

# Análise de Credit Score

Trabalho Final - Digital House Coding School

Modelos de classificação - Credit Score

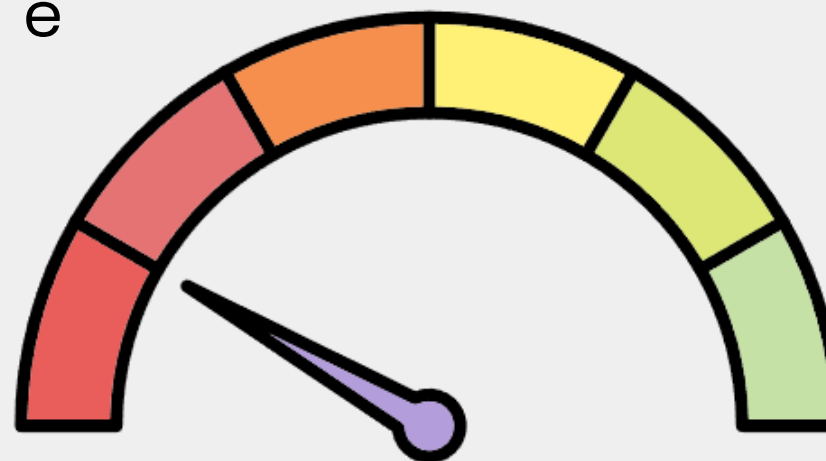
Grupo 1 - Augusto Ribeiro, Eduardo Henrique,  
Felipe Ferraz, Lucas Ribeiro, Thiago Ferro, Inês Dantas

# O que é o credit score?

Credit Score é uma pontuação resultante dos hábitos de pagamento e relacionamento do consumidor com o mercado de crédito. Fatores como pagar contas em dia, histórico de dívidas negativadas, saldo devedor etc, representam o risco de inadimplência e saúde financeira geral de uma pessoa.

O score vai de 0 a 1.000 pontos. Esses pontos foram divididos em nosso dataset dentre três classes (Classificação multiclasse):

- Poor (ruim) - alto risco de inadimplência;
- Standard (padrão) - médio risco;
- Good (bom) - baixo risco.



# Dados

Os dados escolhidos foram obtidos no Kaggle e são fictícios. O objetivo é construir um sistema inteligente para segregar as pessoas em faixas de pontuação de crédito para reduzir os esforços manuais.

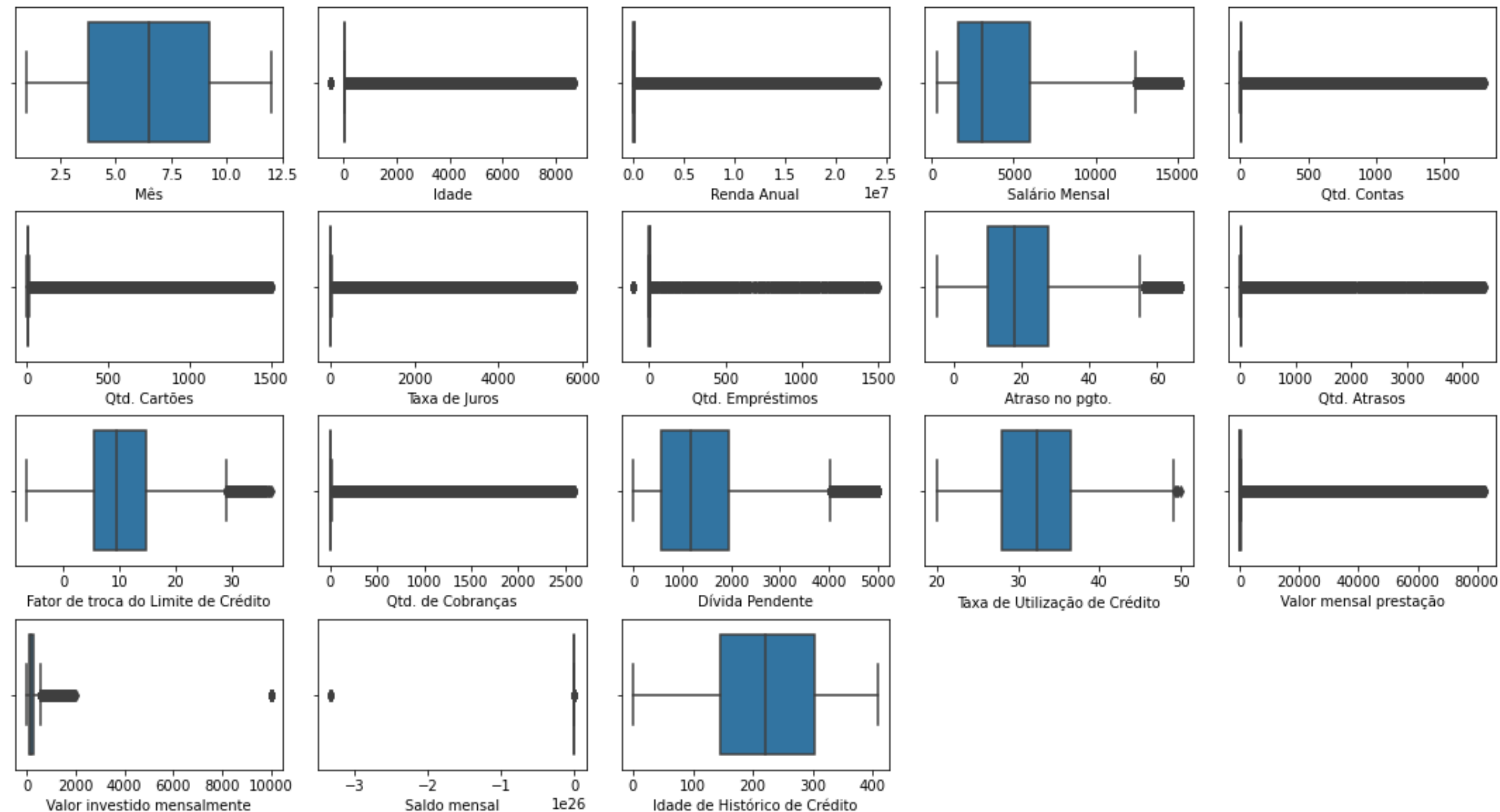
- ID - Identificador único de entrada (observação)
- Customer\_ID - ID único de cliente
- Month - Mês do ano
- Name - nome do cliente
- Age - Idade do cliente
- SSN - Social Security Number (CPF no Brasil)
- Occupation - Ocupação do cliente
- Monthly\_Inhand\_Salary - Salário mensal do cliente
- Num\_Bank\_Accounts - Quantidade de contas em bancos
- Num\_Credit\_Card - Quantidade de cartões de crédito
- Interest\_Rate - Taxa de juros do cartão de crédito
- Num\_of\_Loan - Quantidade de empréstimos feitos no banco
- Type\_of\_Loan - Tipo de empréstimo feito pelo cliente
- Num\_Credit\_Inquiries - Quantidade de "cobranças" no cartão
- Total\_EMI\_per\_month - Pagamento fixo em dólares por mês
- Outstanding\_Debt - Restante da dívida a ser paga
- Delay\_from\_due\_date - Qtd. de dias de atraso no pagamento do cartão
- Changed\_Credit\_Limit - Variação percentual de limite do cartão de crédito
- Credit\_History\_Age - Tempo de histórico de crédito do cliente
- Monthly\_Inhand\_Salary - Salário mensal do cliente
- Credit\_Mix - Qualidade da variedade de crédito do cliente (crédito rotativo e parcelado)
- Num\_of\_Delayed\_Payment - Média de pagamentos atrasados pelo cliente
- Credit\_Utilization\_Ratio - Taxa de utilização do cartão de crédito
- Payment\_of\_Min\_Amount - Se o cliente pagou ou não o mínimo da dívida
- Monthly\_Balance - Saldo mensal do cliente
- Payment\_Behaviour - Comportamento de pagamento do cliente
- **Credit Score - Pontuação da credibilidade do cliente**

