

Previsão de *Default* em Empréstimos para Pequenos Negócios



06/09/2020



MBA Analytics em Big Data

Jaime Hikaru Mishima

Coordenadores:

Profª Drª Alessandra de Ávila Montini

Profª Dr. Adolpho Walter Pimazoni Canton

Agenda

1. Objetivo do Trabalho
2. Contextualização do Problema
3. Base de Dados
 - i. Base Original & Tamanho
 - ii. Filtros
 - iii. Descrição das Variáveis
 - iv. Criação de Variáveis
 - v. Tratamento
4. Análise Exploratória de Dados
5. Modelagem
 - i. Métodos de Seleção
 - ii. Estatística Tradicional
 - iii. Interpretação
6. Conclusões
 - i. Impacto para negócios

1. Objetivo do Trabalho

COMECE PELO "POR QUÊ"

4

Por quê?

Como?

O quê?

1 Por quê?

Pequenas empresas contribuem para a economia local por meio de crescimento e inovação para a comunidade, **estimulando o crescimento econômico** e **provendo geração de empregos**. Nesse contexto, **facilitar o acesso ao crédito** aos pequenos negócios é importante para gerar renda à população e reduzir desigualdades sociais.

2 Como?

Por meio de **dados históricos de empréstimos** intermediados pela SBA (agência norte-americana intermediadora) o objetivo é **entender quais variáveis** -cadastrais, temporais, financeiras, negócio - **influenciam no sucesso ou não de um empréstimo**.

3 O quê?

Prever o pagamento de empréstimos do tipo 7a para pequenas empresas. Esses empréstimos são de no máximo \$2MM com a garantia da SBA (agência norte-americana intermediadora) de até \$1.5MM (75%). Buscamos responder: **Qual o risco da empresa que está ofertando o crédito? Dado o risco, devo emprestar ou não?**



2. Contextualização do Problema

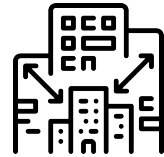
CASO DE USO & ELEGIBILIDADE

5

Para o estudo foram analisados **557k empréstimos** do tipo 7a realizados entre **Outubro de 2009 a Outubro de 2019**.

Os empréstimos do tipo 7a são voltados para comprar um negócio ou obter capital de giro. Historicamente o programa foi concebido para empréstimos de alto risco para aquisições.

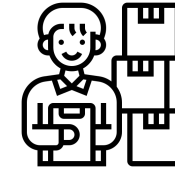
Casos de Uso do Empréstimo



Expansão



Capital de Giro



Compra de Inventário



Móveis e Utensílios

Elegibilidade ao programa



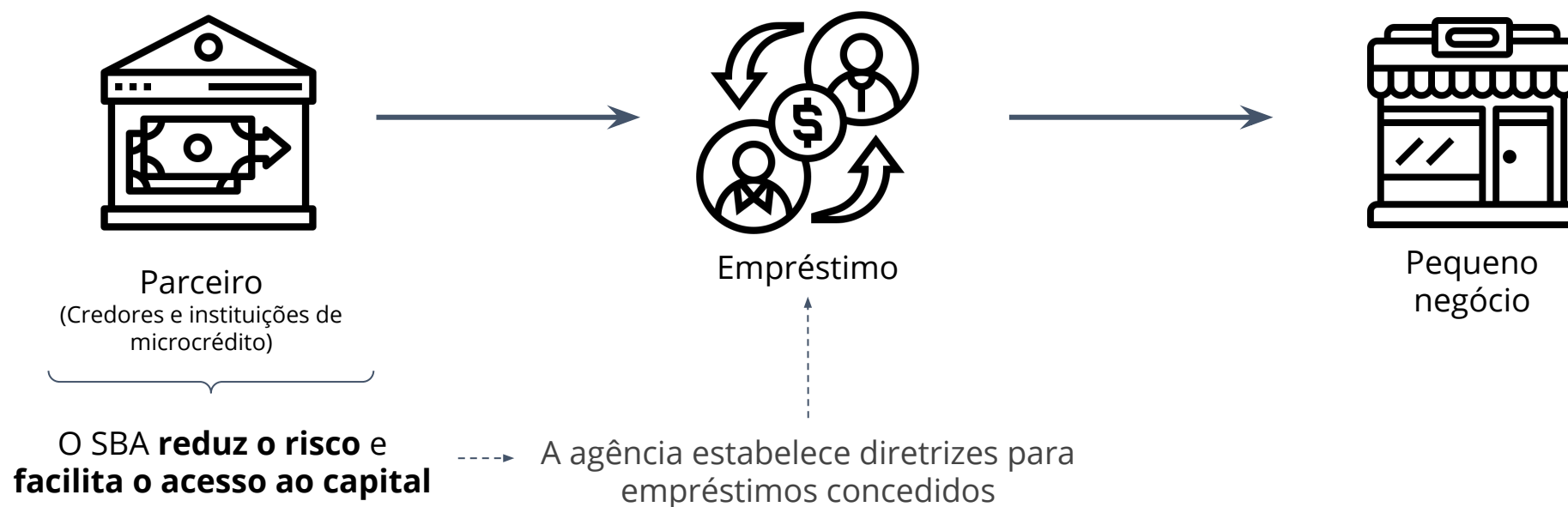
Vendas anuais entre USD 750k e USD 33.5MM para empresas do **varejo, serviços e agricultura**.

2. Contextualização do Problema

FUNCIONAMENTO

Num contexto histórico de empréstimos na última década, o intuito é entender **quais fatores aumentam o risco de *default*** e **como diminuir o risco em empréstimos futuros**.

Default: Quando o tomador do empréstimo é incapaz de fazer pagamentos pontuais, perde pagamentos ou para de fazer pagamentos do empréstimo tomado.



A Small Business Administration (SBA) é uma agência do governo Norte Americano que provê **suporte para empreendedores e pequenas empresas**. O objetivo da SBA é trabalhar com credores para possibilitar empréstimos para pequenas empresas.

3.i Base de Dados

BASE ORIGINAL & TAMANHO

7

1 tabela² contendo **557.542 empréstimos** (registros).

Período de extração: de **01/10/2009 a 30/09/2019**.

Foi selecionado aleatoriamente **70% da base para o treinamento e 30% da base para teste** dos modelos.

32 Variáveis:

- 19 Qualitativas Nominal - ex: endereço, tipo de entrega do empréstimo

- 3 Qualitativas Ordinal - ex: status empréstimo

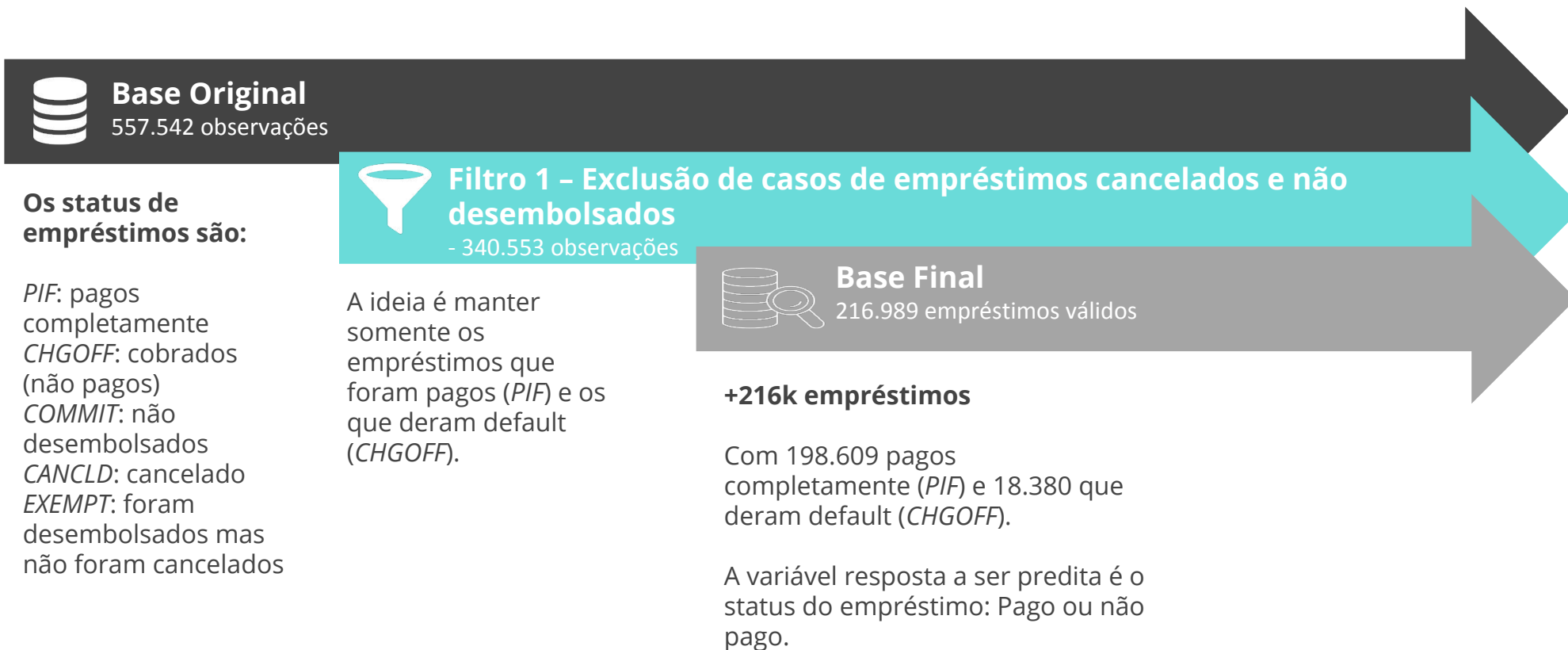
- 8 Quantitativas Contínua - ex: valor total empréstimo, data aprovação empréstimo

