFC model ranks features according to how much each feature decreases the weighted impurity in a decision tree (averaged across all the decision tress in the model).

Feature importance for the three models is captured in the Figure 1. Since these values add up to one, this allows a straightforward comparison across models. The main focus is on three metrics, summarized in the Table 2: feature ranking, variation and utilization.

Sensitivity Analysis

Sensitivity analysis of the neural network examines which change in input feature the output is the most sensitive to. One way to analyze which input feature makes most impact on the final output is to examine occlusion sensitivity. This method has been widely utilized in image classification approaches, see [4] for details. For this approach, different portions of the input image are systematically occluded with a grey square and output of the model is monitored. If a particular occlusion had caused network output to change dramatically, it would indicate the portion of the image was important for the classification.

In this project, the occlusion sensitivity method is adapted for the credit card payment performance example. Here instead of the grey square, each input feature is in turn replaced by the average in the dataset and a corresponding output is generated. Resulting sigmoid scores, one for each feature changed, are then plotted in a heatmap, see Figure 2.

Profile Analysis

Furthermore, we consider modified unary quantitative input influence QII $= S_{orig} - S(x_i)$, the difference between the original sigmoid score and the one resulting in the occlusion of a feature, x_i. Following the approach used in [1], modified QII's are used as a distance metric for clustering to identify profile groups in the model. Three main profile clusters for the FFNN model are shown in Figure 3.

For example, one of the main clusters can be described by individuals who are graduate school-level educated women in their early 30s with a large limit balance, who have been paying their bills on time in the last six months. For these obligors marital status is not an important feature, indicating a mixture of married and single candidates.

Models	Accuracy	Recall	Precision	F1	AUC	Gini Coefficient
FFNN	**82%**	31%	**71%**	43%	77.10%	**54.30%**
LR	81%	23%	68%	34%	**72.00%**	44.00%
RFC	81%	**36%**	64%	**46%**	65.00%	30.00%

Table 1 Performance metrics for the models.

Models	Rank 1	Rank 2	Rank 3	Variation	Utilization
FFNN	PAY_2	PAY_1	LIMIT_BAL	0.77%	Full
LR	GENDER	PAY_AMT2	PAY_1	4.05%	EDUCATION, AGE, BILL_AMT_6, PAY_AMT_3 underutilized
RFC	PAY_1	AGE	LIMIT_BAL	1.93%	Full

Table 2 Performance metrics for the models.