# EDA e Data Cleaning



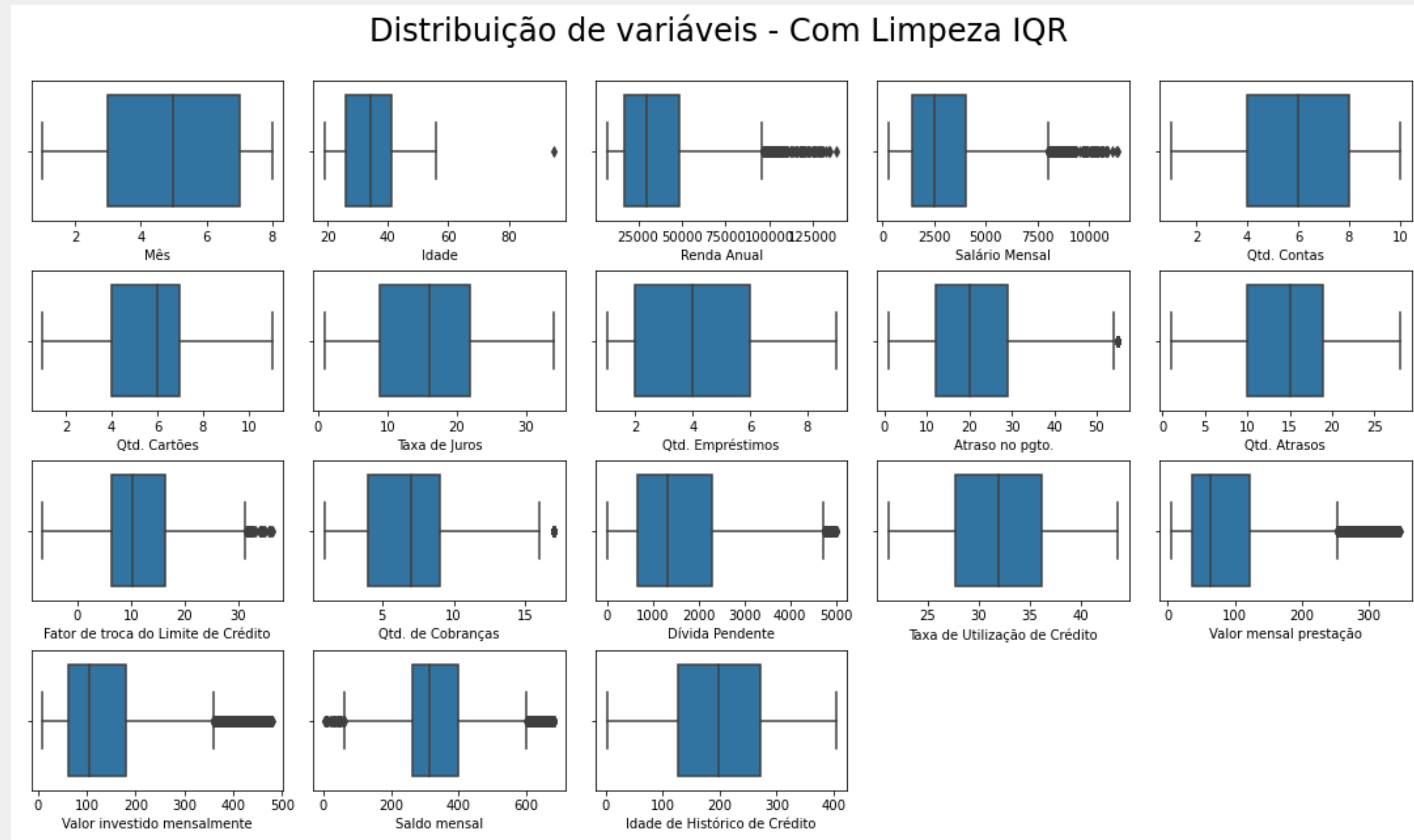
# EDA e Data Cleaning

Distribuição de variáveis - Sem Limpeza IQR



# EDA e Data Cleaning

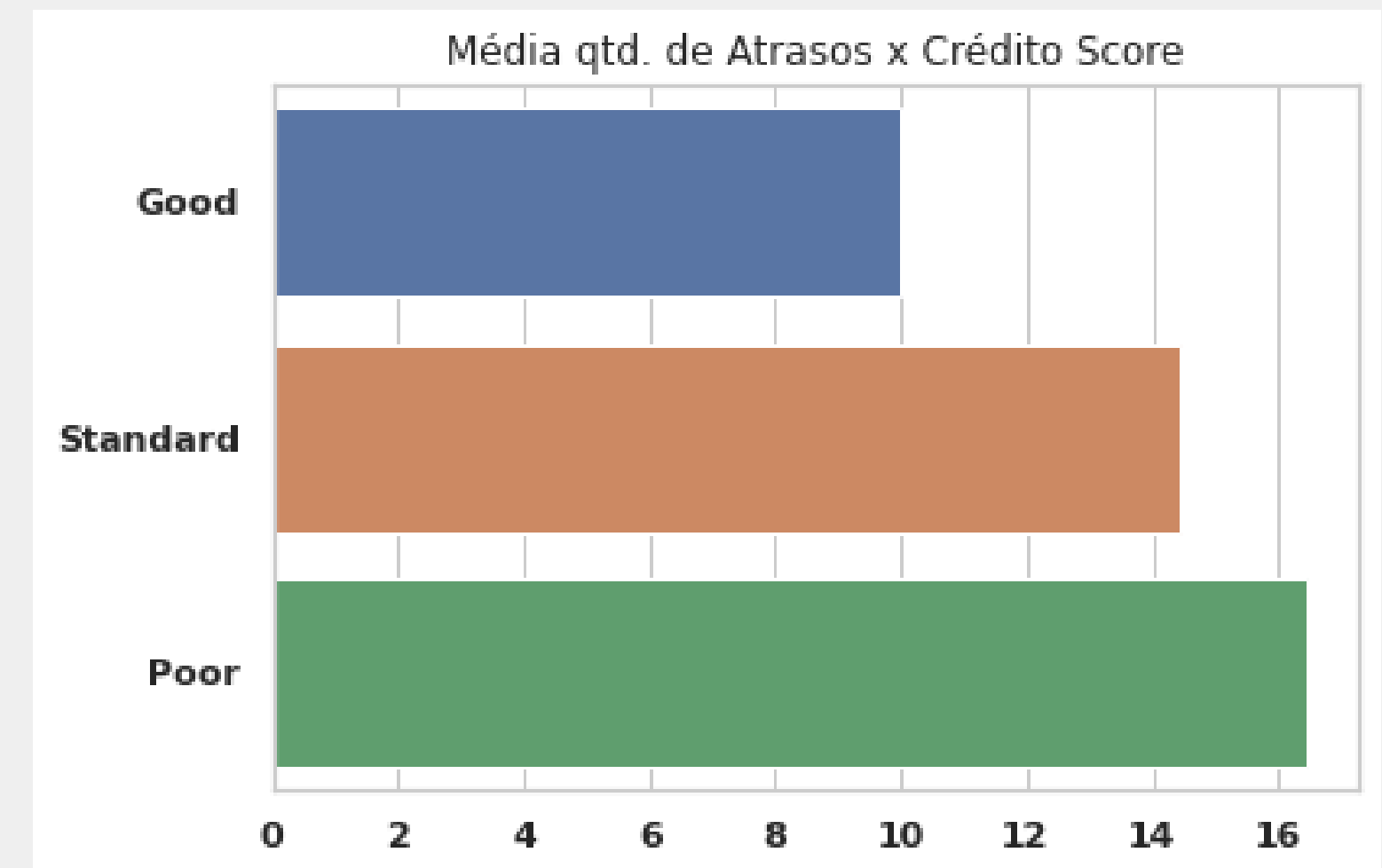
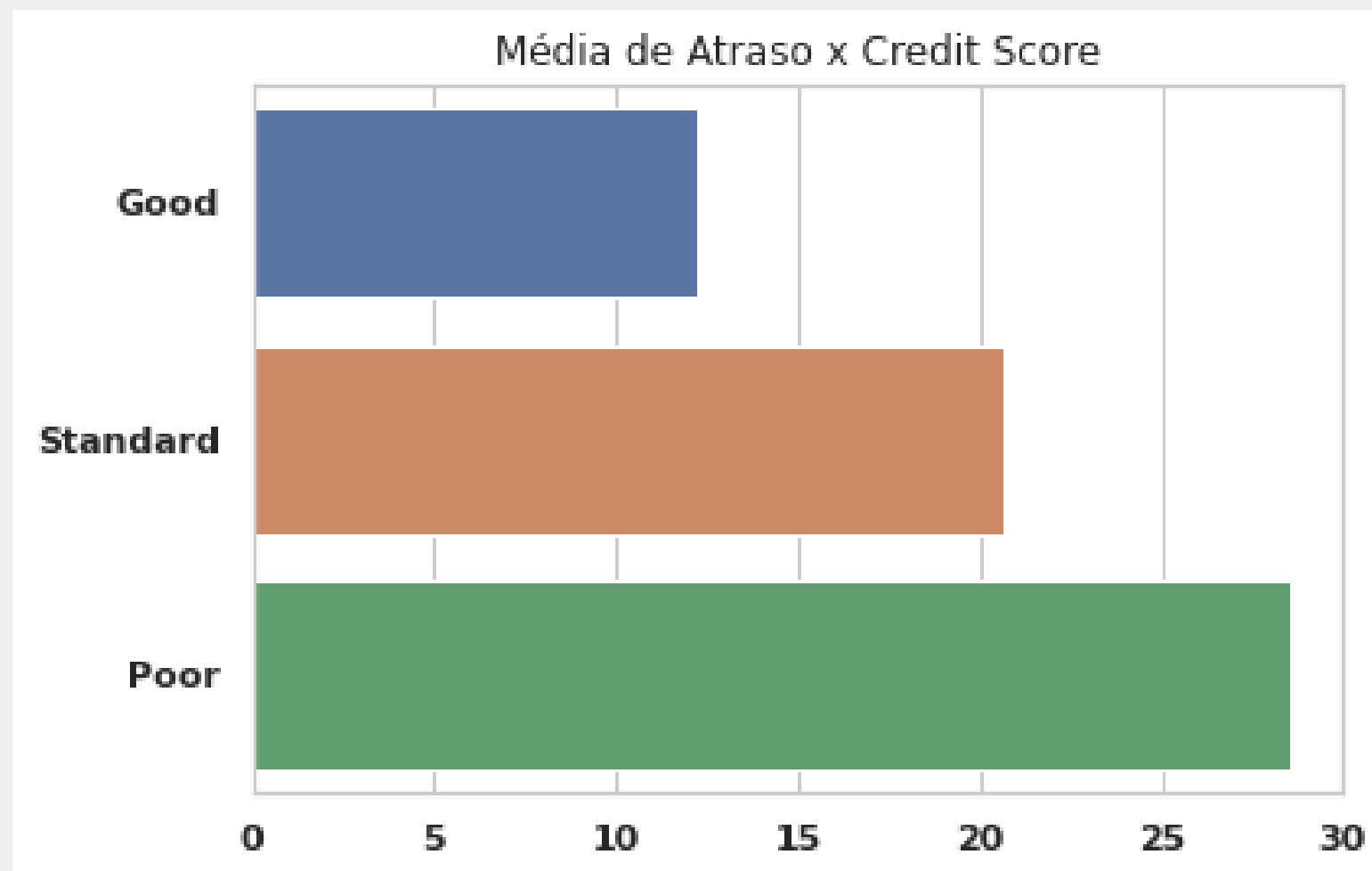
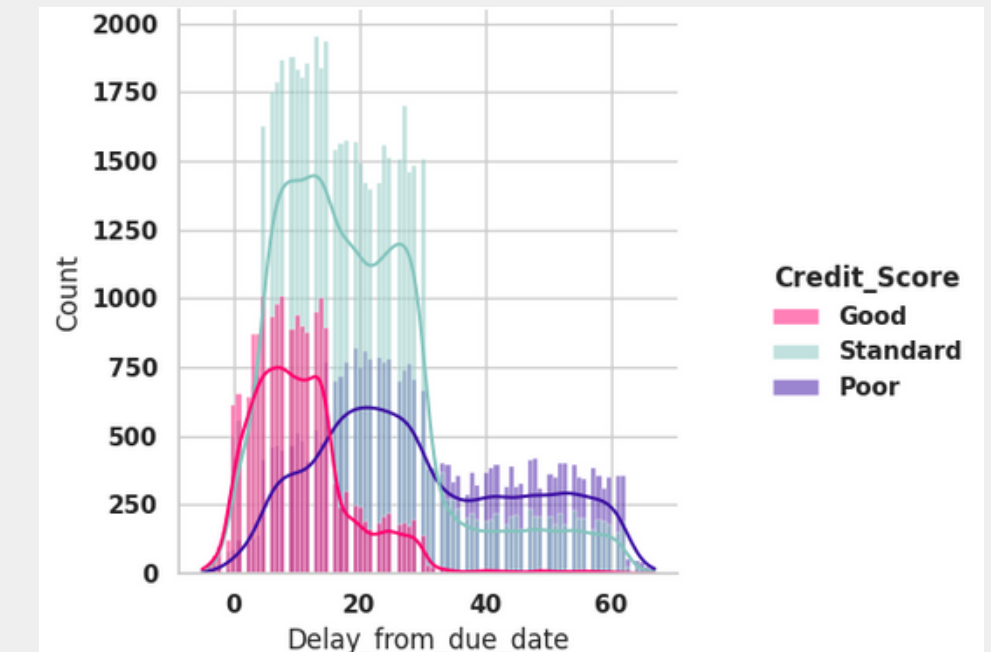
- 50% das idades estão na faixa de [26; 41];
- 75% das pessoas ganham menos de 4 mil;
- Em média, as pessoas tem 6 cartões e 4 empréstimos feitos;
- Em média, as pessoas tem saldo devedor de 1600;
- 50% das pessoas investem mensalmente na faixa de [62; 181].





# EDA e Data Cleaning

- Pessoas com Credit Score mais baixo tem uma média de atraso no pagamento das parcelas maior. Assim como uma maior quantia de atrasos.



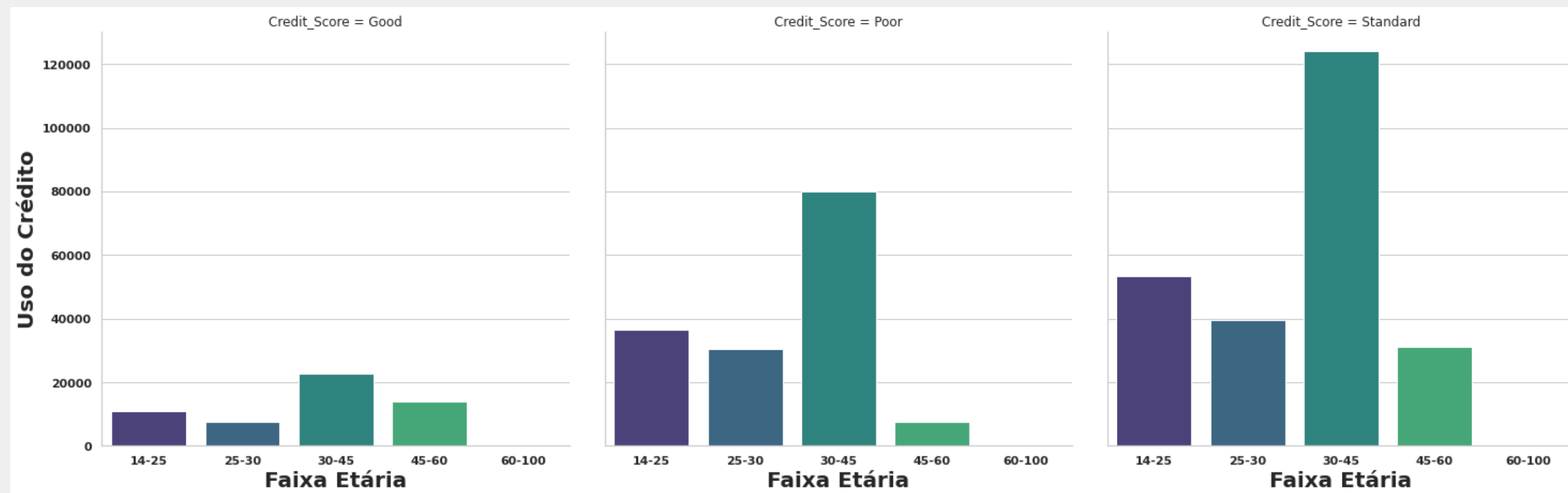
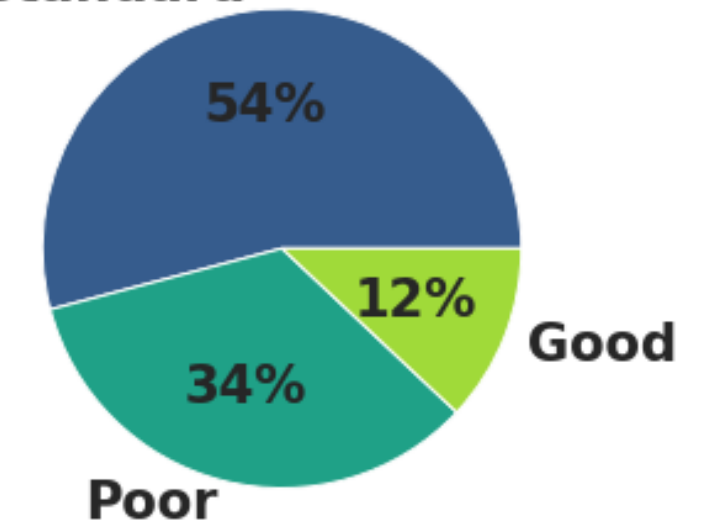
# EDA e Data Cleaning

## Credit Score e Uso do Crédito

- Maioria do Credit Score é Standard (54%), **target desbalanceada**;
- Idades entre 35 e 45 anos usam mais o Cartão de Crédito;