- 2 Quantitativas Discreta - ex: tamanho contrato (em meses), empregos criados

² Base obtida em [relatório 2010 - Present SBA 7\(a\) Loan Data](#)

3.ii. Base de Dados

FILTROS



3.iii. Base de Dados

DESCRIÇÃO VARIÁVEIS

9



Variáveis Cadastrais

- Nome do tomador do empréstimo e do banco
- Endereço do mutuário e banco: rua, cidade, estado, zip
- Município e estado do Projeto



Variáveis Temporais

- Data de aprovação do empréstimo
- Ano fiscal de aprovação
- Data do primeiro desembolso do empréstimo



Variáveis Financeiras

- Valor total do empréstimo
- Total de garantia do empréstimo do SBA
- Taxa de juros inicial
- Tempo do contrato (em meses)
- Tipo de entrega do empréstimo: definição e regras do SBA.
- RevolverStatus: Empréstimo a prazo (0) ou linha de crédito (1)



Variáveis do Negócio

- Código e descrição de classificação da indústria
- Código e nome da franquia
- Descrição do subprograma
- Tipo do negócio
- **Target: status do empréstimo (pagou ou não)**
- Empregos criados
- Tipo do empréstimo: no caso todos são do tipo 7a

A **variável resposta** para a modelagem é o **status**. Ela indica se o pagamento do empréstimo foi realizado (0) ou não (*default*, status 1).

Legenda

● Qualitativa Nominal ● Qualitativa Ordinal ● Quantitativa Contínua ● Quantitativa Discreta

3.iv. Base de Dados

CRIAÇÃO DE VARIÁVEIS

10

Baseados nas variáveis da base original, foram criadas features para melhor entender possíveis comportamentos de default.



sameStateLoan

Indica se um mutuário e o banco são do mesmo estado. Regra: 1 quando estado do mutuário (BorrState) é igual ao estado do banco (BankState), 0 caso contrário.



sameCityLoan

Indica se um mutuário e o banco são da mesma cidade. Regra: 1 quando cidade do mutuário (BorrCity) é igual a cidade do banco (BankCity), 0 caso contrário.



shareSBAGuaranteedTotalApproval

Porcentagem do valor emprestado que é garantido pela SBA. $(SBAGuaranteedApproval) / (GrossApproval)$



daysFromApprovalToFirstDisbursement

Tempo em dias da aprovação do empréstimo até o primeiro desembolso. $(FirstDisbursementDate) - (ApprovalDate)$



Franchise

Variável para indicar se um empréstimo foi para uma franquia. Regra: 1 se tem FranchiseCode, 0 caso contrário.

3.iv. Base de Dados

CRIAÇÃO DE VARIÁVEIS

11

A cardinalidade do código da indústria é elevada (1202 tipos de classificações). Com isso foi criada uma nova variável para indicar o setor da indústria, conforme abaixo:



NaicsCode

Sigla de classificação da Indústria Norte-Americana. Cada negócio é classificado baseado em um código de 6 dígitos.

72	2	5	1	4
----	---	---	---	---

722514: **Indústria Nacional** - Cafeterias, Buffets Grill e Buffets

72251: **Indústria** - Restaurantes e Outros lugares de alimentação

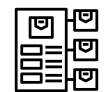
7225: **Grupo** - Restaurantes e Outros lugares de alimentação

722: **Subsetor** - Serviços de comida e bebidas

72: **Setor** - Acomodação e Serviços de comida

NaicsSector

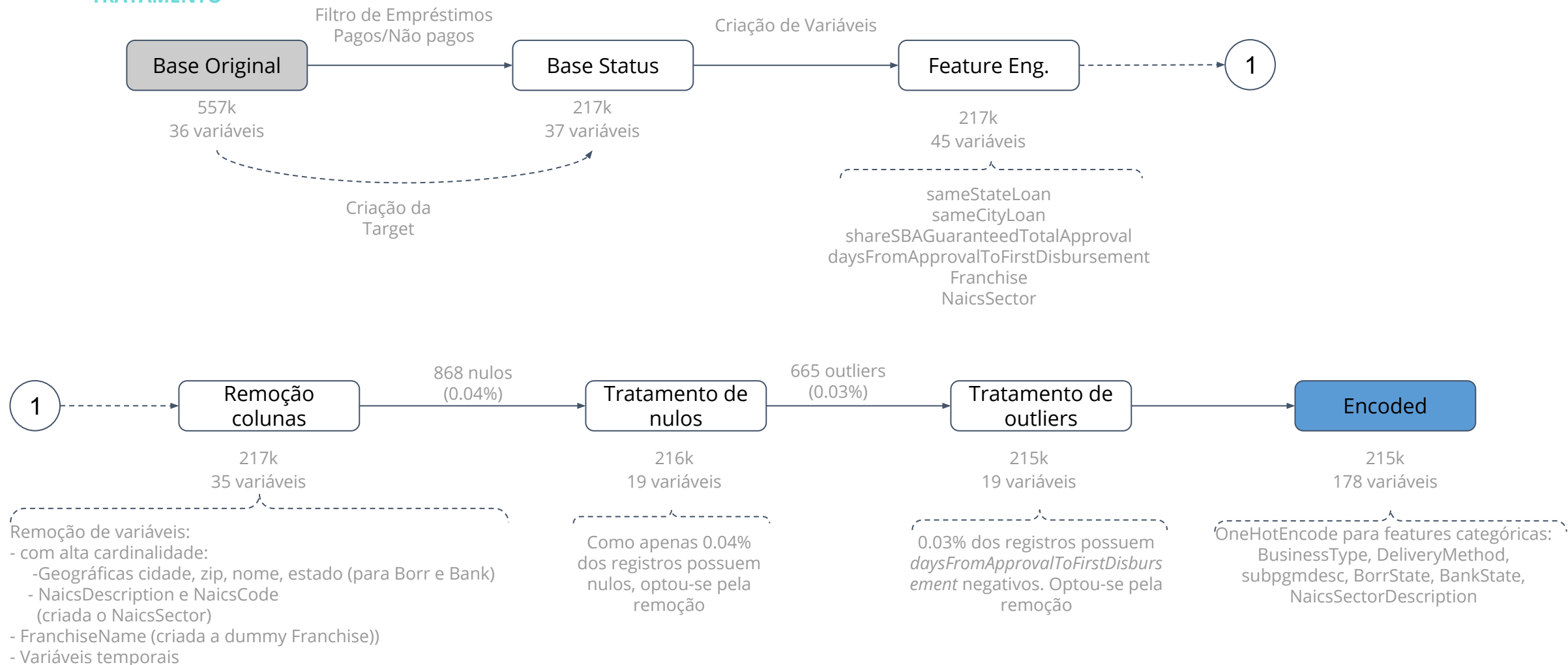
Setor de acordo com a classificação da Indústria Norte-Americana. Ao total são 20 setores.