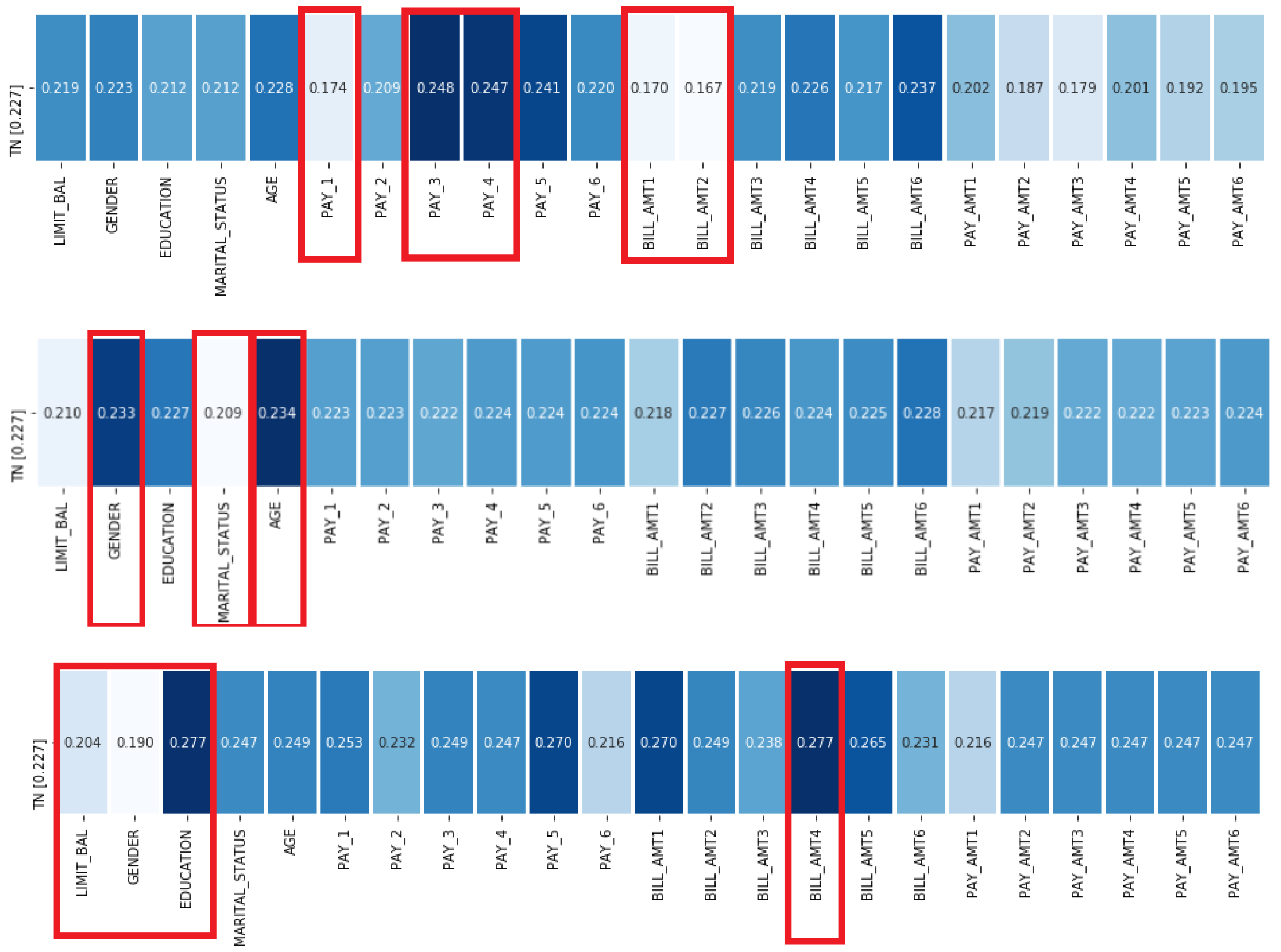

Figure 2 Sensitivity analysis for a TN candidate for FFNN, LR and RFC models respectively.

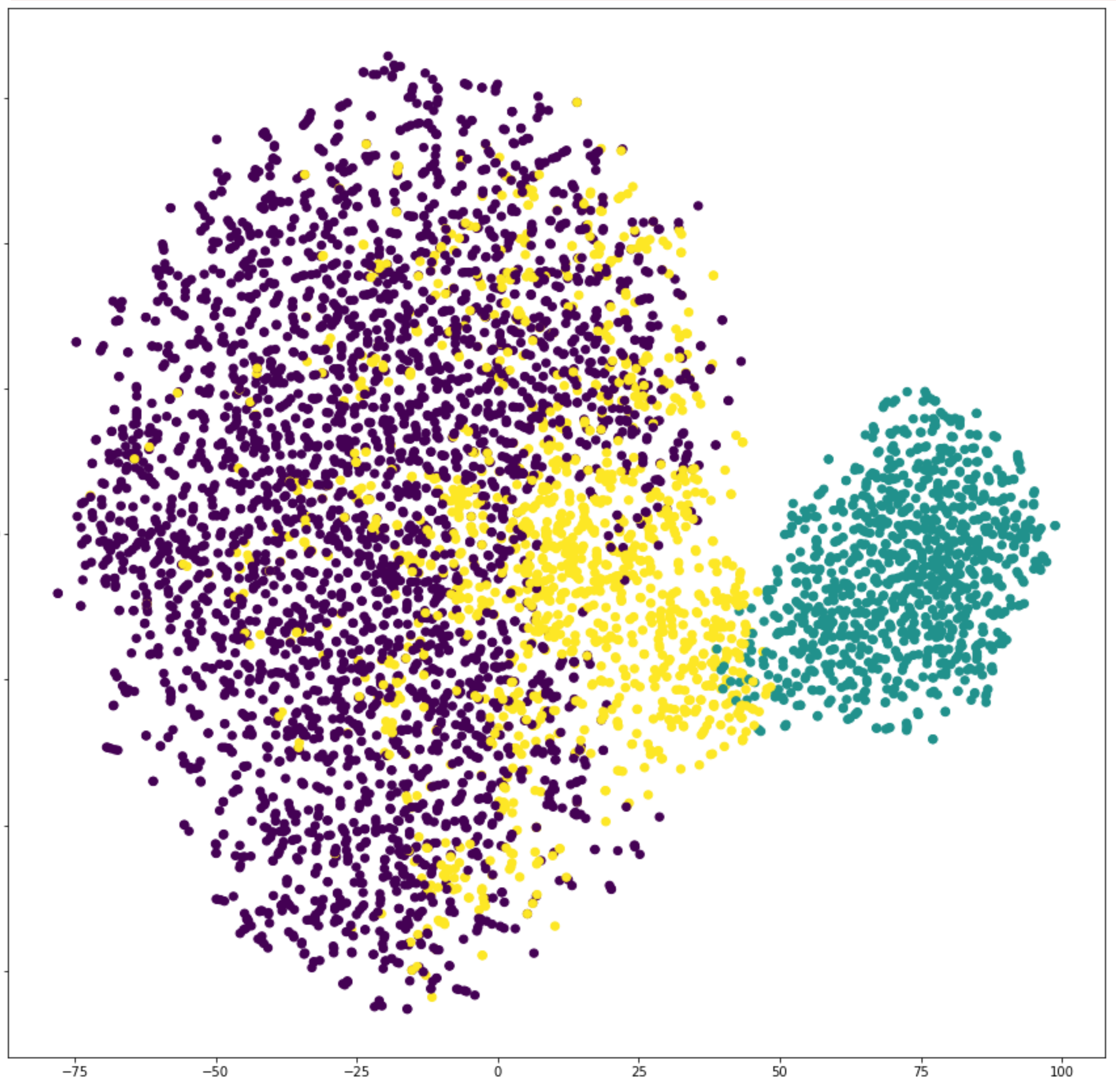

Figure 3 TSNE plot of three main modified QII based profiles in FFNN.