## Distribuição de Score

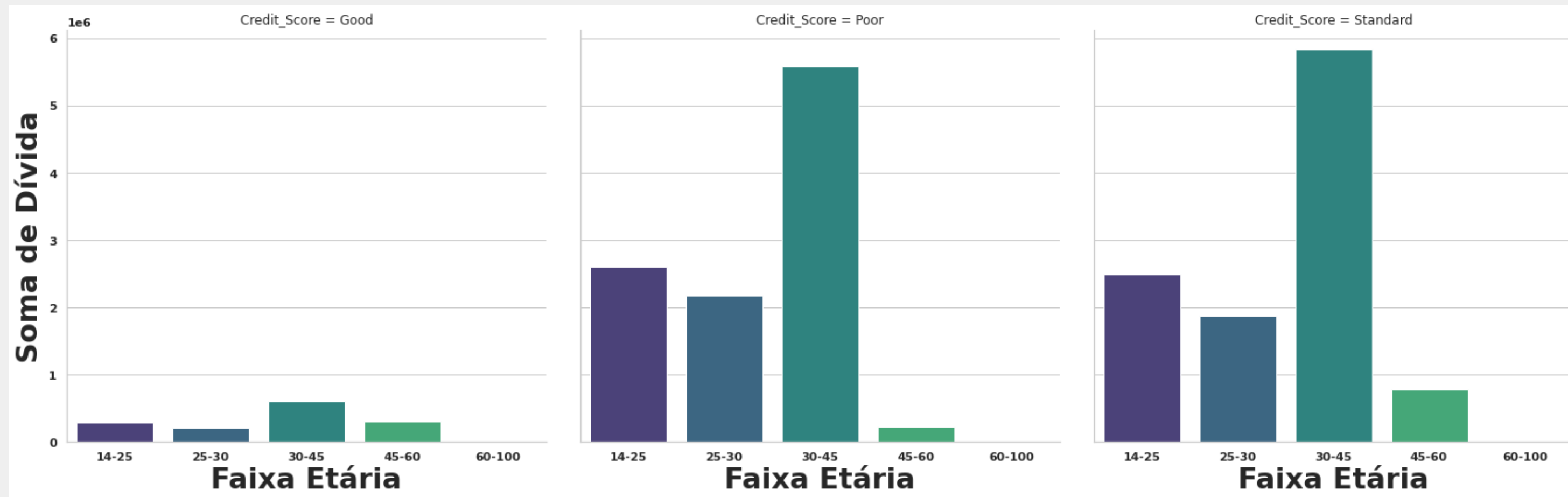
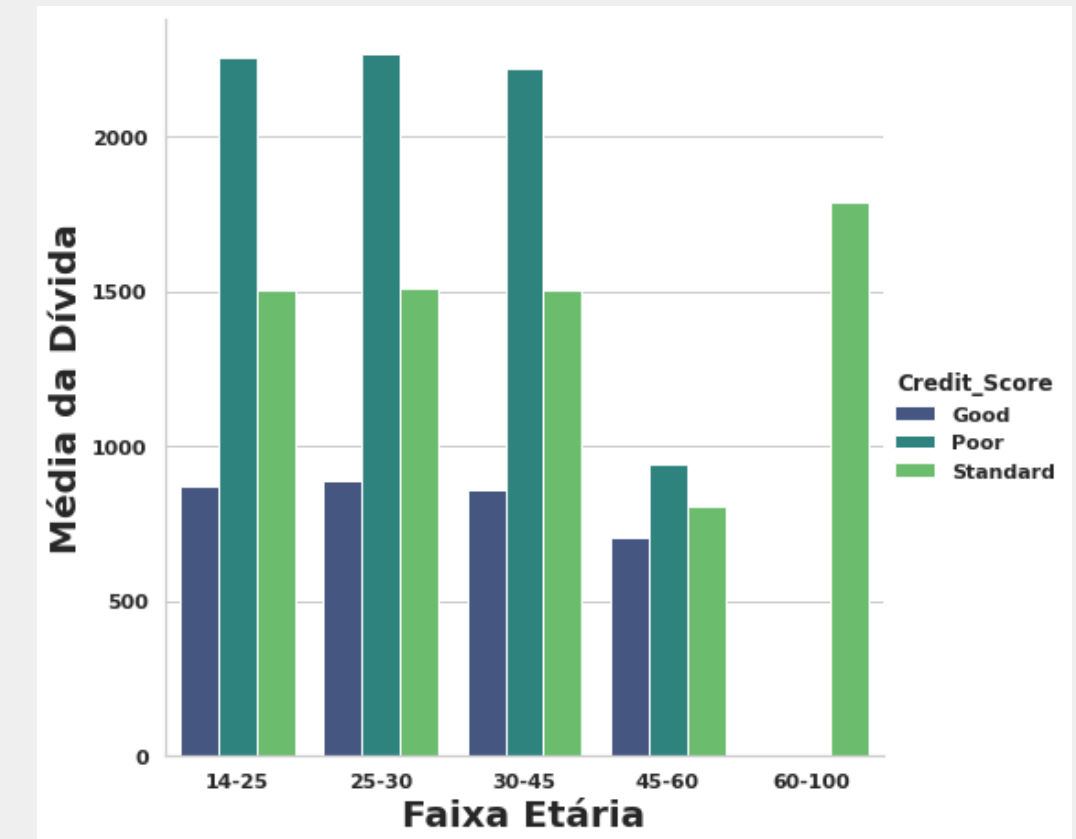
### Standard





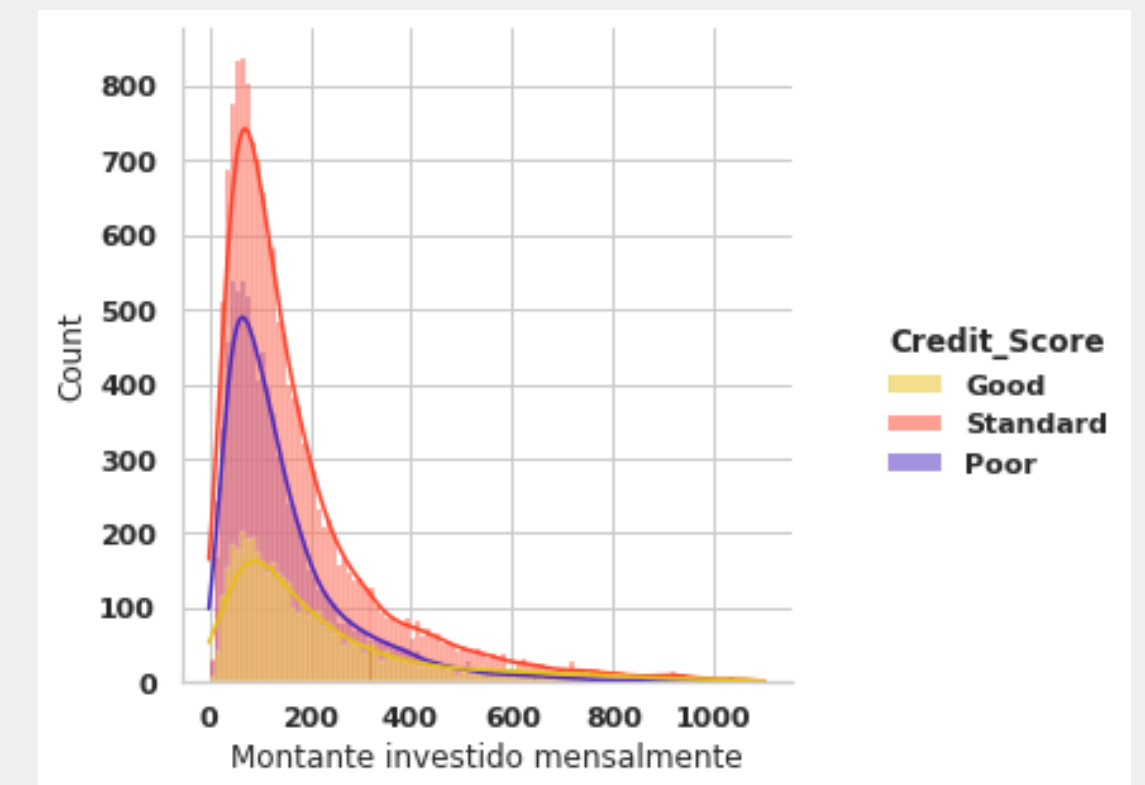
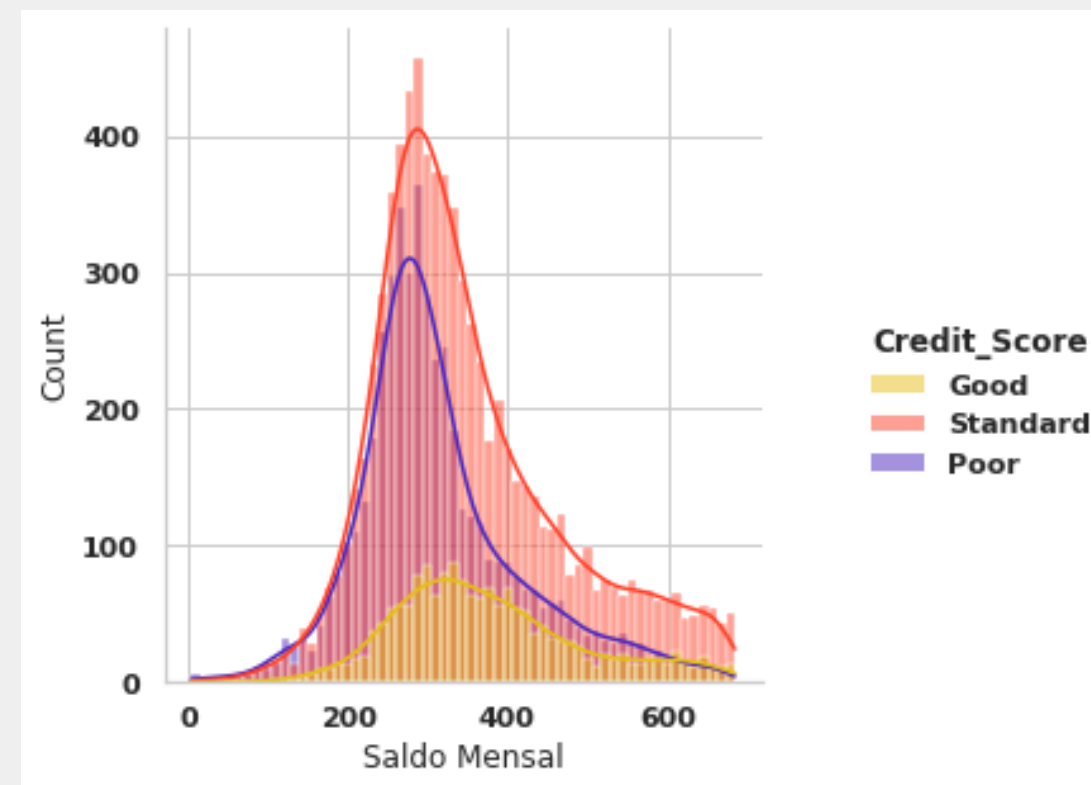
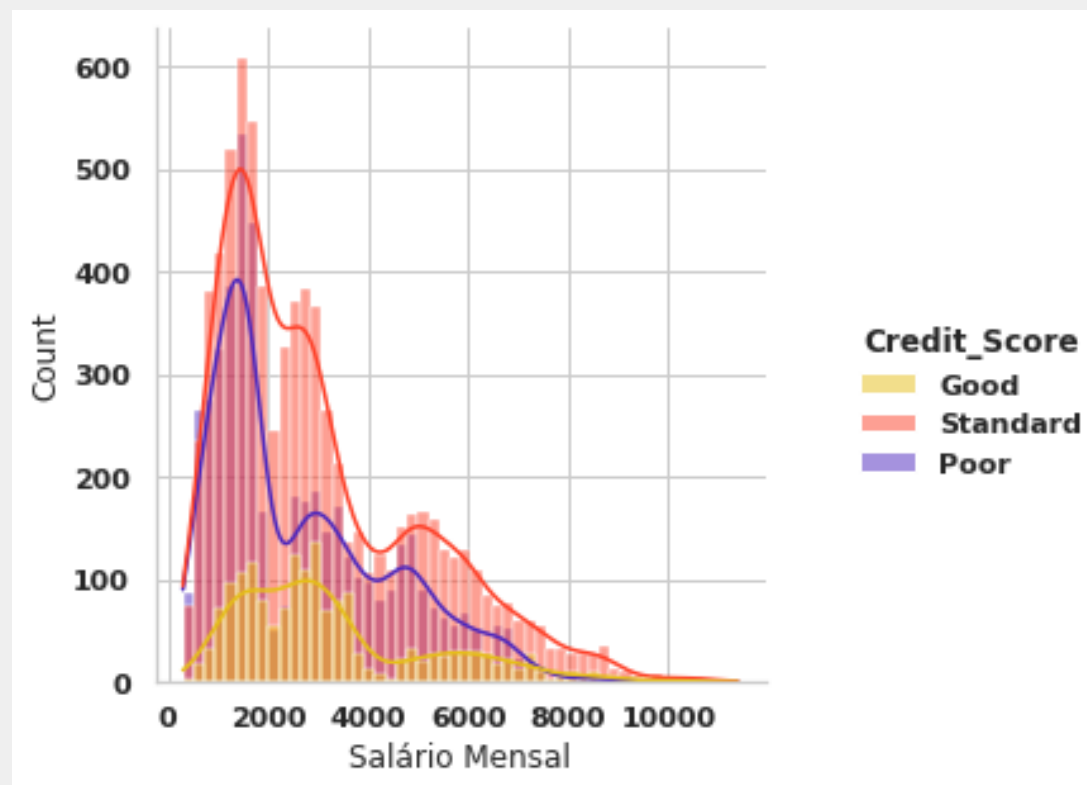
# EDA e Data Cleaning

- Pessoas com alta dívida tendem a ter pior Credit Score;
- Idades entre 25 e 45 anos possuem mais dívida, provavelmente pelo maior uso do crédito.