3.v. Base de Dados

TRATAMENTO

12

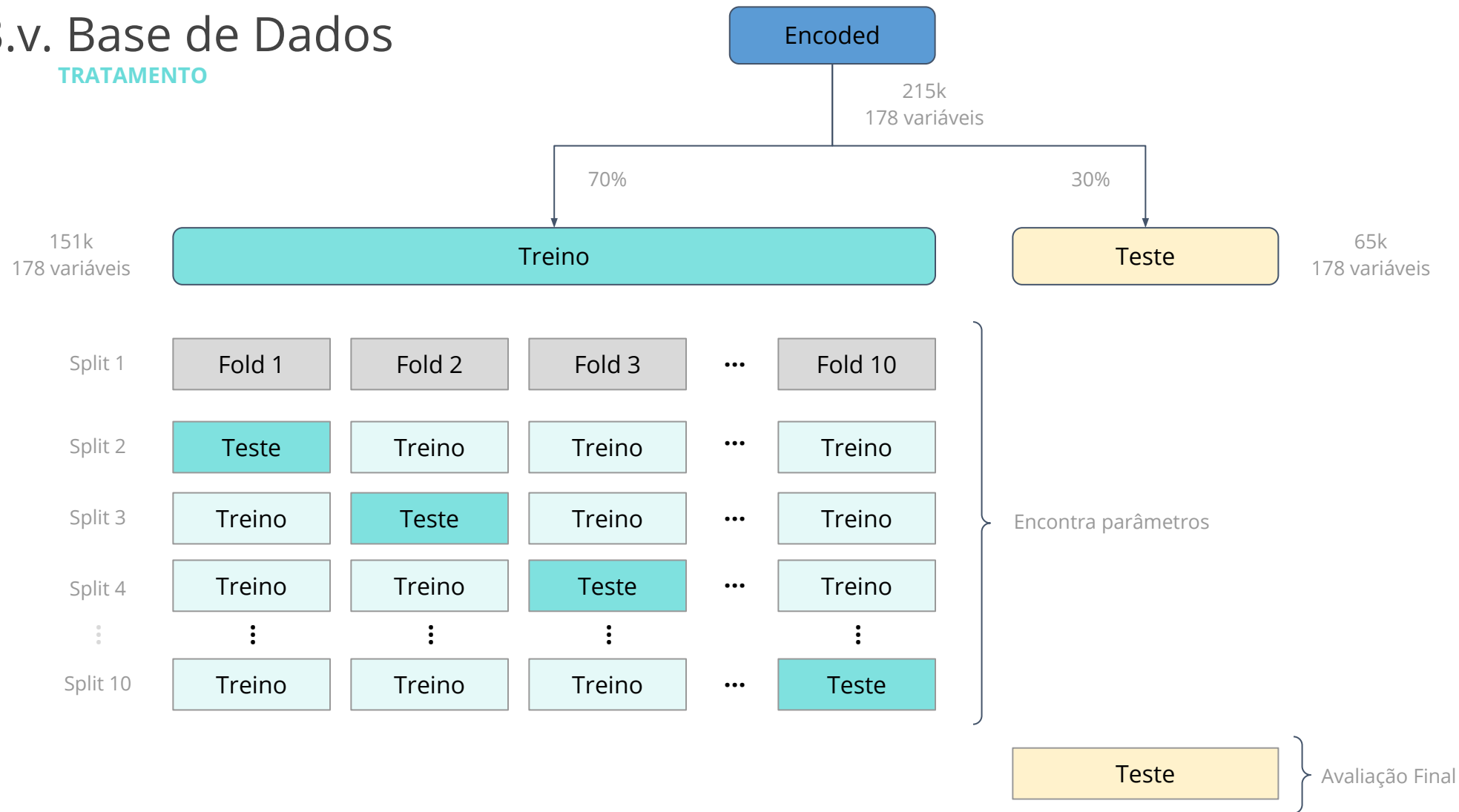


O fluxo acima mostra o processo de geração da ABT, da base original (557k registros e 36 variáveis) até a base final com variáveis categóricas encodadas com uso do OneHotEncoder (215k registros e 178 variáveis).

3.v. Base de Dados

TRATAMENTO

13



A base gerada do fluxo anterior (Encoded) foi separada em **Treino** e **Teste** com **70% e 30% de tamanho**, respectivamente. Para a base de treino, foi realizado **Cross-Validation com 10 folds**. E os resultados foram avaliados no grupo de teste.

Análise Exploratória de Dados

4. Análise Exploratória de Dados

ANÁLISE DE DISTRIBUIÇÃO

Tabela 1. Valores de média, desvio padrão (DP) e distribuição de percentis.

Variável	Média	DP	P ₀	P ₂₅	P ₅₀	P ₇₅	P ₁₀₀
GrossApproval (USD)	314K	598K	1.0K	30K	100K	300K	5.0M
SBAGuaranteedApproval (USD)	234K	470K	0.5K	18K	53K	205K	5.3M
InitialInterestRate (%)	6.3	1.4	0.0	5.5	6.0	7.1	12.5
TermInMonths	104	73	0	60	84	120	360
RevolverStatus	0.35	0.48	0	0	0	1	1
JobsSupported	11.1	22.1	0	2	5	12	2150
Target	0.08	0.28	0.0	0.0	0.0	0.0	1.0
sameStateLoan	0.5	0.5	0.0	0.0	0.0	1.0	1.0
sameCityLoan	0.1	0.3	0.0	0.0	0.0	0.0	1.0
shareSBAGuaranteedTotalApproval	0.64	0.16	0.15	0.50	0.50	0.75	2.25
daysFromApprovalToFirstDisbursement	50	130	-3648	0	9	43	3652
Franchise	0.07	0.25	0	0	0	0	1

A tabela 1 mostra a distribuição das variáveis numéricas. O objetivo é entender um pouco do processo gerador dos dados.

4. Análise Exploratória de Dados

ANÁLISE DE DISTRIBUIÇÃO

Tabela 1. Valores de média, desvio padrão (DP) e distribuição de percentis.

Variável	Média	DP	P ₀	P ₂₅	P ₅₀	P ₇₅	P ₁₀₀
GrossApproval (USD)	314K	598K	1.0K	30K	100K	300K	5.0M
SBAGuaranteedApproval (USD)	234K	470K	0.5K	18K	53K	205K	5.3M
InitialInterestRate (%)	6.3	1.4	0.0	5.5	6.0	7.1	12.5
TermInMonths	104	73	0	60	84	120	360
RevolverStatus	0.35	0.48	0	0	0	1	1
JobsSupported	11.1	22.1	0	2	5	12	2150
Target	0.08	0.28	0.0	0.0	0.0	0.0	1.0
sameStateLoan	0.5	0.5	0.0	0.0	0.0	1.0	1.0
sameCityLoan	0.1	0.3	0.0	0.0	0.0	0.0	1.0
shareSBAGuaranteedTotalApproval	0.64	0.16	0.15	0.50	0.50	0.75	2.25
daysFromApprovalToFirstDisbursement	50	130	-3648	0	9	43	3652
Franchise	0.07	0.25	0	0	0	0	1

- O média do valor total do empréstimo (GrossApproval) foi de 314k USD. 50% dos empréstimos foram de até 100k USD e 75% até 300k USD. O máximo valor de empréstimo foi 5MM USD.
- A média garantida pelo SBA sobre o valor do empréstimo foi de 234k USD.
- A garantia média dada pelo SBA foi de 64%.

4. Análise Exploratória de Dados

ANÁLISE DE DISTRIBUIÇÃO

Tabela 1. Valores de média, desvio padrão (DP) e distribuição de percentis.