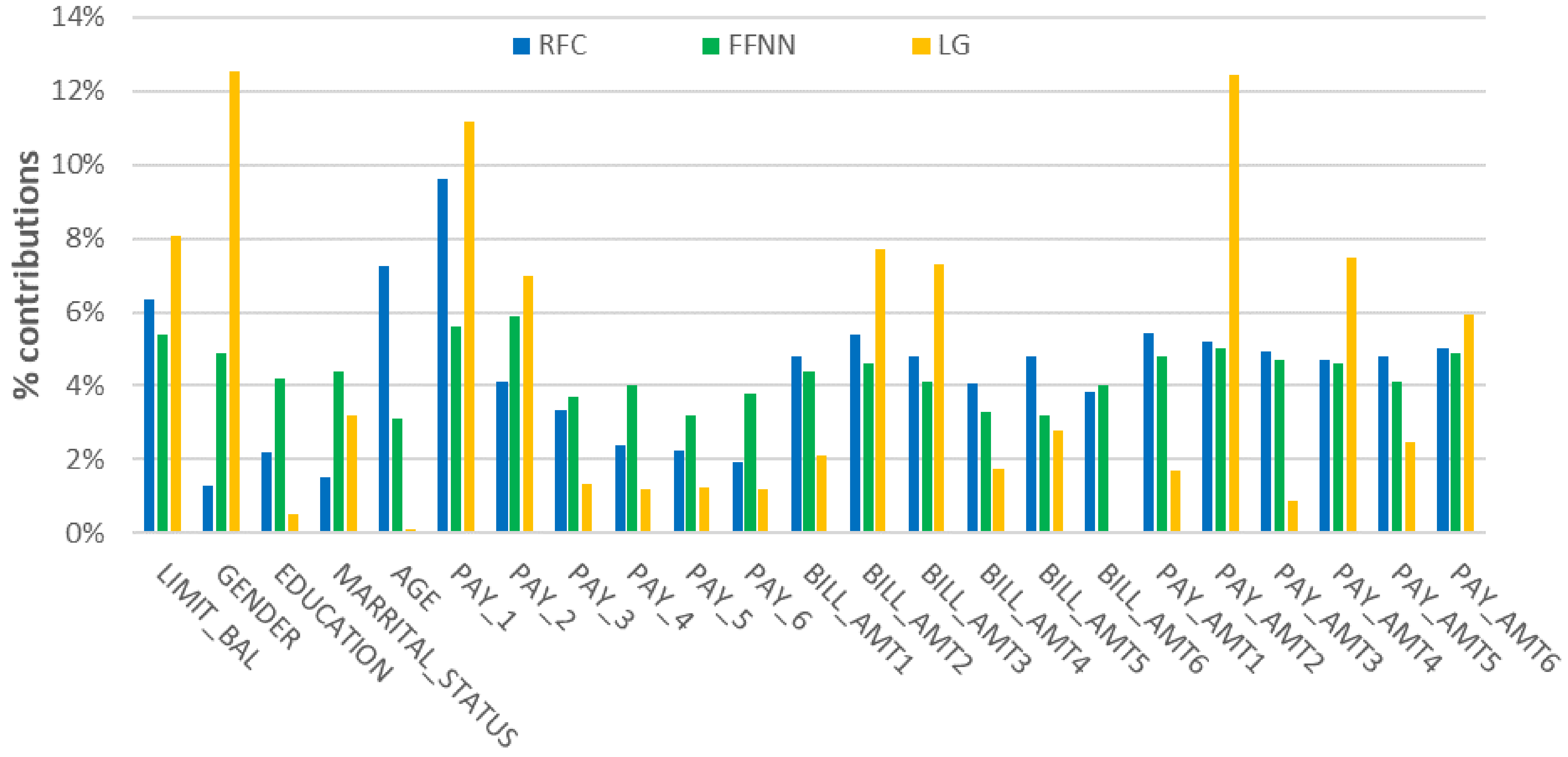

Figure 1 Features' importance across the models.

[1] Bracke, P., Datta, A., Jung, C., Sen, S., (2019) Machine learning explainability in finance: an application to default risk analysis, Staff Working Paper No. 81, Bank of England.

[2] Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[3] Montavon, G., Bach, S., Binder, A., Samek, W., Muller, K. (2017). Explaining NonLinear Classification Decisions with Deep Taylor Decomposition, Pattern Recognition, 65, 211-222.

[4] Zeiler M.D., Fergus R. (2014). Visualizing and Understanding Convolutional Networks. ECCV.

[5] Poster video link: https://youtu.be/hdCaVNb5K7M