# EDA e Data Cleaning

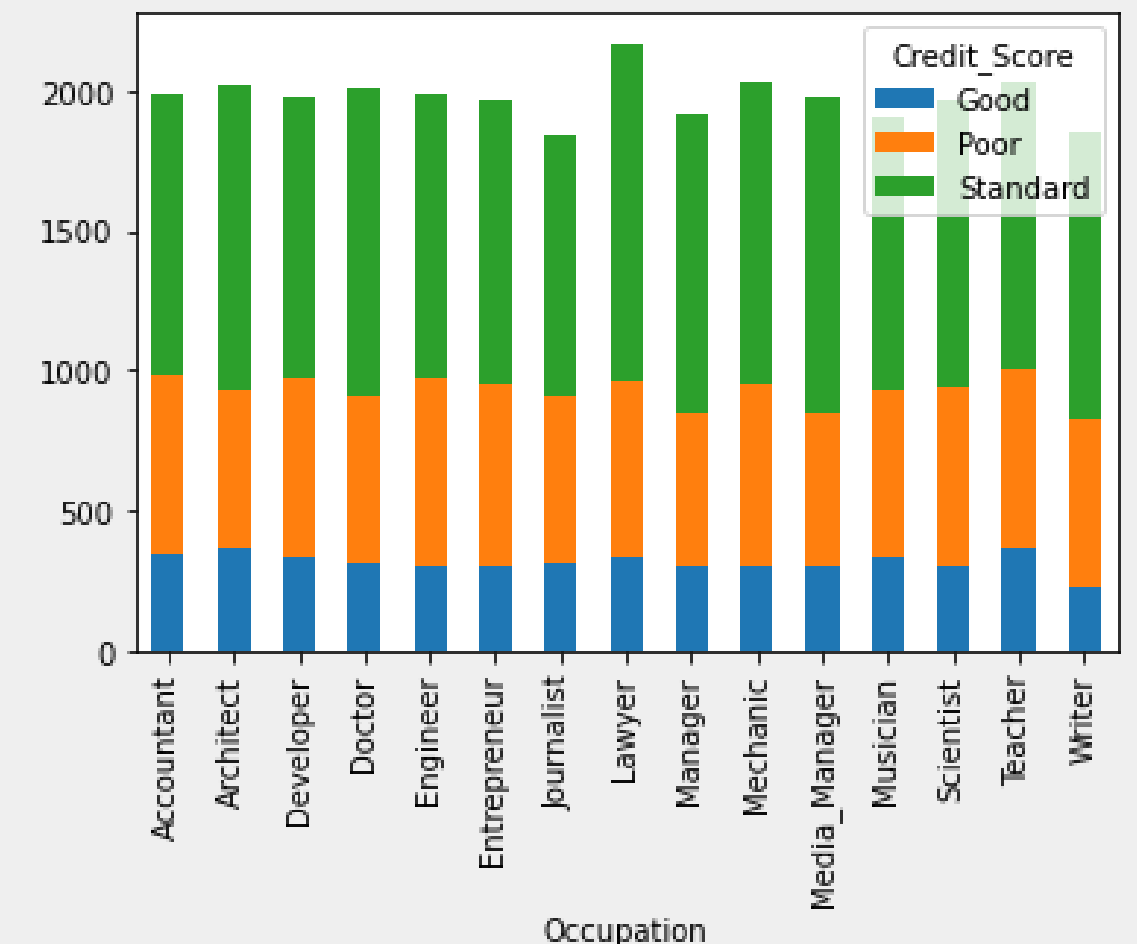
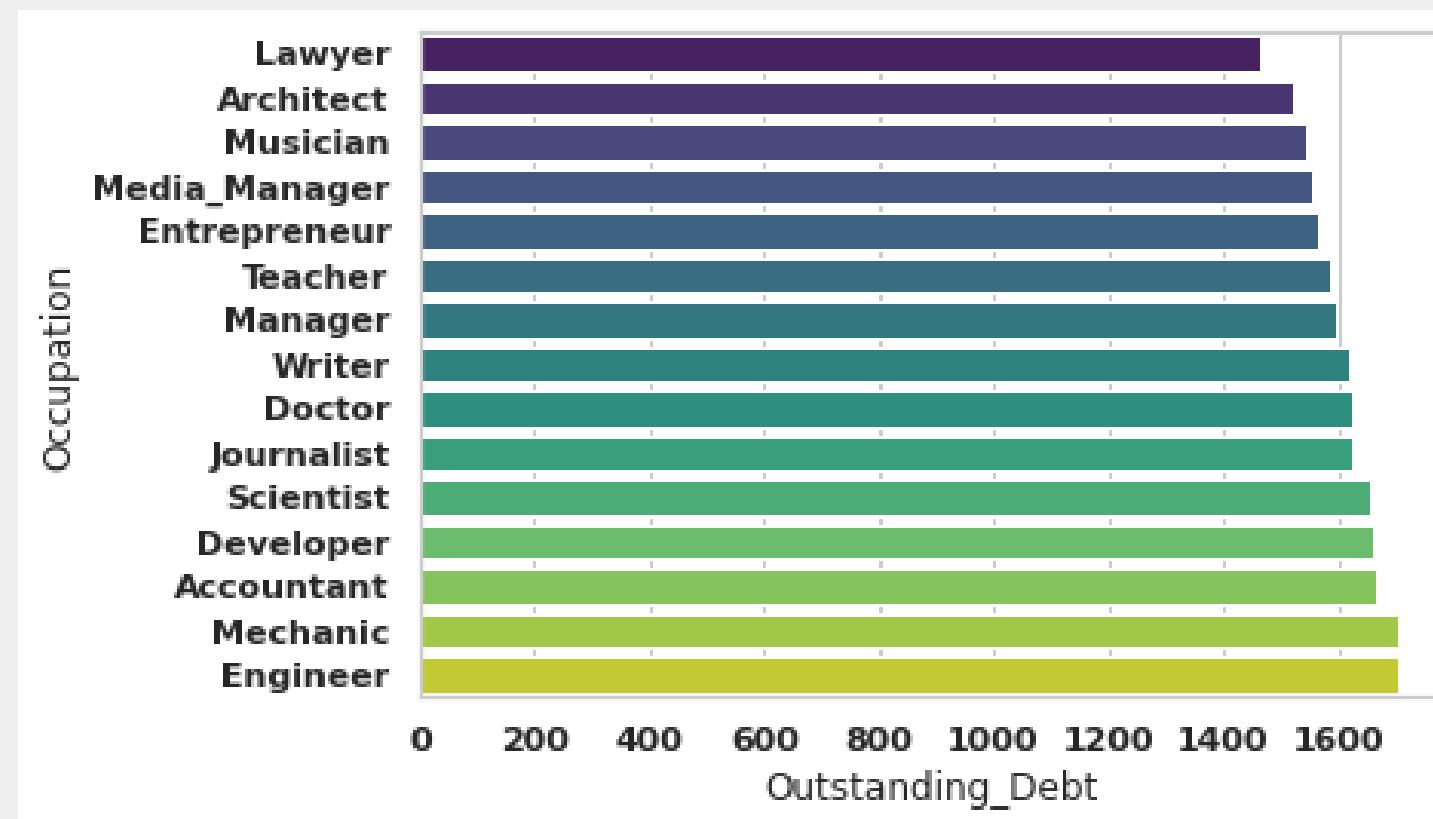
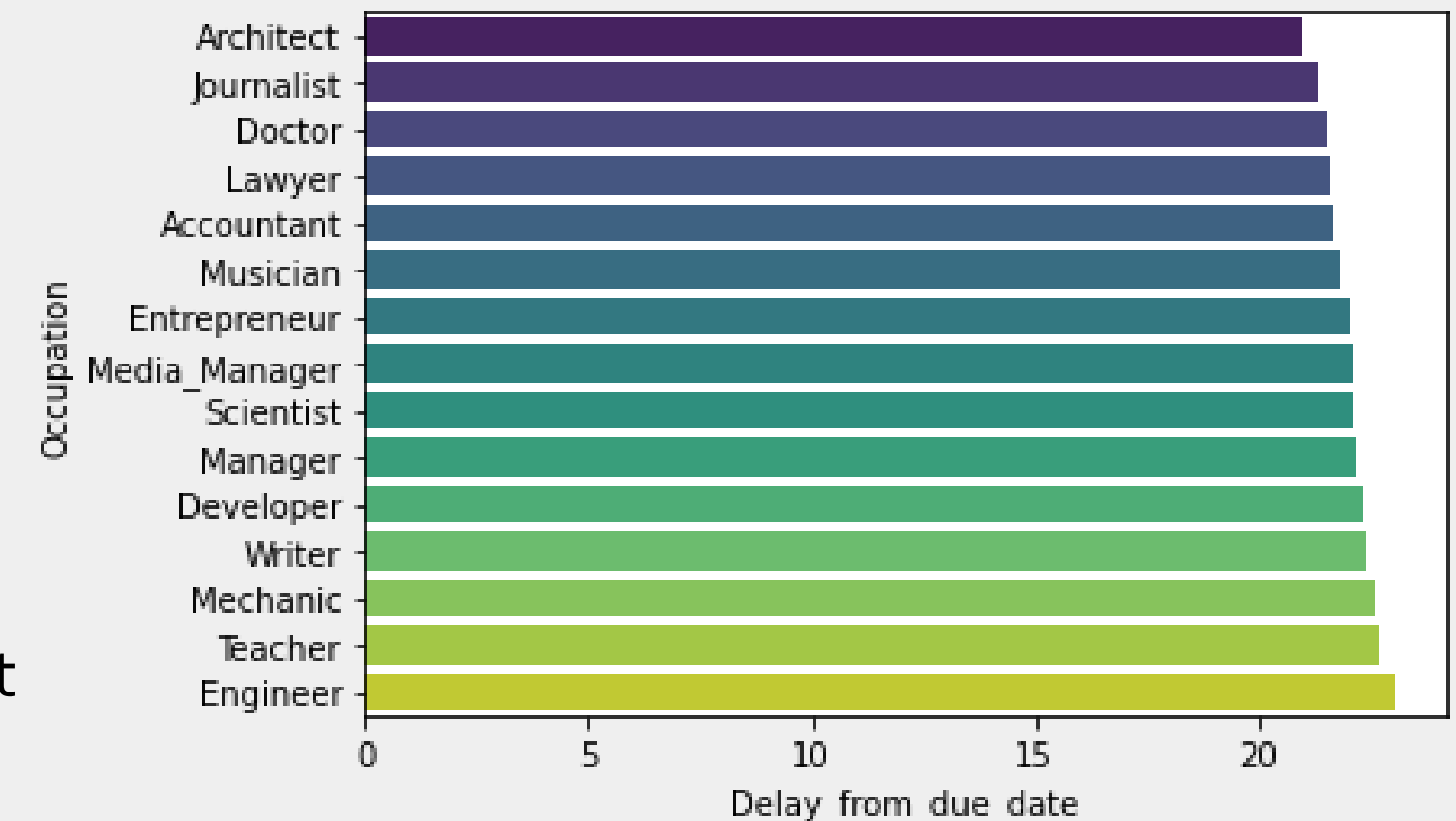
- O quanto a pessoa ganha não implica diretamente no Credit Score, mas sim o quão bem ela gerencia suas finanças;
- Pessoas com bom Credit Score costumam guardar mais dinheiro ao final do mês.



# EDA e Data Cleaning

## Profissão

- Apresenta distribuição uniforme em atrasos e Dívida;
- Acompanham a distribuição desbalanceada de Credit Score, ou seja, não influencia na target.



# Feature Engineering



# Encoding e Normalização

## Catégoricas (pd.DataFrame.factorize())

- Credit\_Mix (mix de crédito, Payment\_of\_Min\_Amount (pagamento do mínimo da dívida)

## Variáveis Dropadas

- Customer\_ID, Name, SSN, Type\_of\_Loan, Type\_of\_Loan\_ajustado, Credit\_History\_Age, Credit\_Score, Month, Monthly\_Balance, Occupation, Payment\_Behaviour

## Numéricas (Standard Scaler)

- Utilizamos o Standard Scaler nas variáveis numéricas que não foram dropadas deixando-as com média 0 e desvio padrão 1

**Target multiclasse → 0 = Good, 1 = Poor, 2 = Standard**

## Factorize de Credit Mix

```
Standard    10882
Good         5781
Bad          5255
Name: Credit_Mix, dtype: int64
```

```
1    10882
0     5781
2     5255
Name: Credit_Mix, dtype: int64
```

# Feature Selection

## Tabela StatsModels Summary

Passam no teste e tem coeficiente relevante:

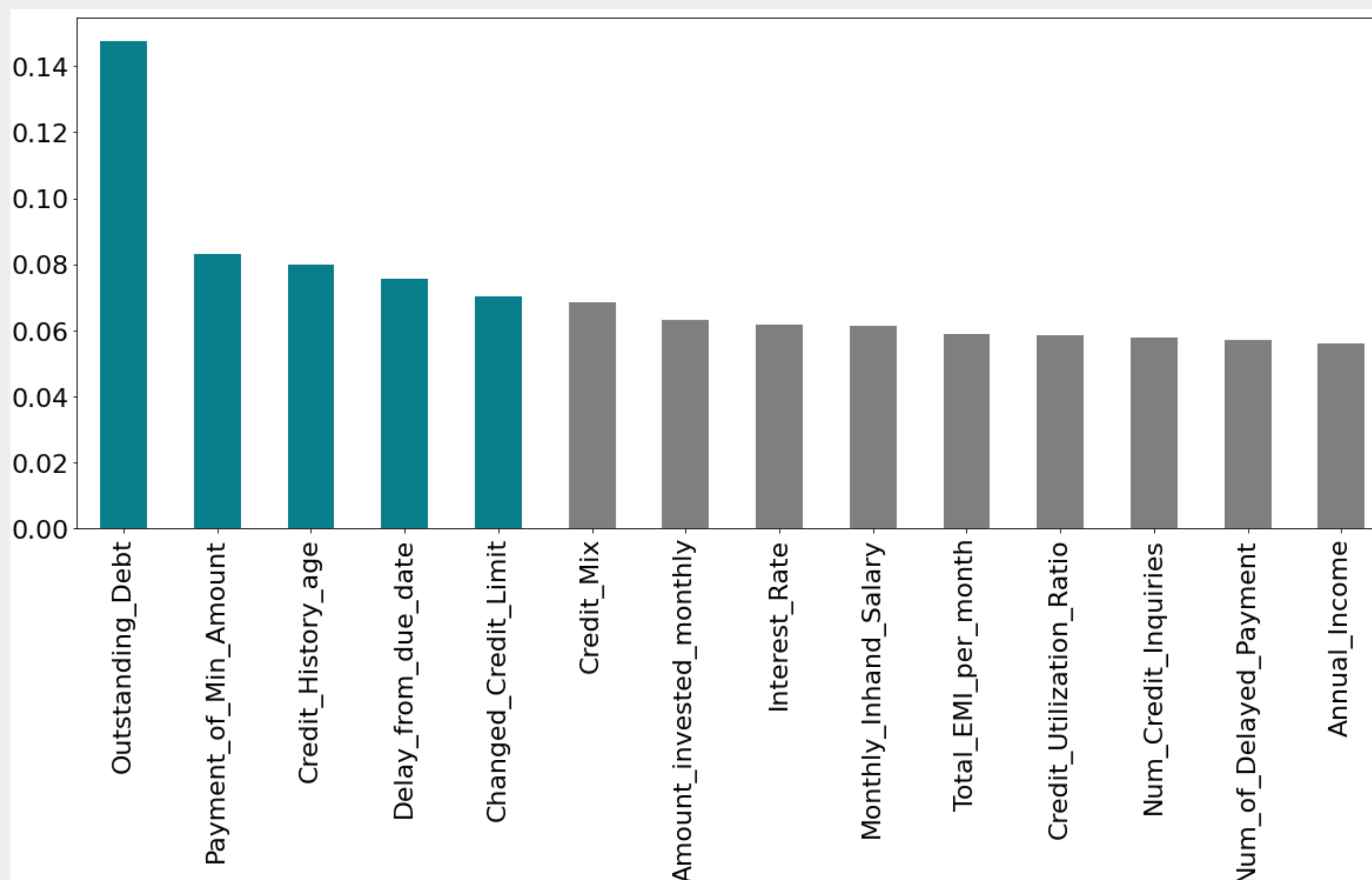
- Salário Mensal (Monthly\_Inhand\_Salary);
- Dias de atraso de pagamento do cartão (Delay\_from\_due\_date);
- Fator de troca de limite do cartão (Changed\_Credit\_Limit);
- Saldo devedor (Outstanding\_Debt);
- Mix de Crédito (Credit\_Mix);
- Tempo de histórico com cartão (Credit\_History\_Age); e
- Pagamento do montante mínimo (Payment\_of\_Min\_Amount).