Variável	Média	DP	P ₀	P ₂₅	P ₅₀	P ₇₅	P ₁₀₀
GrossApproval (USD)	314K	598K	1.0K	30K	100K	300K	5.0M
SBAGuaranteedApproval (USD)	234K	470K	0.5K	18K	53K	205K	5.3M
InitialInterestRate (%)	6.3	1.4	0.0	5.5	6.0	7.1	12.5
TermInMonths	104	73	0	60	84	120	360
RevolverStatus	0.35	0.48	0	0	0	1	1
JobsSupported	11.1	22.1	0	2	5	12	2150
Target	0.08	0.28	0.0	0.0	0.0	0.0	1.0
sameStateLoan	0.5	0.5	0.0	0.0	0.0	1.0	1.0
sameCityLoan	0.1	0.3	0.0	0.0	0.0	0.0	1.0
shareSBAGuaranteedTotalApproval	0.64	0.16	0.15	0.50	0.50	0.75	2.25
daysFromApprovalToFirstDisbursement	50	130	-3648	0	9	43	3652
Franchise	0.07	0.25	0	0	0	0	1

- A taxa de juros média foi de 6.3% (*InitialInterestRate*), com um desvio de 1.4%.
- A duração média do empréstimo é de 104 meses, ou seja 8 anos e 8 meses (*TermInMonths*).
- 8% dos empréstimos deram *default*.
- 7% dos empréstimos foram de franquias.

4. Análise Exploratória de Dados

ANÁLISE DE DISTRIBUIÇÃO

Tabela 1. Valores de média, desvio padrão (DP) e distribuição de percentis.

Variável	Média	DP	P ₀	P ₂₅	P ₅₀	P ₇₅	P ₁₀₀
GrossApproval (USD)	314K	598K	1.0K	30K	100K	300K	5.0M
SBAGuaranteedApproval (USD)	234K	470K	0.5K	18K	53K	205K	5.3M
InitialInterestRate (%)	6.3	1.4	0.0	5.5	6.0	7.1	12.5
TermInMonths	104	73	0	60	84	120	360
RevolverStatus	0.35	0.48	0	0	0	1	1
JobsSupported	11.1	22.1	0	2	5	12	2150
Target	0.08	0.28	0.0	0.0	0.0	0.0	1.0
sameStateLoan	0.5	0.5	0.0	0.0	0.0	1.0	1.0
sameCityLoan	0.1	0.3	0.0	0.0	0.0	0.0	1.0
shareSBAGuaranteedTotalApproval	0.64	0.16	0.15	0.50	0.50	0.75	2.25
daysFromApprovalToFirstDisbursement	50	130	-3648	0	9	43	3652
Franchise	0.07	0.25	0	0	0	0	1

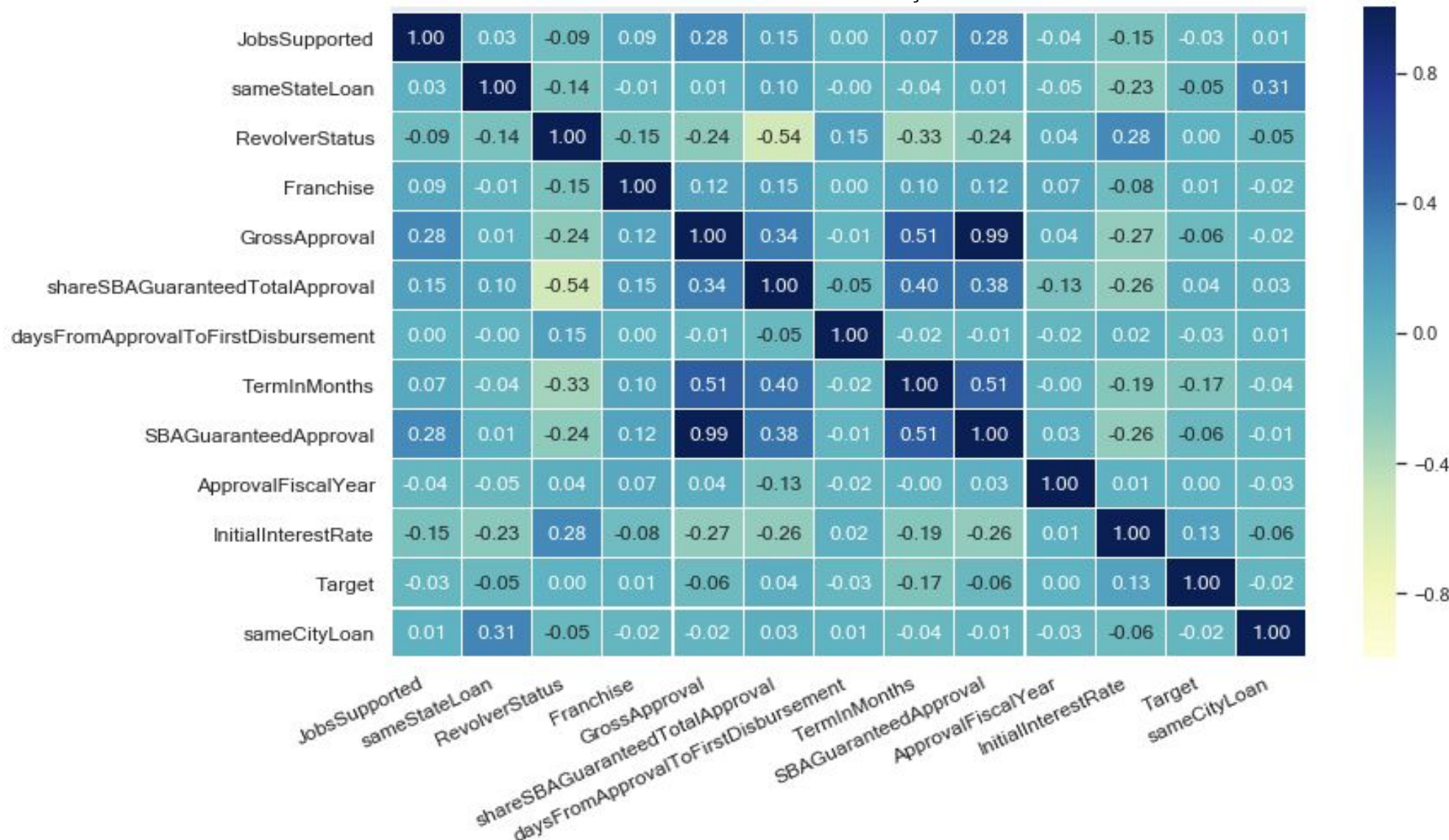
- 35% dos empréstimos foram do tipo linha de crédito rotativo (*RevolverStatus* igual a 1). 65% foram a prazo (*RevolverStatus* igual a 0).
- Os empréstimos propiciaram a criação, em média, de 11.1 empregos (Variável *JobsSupported*). 75% dos empréstimos geraram até 12 empregos.

4. Análise Exploratória de Dados

CORRELAÇÃO DE VARIÁVEIS

19

Tabela 2. Matriz de correlação de Pearson

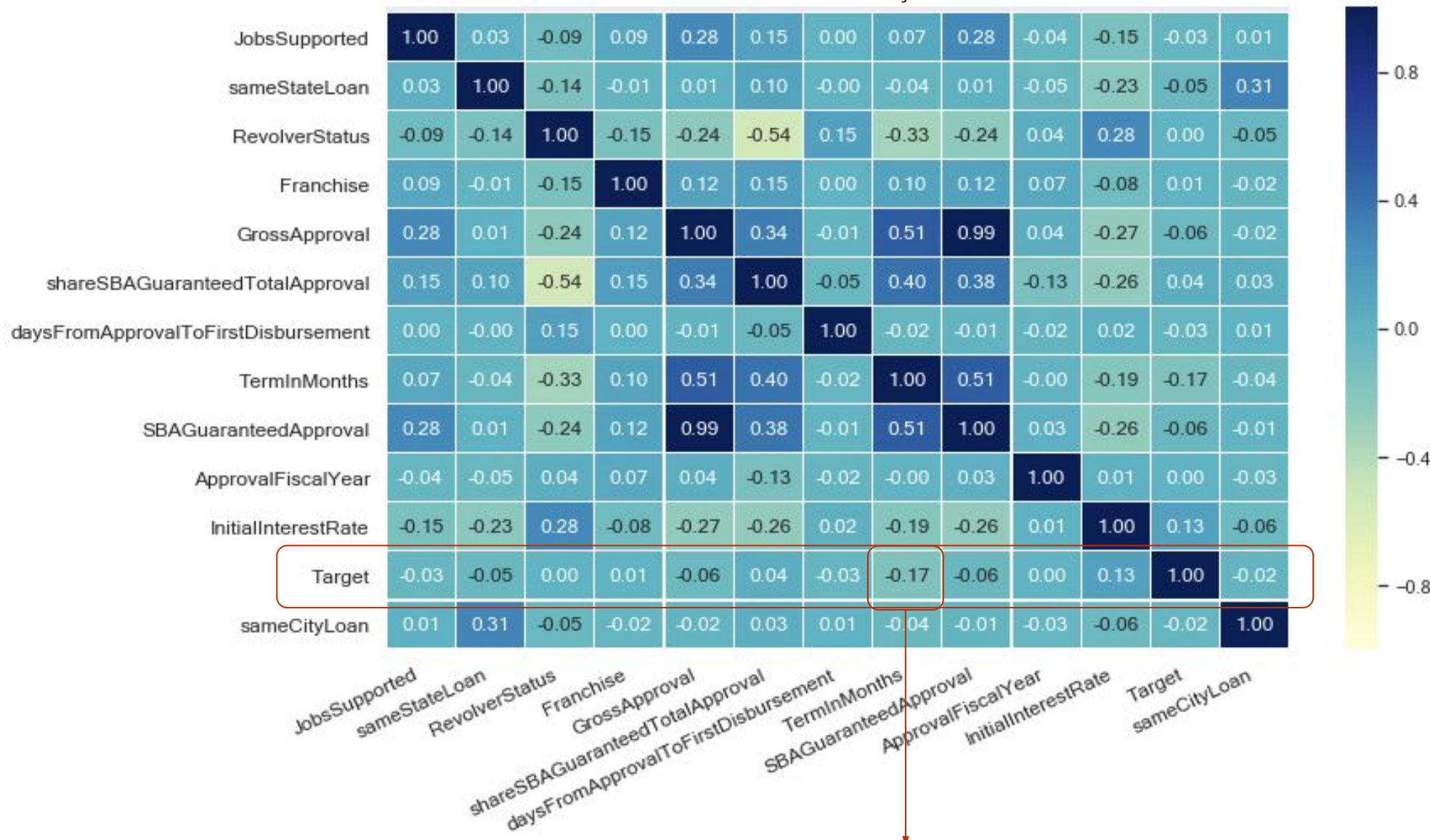


Na tabela 2 temos as **correlações de Pearson entre as variáveis**. Ela mede a correlação linear entre duas variáveis, tendo um valor entre +1 e -1. +1 indica correlação linear perfeita positiva, 0 diz que não há correlação e -1 indica uma correlação perfeita negativa.

4. Análise Exploratória de Dados

CORRELAÇÃO DE VARIÁVEIS

Tabela 2. Matriz de correlação de Pearson



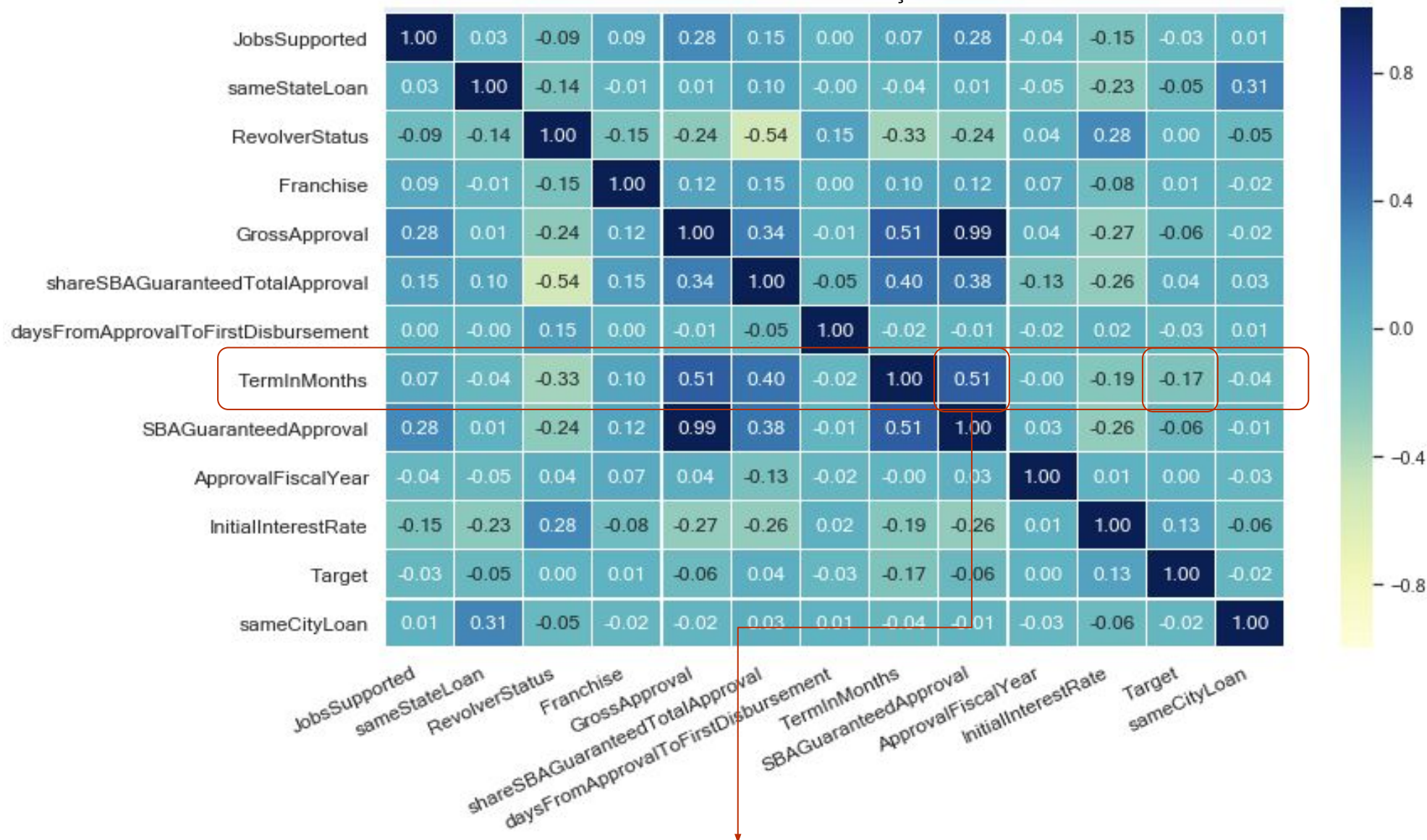
O tempo de duração do empréstimo tem uma alta correlação negativa com a **target** (-17%)