	coef	std err	t	P> t	[0.025	0.975]
Annual_Income	-0.0015	0.006	-0.255	0.799	-0.013	0.010
Monthly_Inhand_Salary	0.0630	0.006	10.299	0.000	0.051	0.075
Interest_Rate	0.0009	0.006	0.156	0.876	-0.010	0.012
Delay_from_due_date	-0.2421	0.007	-33.600	0.000	-0.256	-0.228
Num_of_Delayed_Payment	-0.0053	0.006	-0.932	0.351	-0.017	0.006
Changed_Credit_Limit	0.0516	0.007	7.516	0.000	0.038	0.065
Num_Credit_Inquiries	-0.0014	0.006	-0.244	0.807	-0.012	0.010
Outstanding_Debt	-0.3321	0.009	-38.266	0.000	-0.349	-0.315
Credit_Utilization_Ratio	0.0104	0.006	1.792	0.073	-0.001	0.022
Total_EMI_per_month	-0.0064	0.006	-1.121	0.262	-0.018	0.005
Amount_invested_monthly	0.0032	0.006	0.554	0.580	-0.008	0.014
Credit_History_age	0.1942	0.008	24.861	0.000	0.179	0.210
Credit_Mix	1.0444	0.013	82.097	0.000	1.019	1.069
Payment_of_Min_Amount	0.3845	0.019	20.404	0.000	0.348	0.421

Testes de Hipóteses Features

# Feature Selection

**ExtraTreesClassifier.feat\_importance\_ (Coef. Gini)**

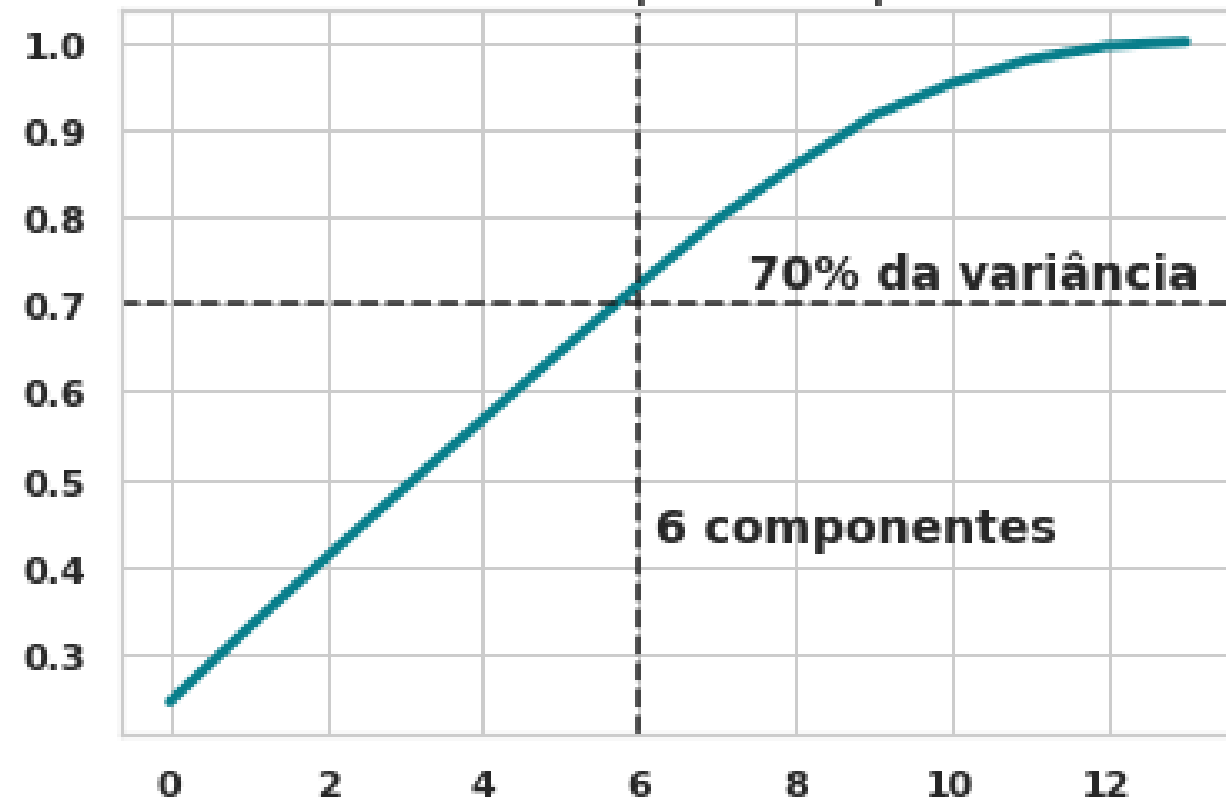




# PCA

- Análise de componentes principais nos permite visualizar quais das variáveis tem maior influência no comportamento dos dados com a target.

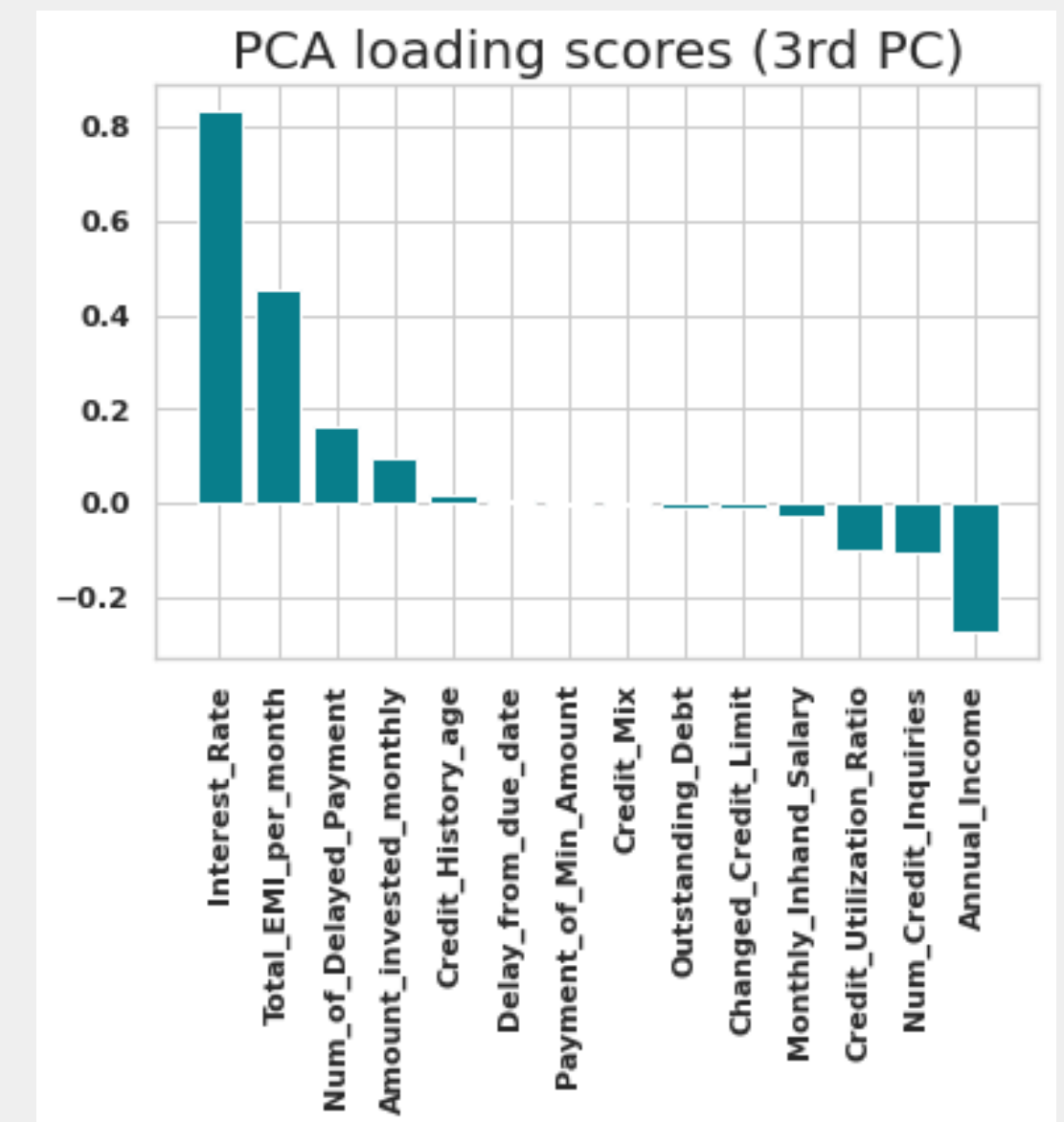
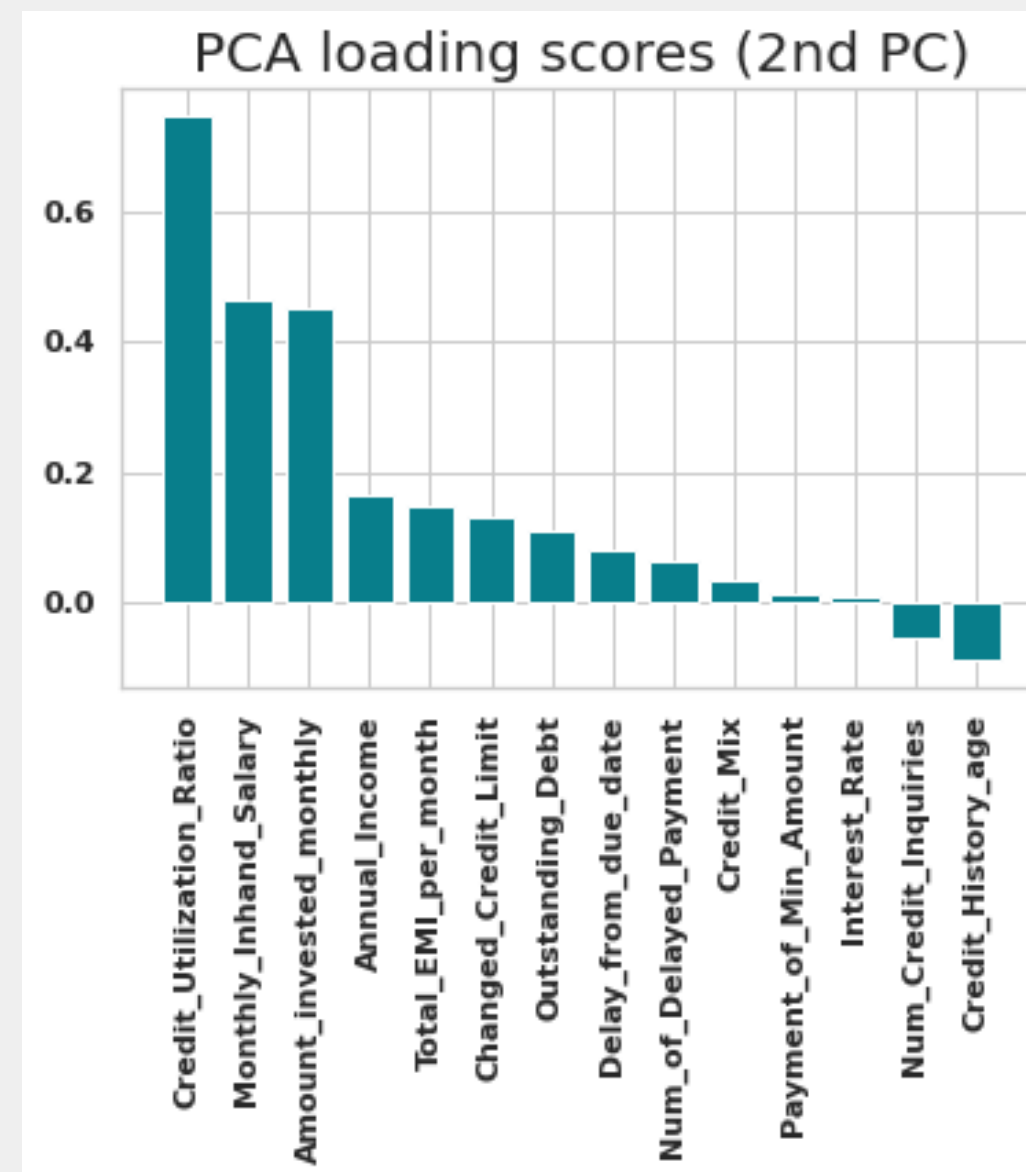
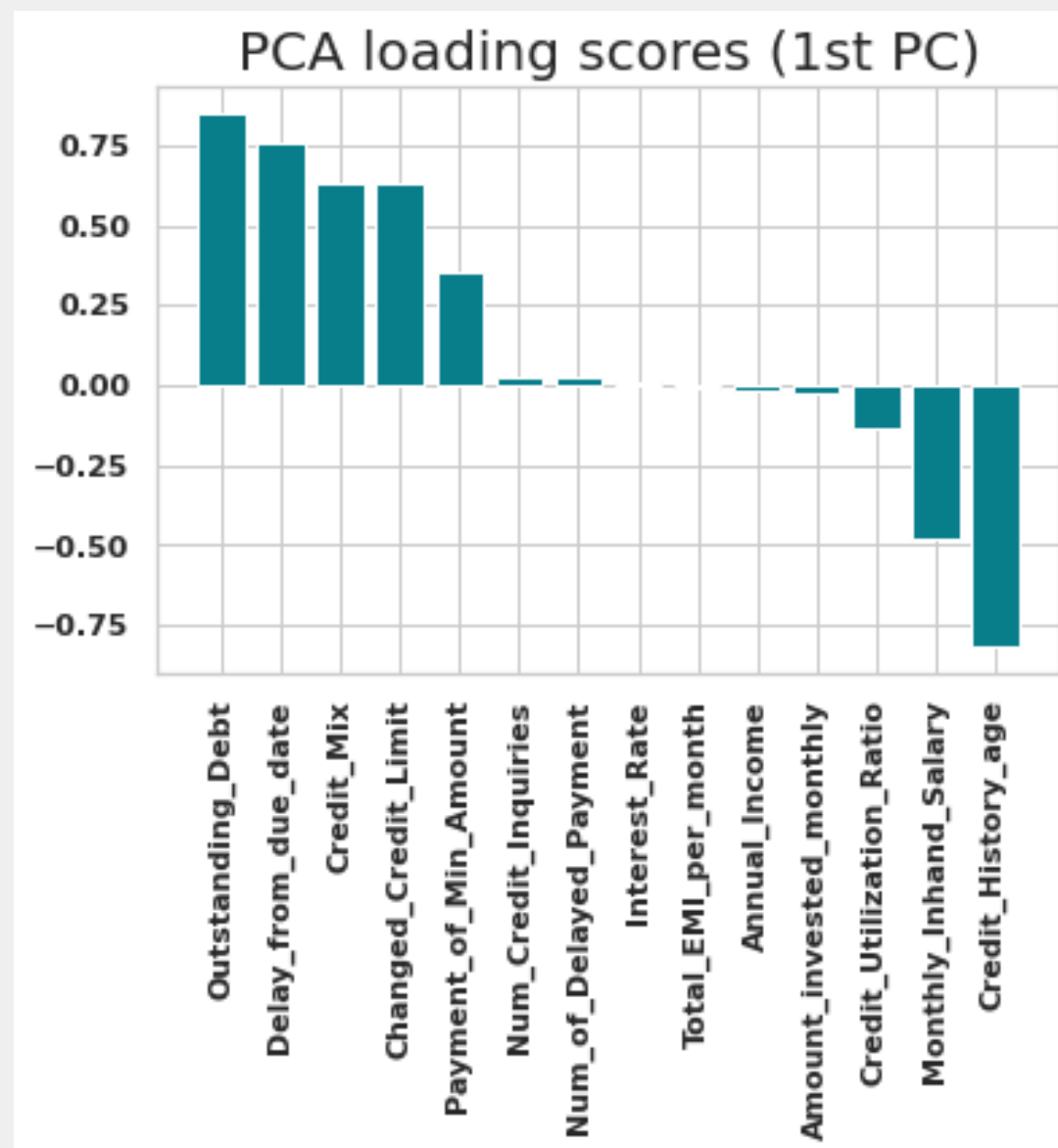
Variância cumulativa explicada por número de PCs