4. Análise Exploratória de Dados

CORRELAÇÃO DE VARIÁVEIS

21

Tabela 2. Matriz de correlação de Pearson

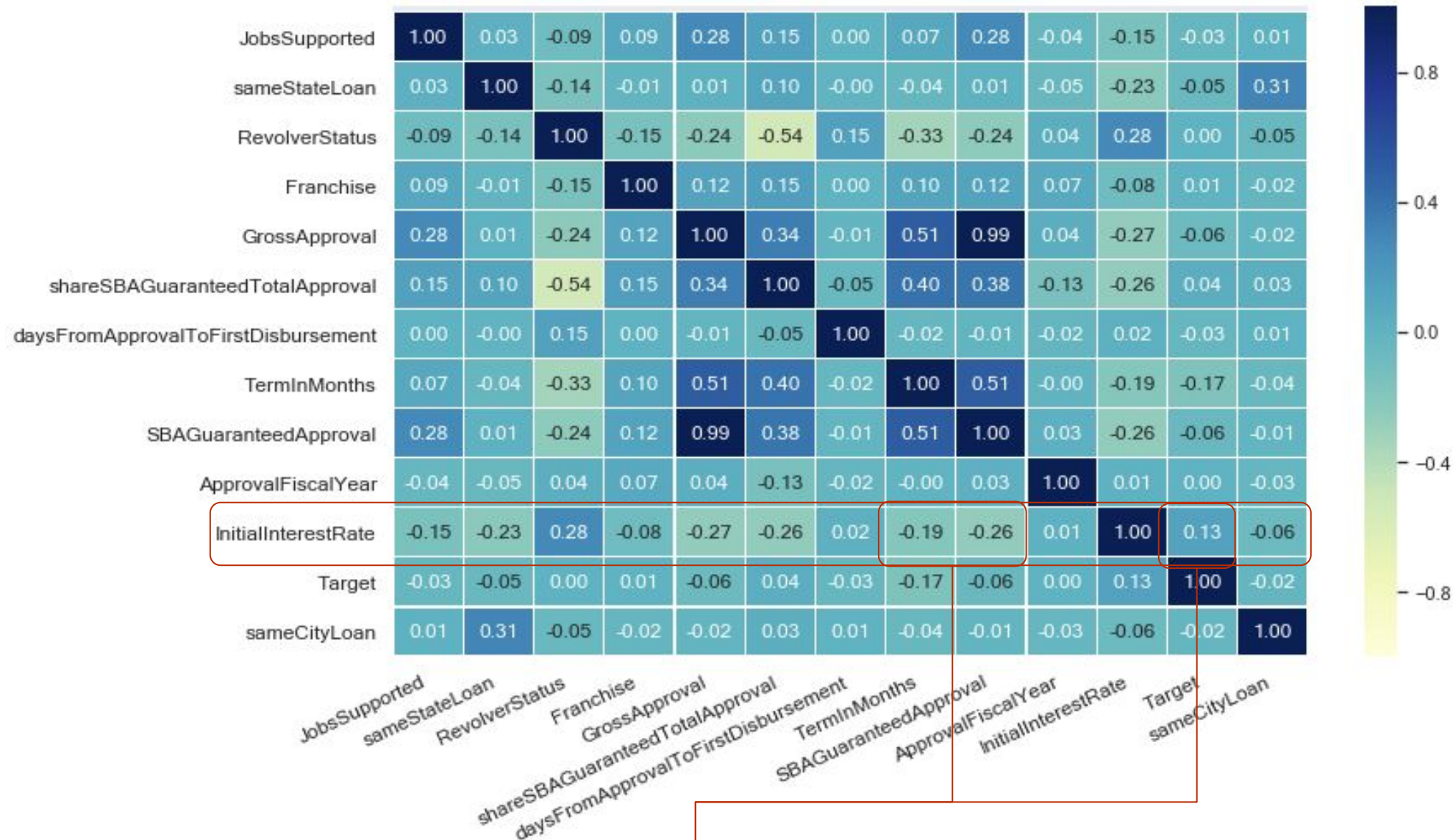


O **tempo de duração do empréstimo** está diretamente relacionado com o **valor total do empréstimo** (51%), bem como o quanto o SBA garante sobre o valor (51%).

4. Análise Exploratória de Dados

CORRELAÇÃO DE VARIÁVEIS

Tabela 2. Matriz de correlação de Pearson

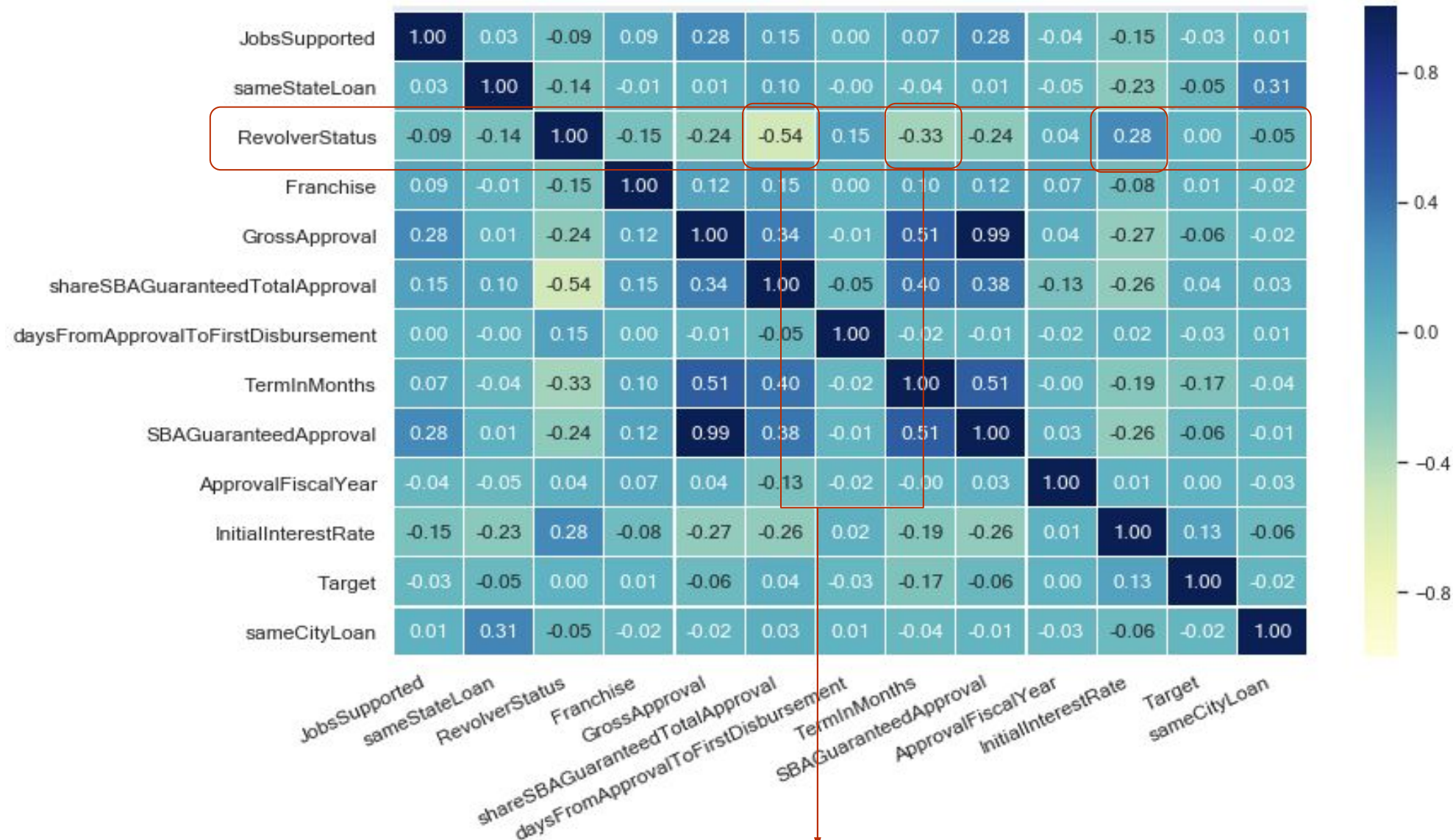


A taxa de juros inicial tem correlação positiva com a target, 13%. E tem correlação negativa com o quanto o SBA garante sobre o empréstimo (26%) e o tempo de duração do empréstimo (-19%)

4. Análise Exploratória de Dados

CORRELAÇÃO DE VARIÁVEIS

Tabela 2. Matriz de correlação de Pearson

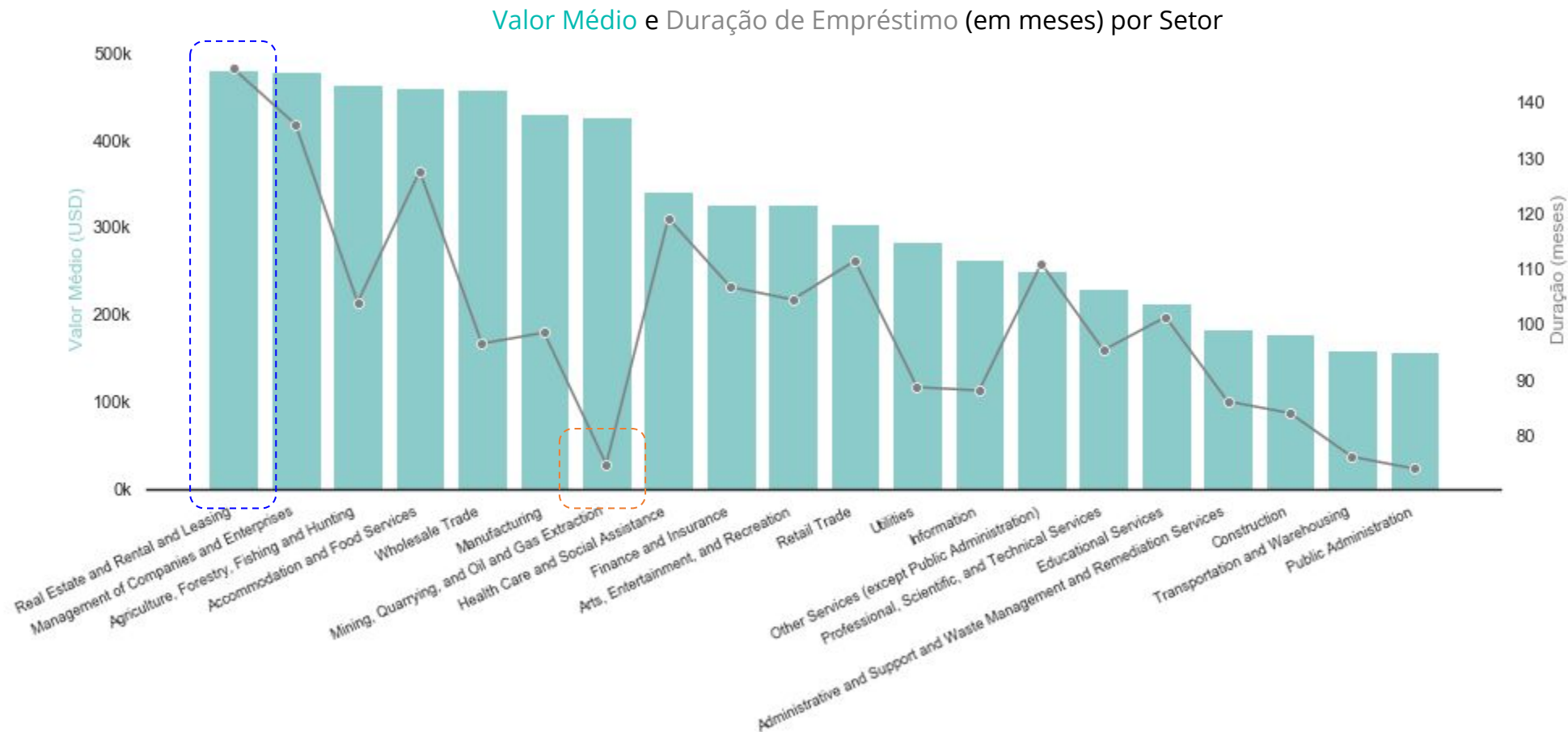


RevolverStatus tem uma correlação negativa com o **tempo de duração do empréstimo** (-33%) e o quanto o **SBA garante sobre o empréstimo** (-54%). Essa métrica diz se o empréstimo foi a prazo (*term*) ou rotativo (*revolving*). Sendo *revolver* igual a 1, faz sentido pois empréstimos rotativos são mais flexíveis mas em compensação tem uma taxa de juros inicial maior (+28% com InitialInterestRate).

4. Análise Exploratória de Dados

SETORES

24



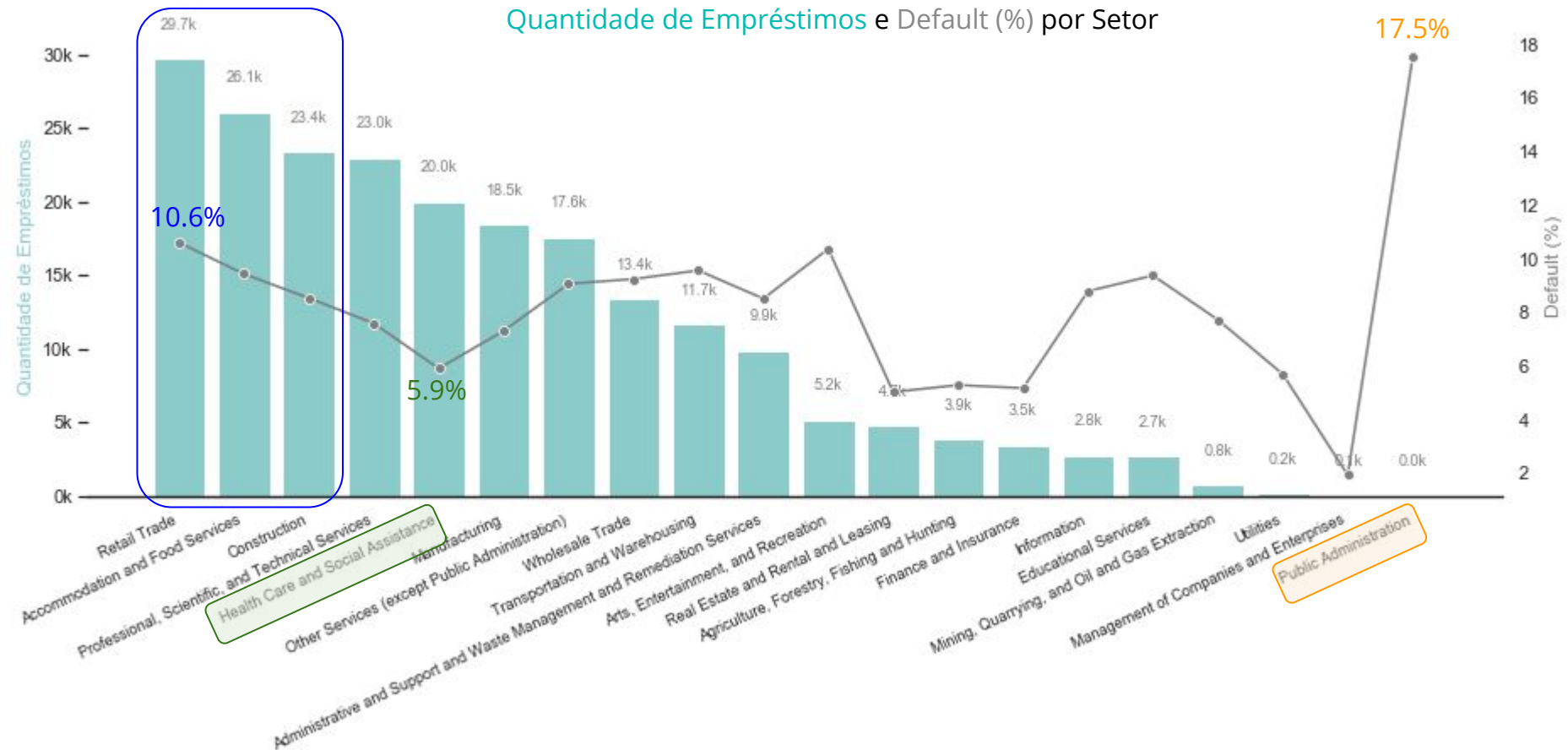
O setor de **Imóveis, Aluguel e Locação** (*Real Estate, Rental and Leasing*) tem o maior valor médio de empréstimo (481k USD) e maior tempo de duração (146 meses).

Mineração, Extração, Óleo (*Mining, Quarrying and Oil*) é um dos setores com menor tempo de duração média do empréstimo, 75 meses.

4. Análise Exploratória de Dados

SETORES

25



Comércio (Retail Trade), **Acomodação e Alimentação** (Accommodation and Food Services) e **Construção** (Constructions) foram os setores que mais realizaram empréstimos, representando 36.5% dos empréstimos.

Administração Pública (Public Administration) teve a maior taxa de default (17.5%) porém ela representou uma pequena parcela dos empréstimos (0.02%).