	PC1	PC2	PC3	PC4	PC5	PC6
Annual_Income	-0.01	0.17	-0.27	0.03	0.48	0.69
Monthly_Inhand_Salary	-0.48	0.46	-0.03	-0.05	0.07	-0.03
Interest_Rate	0.01	0.01	0.83	-0.46	0.13	0.29
Delay_from_due_date	0.76	0.08	0.01	-0.03	0.00	-0.02
Num_of_Delayed_Payment	0.03	0.07	0.16	-0.12	0.40	-0.51
Changed_Credit_Limit	0.63	0.13	-0.01	0.02	0.02	-0.01
Num_Credit_Inquiries	0.03	-0.05	-0.10	-0.03	0.09	0.31
Outstanding_Debt	0.85	0.11	-0.01	-0.01	0.01	-0.01
Credit_Utilization_Ratio	-0.14	0.75	-0.10	-0.12	0.22	-0.20
Total_EMI_per_month	-0.01	0.15	0.46	0.87	0.13	0.04
Amount_invested_monthly	-0.02	0.45	0.09	-0.01	-0.72	0.20
Credit_History_age	-0.81	-0.09	0.02	0.00	-0.02	-0.01
Credit_Mix	0.63	0.04	-0.01	-0.01	0.00	-0.00
Payment_of_Min_Amount	0.36	0.01	-0.00	0.00	0.00	-0.00

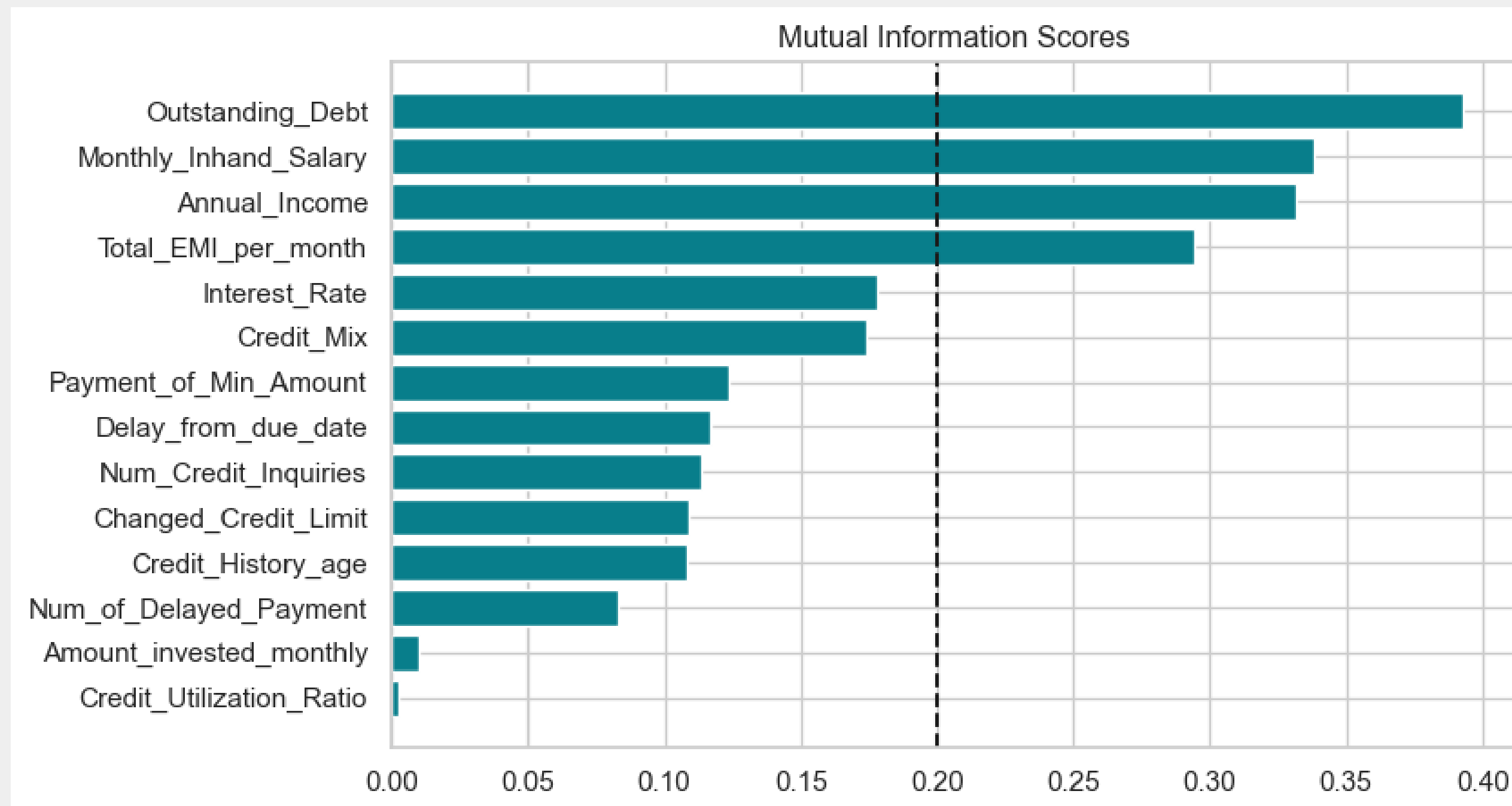
# PCA

- Correlação das 3 primeiras componentes principais com as features:



# Mutual Information Score

A pontuação de Informação Mútua expressa até que ponto a frequência observada de coocorrência entre duas variáveis difere do que seria de esperar, ou seja, o quanto elas se explicam mutuamente.



- 3 Métodos de Feature Selection que conversam entre si, o que trás maior confiança na seleção



# Modelos

Extra Trees

---

Random Forest

---

XGBoost

---

Multi-layer Perceptron

---

Stacking: XGBoost e SVC

# Características Gerais

- Target multi-classe (Poor, Standard, Good);
- Splits com Stratify = y, lida com assimetria da target;
- Todos os modelos foram treinados com features selecionadas;
- Otimização de hiperparâmetros pelo método Bayesiano e com foco em Acurácia e F1-Score (média harmônica entre precision e recall).



$$F_1\text{-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2TP}{2TP + FP + FN}$$

# Comparativo para seleção de Modelos (SKLearn)

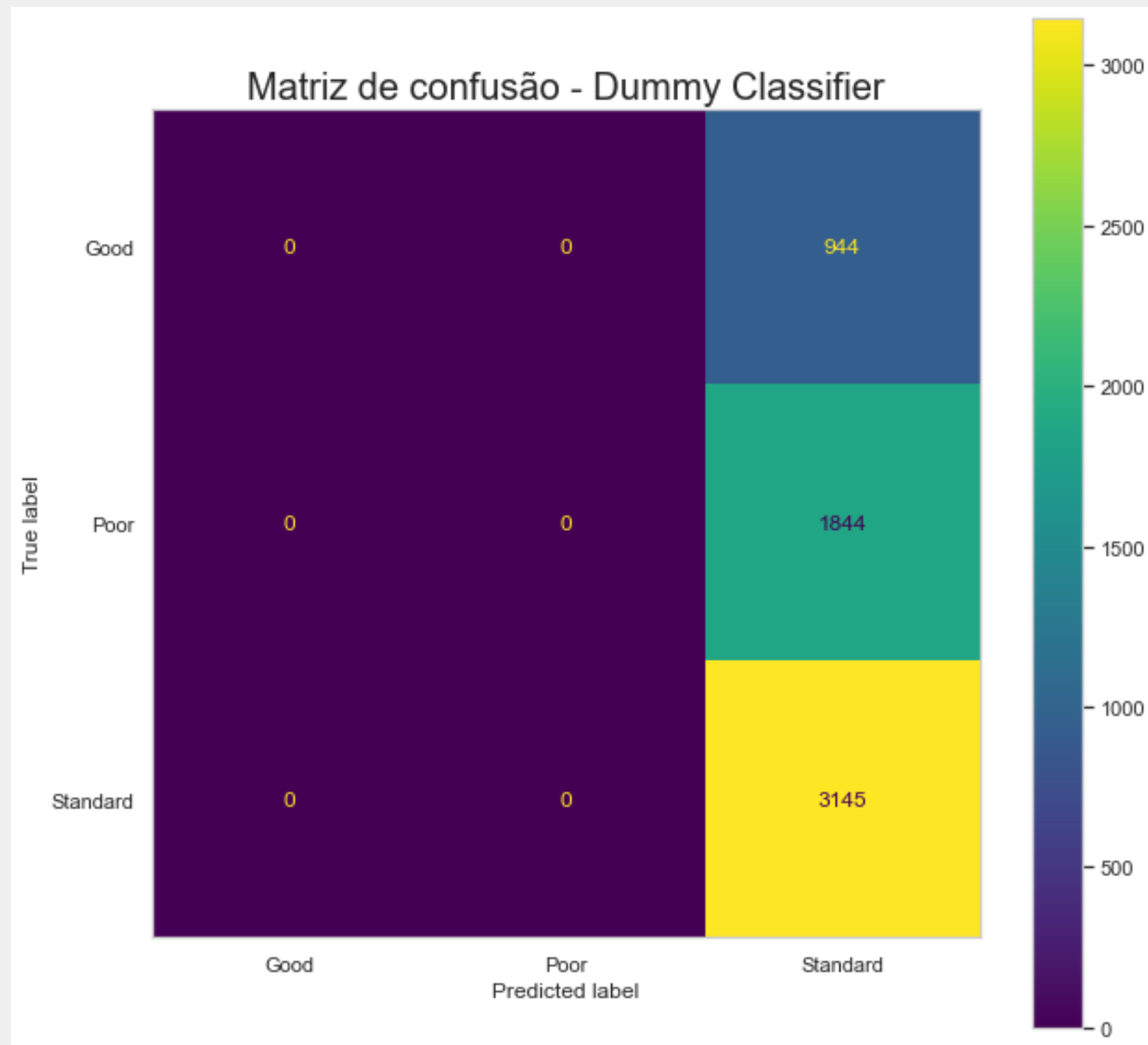
- Através do método Cross Val Score do SKLearn, testamos 7 modelos para escolher os modelos a treinar realmente.

\* Comparação Cross-Validation-Score entre vários modelos:

	Modelo	Score
0	LogisticRegression	0.6108
1	KNeighborsClassifier	0.6647
2	DecisionTreeClassifier	0.6529
3	RandomForestClassifier	0.7470
4	XGBClassifier	0.7248
5	ExtraTreeClassifier	0.6248
6	MLPClassifier	0.6749

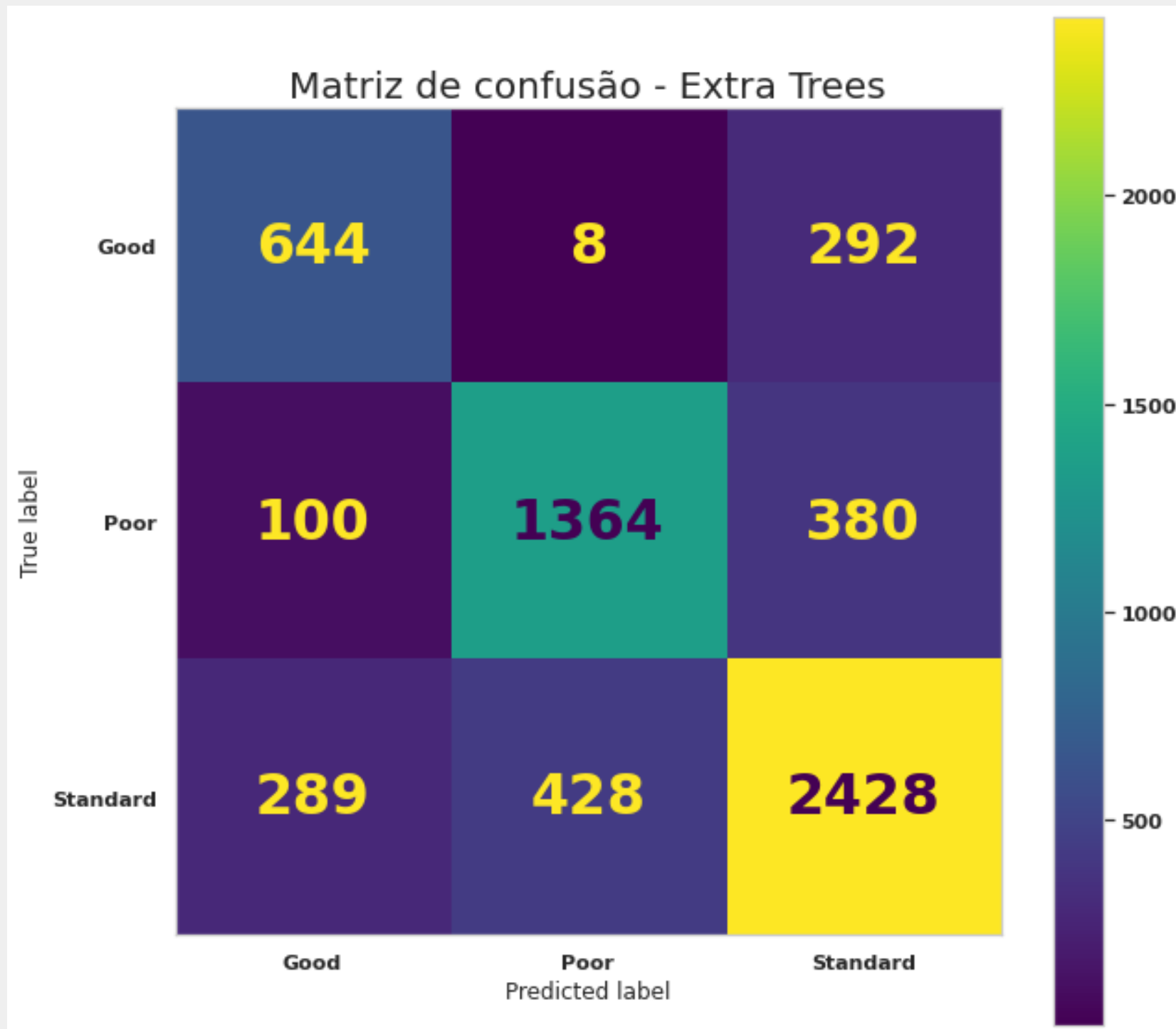


# Modelo Baseline: Dummy Classifier



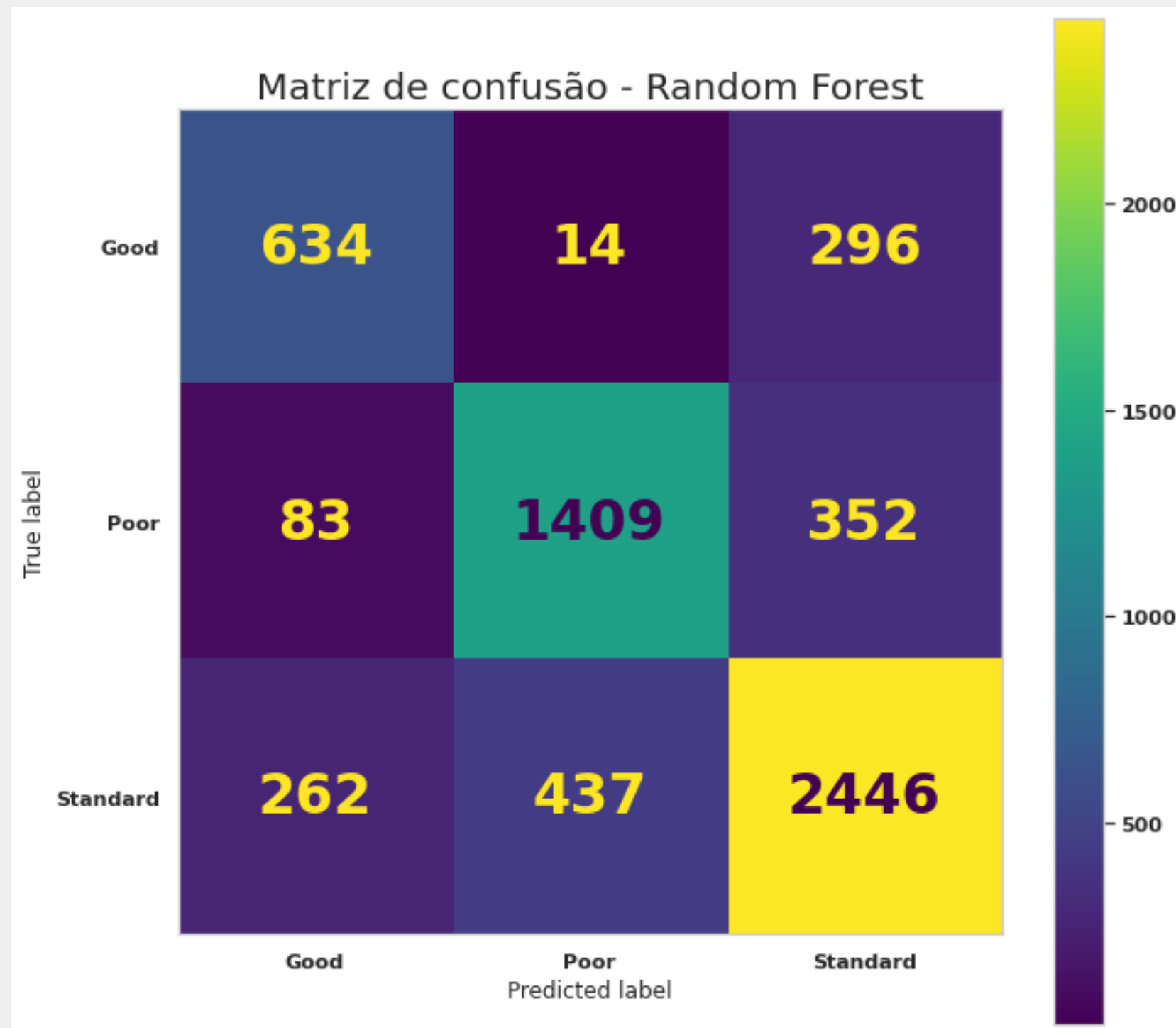
- Modelo que chuta todas as predições como Standard e serve de benchmark na performance dos outros modelos;
- Acurácia: 0.5301;
- OBS.: Não é possível calcular seu Recall e F1-Score, já que teríamos divisão por zero.

# Extra Trees



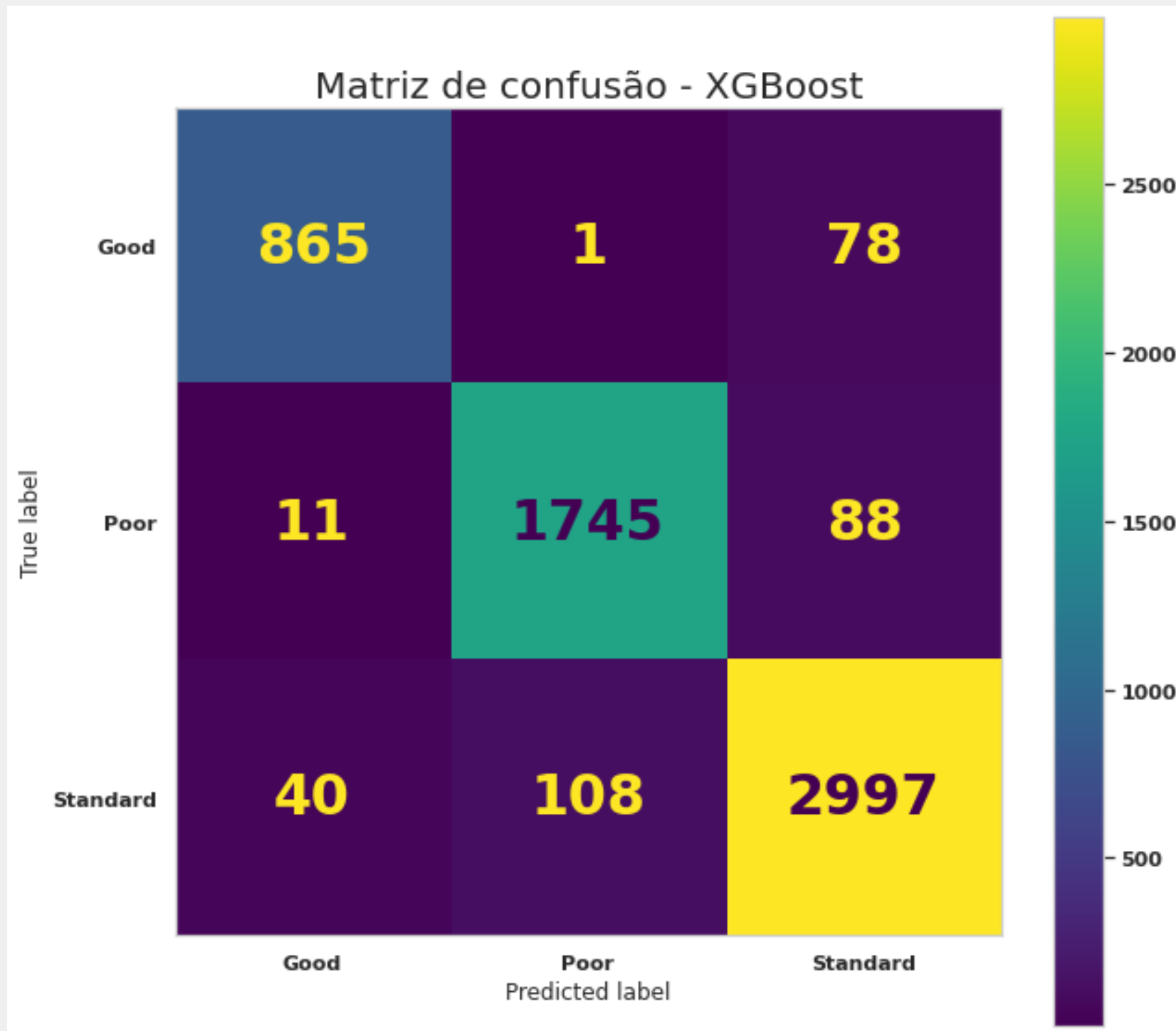
	precision	recall	f1-score	support
0	0.62	0.68	0.65	944
1	0.76	0.74	0.75	1844
2	0.78	0.77	0.78	3145
accuracy			0.75	5933
macro avg	0.72	0.73	0.73	5933
weighted avg	0.75	0.75	0.75	5933

# Random Forest



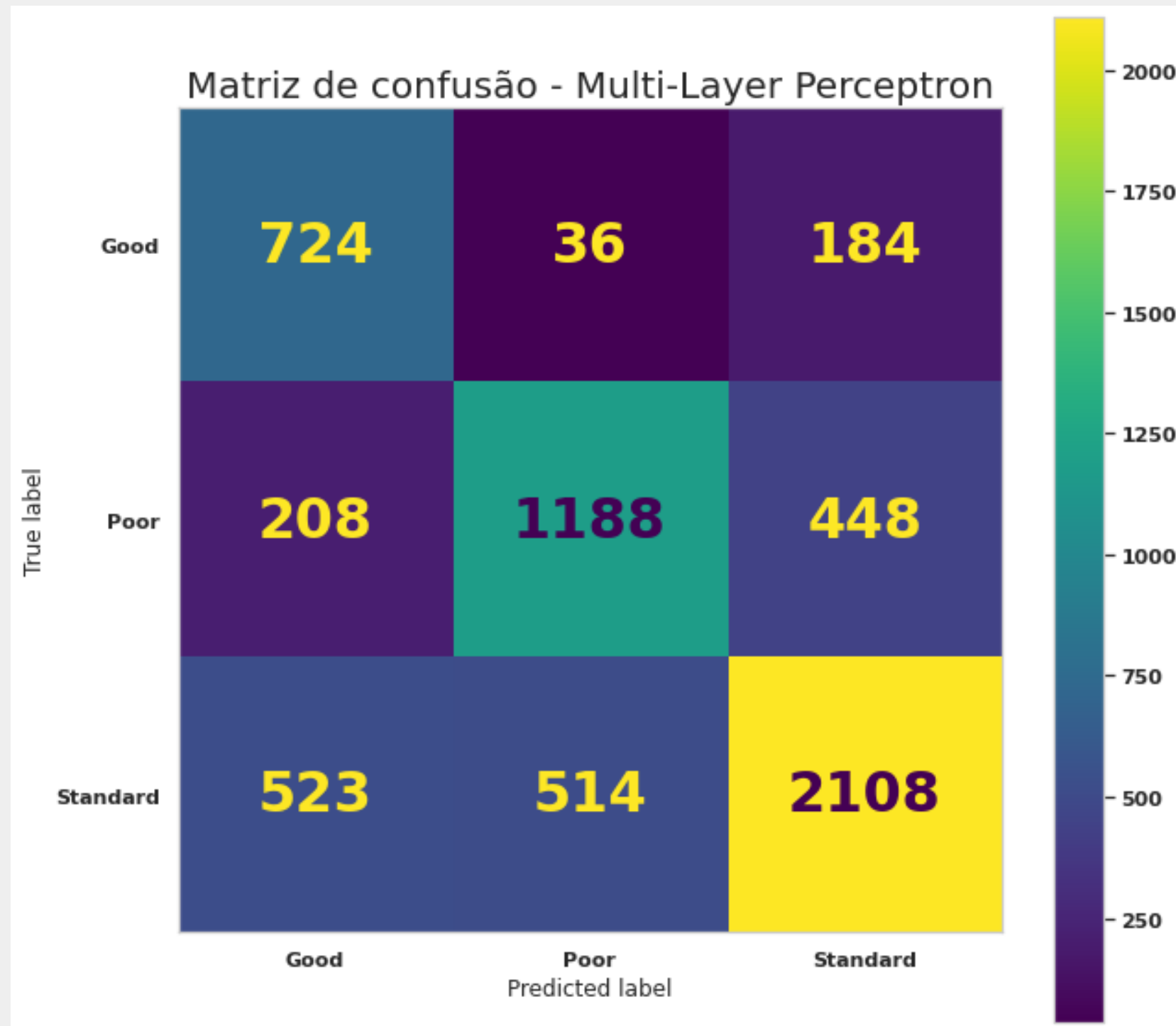
	precision	recall	f1-score	support
0	0.65	0.67	0.66	944
1	0.76	0.76	0.76	1844
2	0.79	0.78	0.78	3145
accuracy			0.76	5933
macro avg	0.73	0.74	0.73	5933
weighted avg	0.76	0.76	0.76	5933

# XGBoost



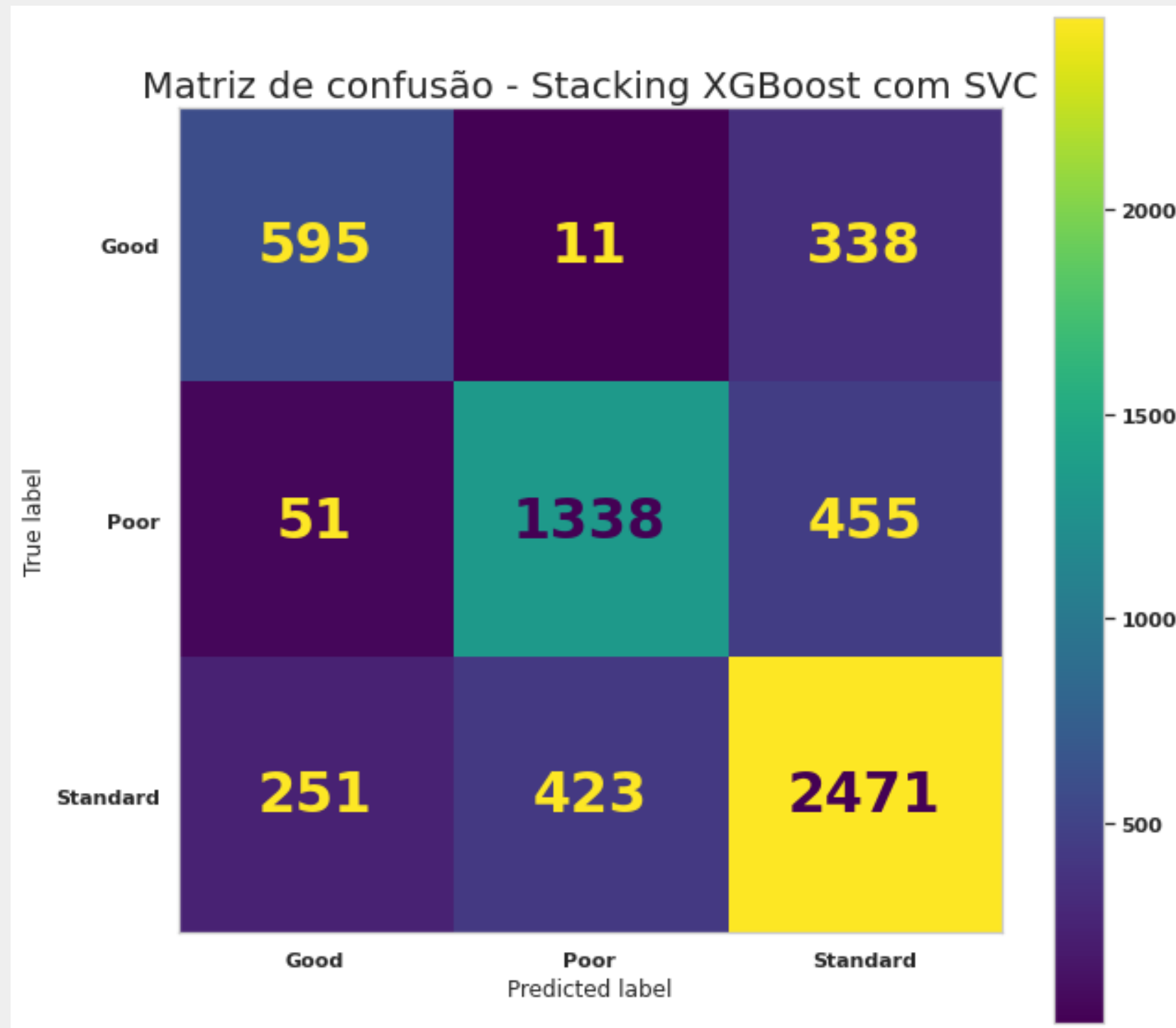
	precision	recall	f1-score	support
0	0.94	0.92	0.93	944
1	0.94	0.95	0.94	1844
2	0.95	0.95	0.95	3145
accuracy			0.95	5933
macro avg	0.94	0.94	0.94	5933
weighted avg	0.95	0.95	0.95	5933

# Multi-Layer Perceptron (SKLearn)



	precision	recall	f1-score	support
0	0.50	0.77	0.60	944
1	0.68	0.64	0.66	1844
2	0.77	0.67	0.72	3145
accuracy			0.68	5933
macro avg	0.65	0.69	0.66	5933
weighted avg	0.70	0.68	0.68	5933

# Stacking de XGBoost com SVC



	precision	recall	f1-score	support
0	0.66	0.63	0.65	944
1	0.76	0.73	0.74	1844
2	0.76	0.79	0.77	3145
accuracy			0.74	5933
macro avg	0.73	0.71	0.72	5933
weighted avg	0.74	0.74	0.74	5933

# Performance

- XGBoost foi o melhor modelo, disparado ( $>$ AUC, Acc, F1;  $<$ FN);
- RF, ExtraTrees e Stacking tiveram performance similar, com maior AUC de RF;
- Dummy Classifier serve de referência para os demais, principalmente durante a otimização de hiperparâmetros.

	Acurácia	Recall	FN	F1-Score	AUROC
XGBoost	0.9451	0.9385	148.0	0.9414	0.9934
Random Forest	0.7566	0.7378	699.0	0.7348	0.8949
Extra Trees	0.7477	0.7313	717.0	0.7259	0.8910
Stacking	0.7423	0.7139	674.0	0.7192	0.8833
Multi Layer Perceptron	0.6776	0.6938	1037.0	0.6611	0.8949
Dummy	0.5301	0.3333	0.0	0.2310	0.5000

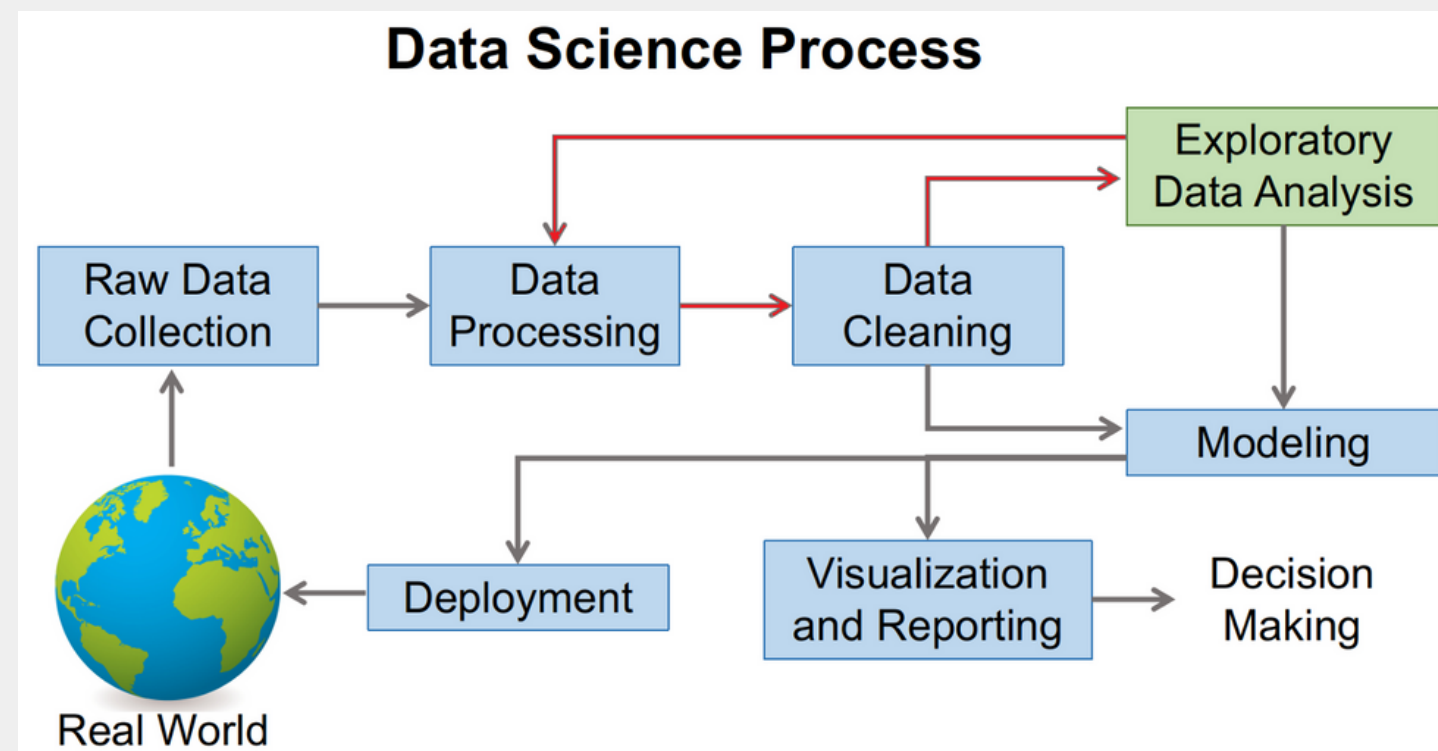
- OBS.: Curvas ROC plotadas no esquema One vs Rest (multi-classe). A variável referência escolhida foi a mais frequente, Standard.





# Conclusões

- A classificação Multi-classe se mostrou uma tarefa à parte já que muitos dos métodos e funções de validação/métricas tiveram que ser substituídos ou reescritos. Não somente, a codificação da target seguiu outro método do que já é costumeiro (binário, OneHot etc);
- Dataset com qualidade baixa de dados, porém com quantia grande de observações. EDA, Data Wrangling e Limpeza de Dados foram pontos de grande foco para lidar com essa situação (vide ciclo iterativo destacado em vermelho).



# Conclusões

- XGBoost com dados processados, normalizados etc teve melhores resultados do que para os dados crus (fora grande tendência de overfitting neste último);
- MLP foi um modelo que trouxe o pior resultado, devido primeiramente ao desconhecimento do mesmo e seus hiperparâmetros por parte dos autores. Utilização das bibliotecas Keras+Tensorflow poderiam trazer o potencial enorme que as redes neurais apresentam;
- Stacking de XGBoost com SVC (e mesmo quando de XGBoost com MLP), também não rendeu bons resultados, já que não foi possível encontrar modelos de soluções complementares. Talvez com uma MLP melhor otimizada esse seria o melhor modelo disparado.



# Fontes



## O que é Score de Crédito?

<https://www.serasa.com.br/score/blog/o-que-e-score-de-credito/>



## Como definir o risco de inadimplência de clientes?

<https://www.cobrefacil.com.br/blog/risco-de-inadimplencia>



## Dados

<https://www.kaggle.com/datasets/parisrohan/credit-score-classification>