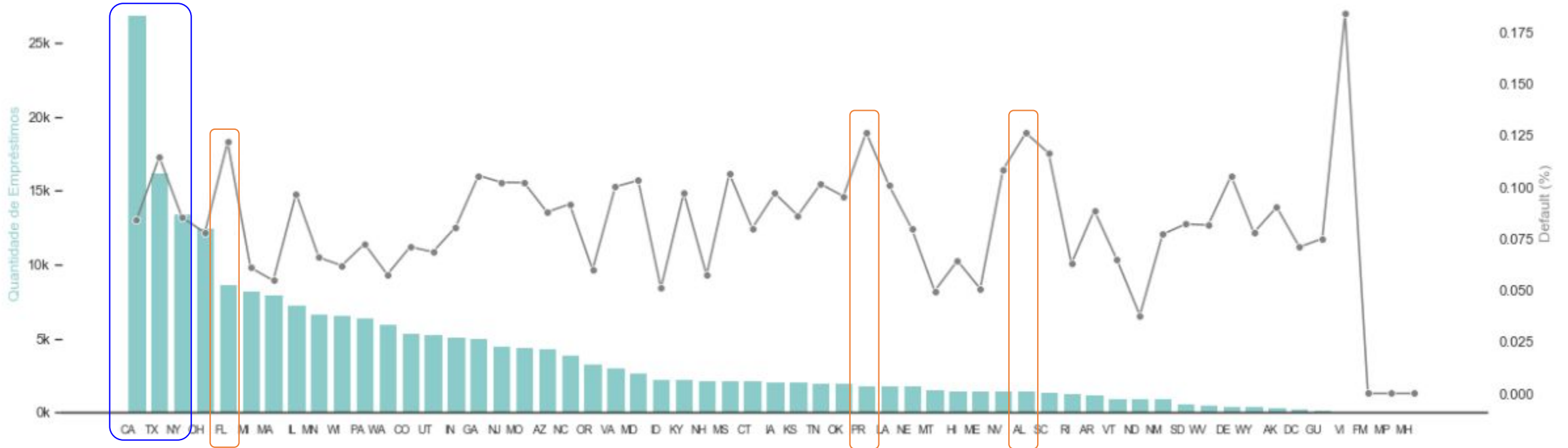
A taxa de default para **comércio** foi de 10.6%. Nos top 5 maiores setores em volume de empréstimos, **Planos de Saúde e Assistência Social** (Health Care and Social Assistance) teve o menor taxa de default, 5.9%.

4. Análise Exploratória de Dados

ESTADO

26

Quantidade de Empréstimos e Default (%) por Estado



CA (Califónia, 12.4%), TX (Texas, 7.5%) e NY (Nova Iorque, 6.2%) foram os estados com maior volume de empréstimos (26% de todos os empréstimos)

PR (Puerto Rico), AL (Alabama) e FL (Flórida) foram os estados com o maior percentual de default (13%, 13% e 12% respectivamente). VI (Virgínia) teve a maior taxa de default (18%) porém somente 38 empréstimos (0.02%) foram realizados.

4. Análise Exploratória de Dados

TIPOS DE NEGÓCIO

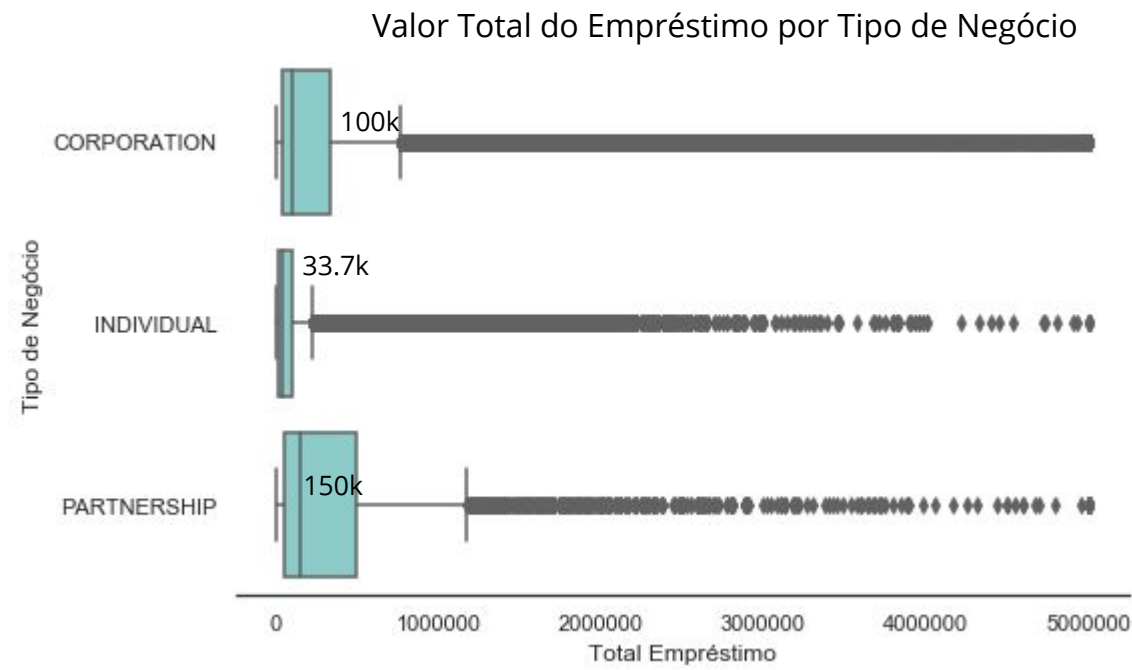


Tabela 3. Valores de média, desvio padrão (DP) e distribuição de percentis.

Tipo de Negócio	Corporation	Individual	Partnership
Quantidade	187k (86.3%)	25.9k (12%)	3.8k (1.8%)
Média	333.1K	154.3K	459.8K
DP	615.5K	373.4K	752.3K
P ₀	1.0K	2.3K	3.0K
P ₂₅	40.0K	17.2K	50.0K
P ₅₀	100.0K	33.7K	150.0K
P ₇₅	331.3K	100.0K	500.0K
P ₁₀₀	5.0M	5.0M	5.0M

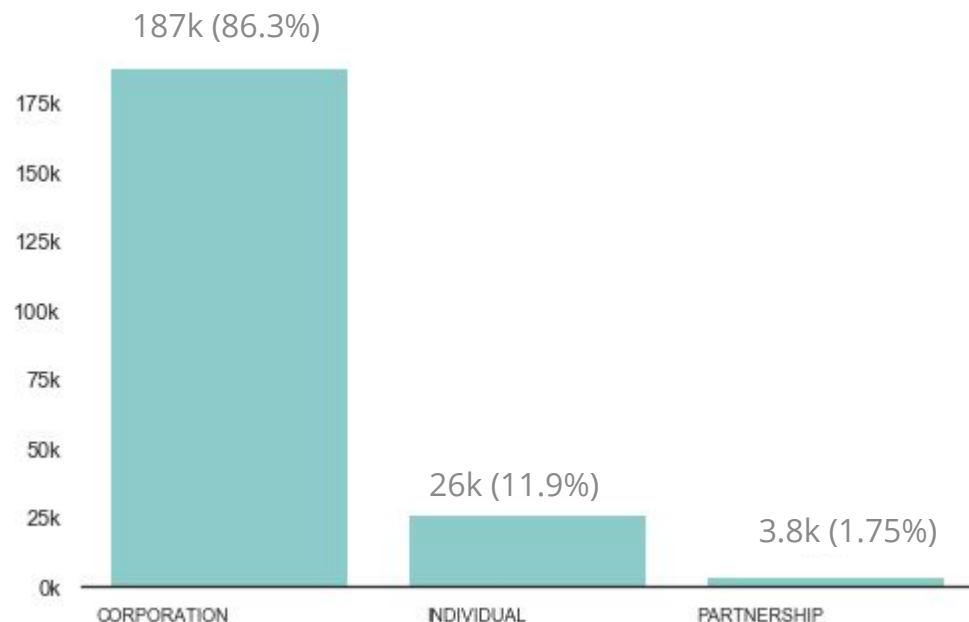
Comparando pelos tipos de negócio, **negócios Individuais têm menor mediana (33.7k USD)** comparado com Corporação (100k USD, Corporation) e Parceria (150k USD, Partnership)

A Tabela 3 mostra que a média dos empréstimos individuais é a menor, em 154k USD. Até 75% desses empréstimos tiveram valor até 100k USD.

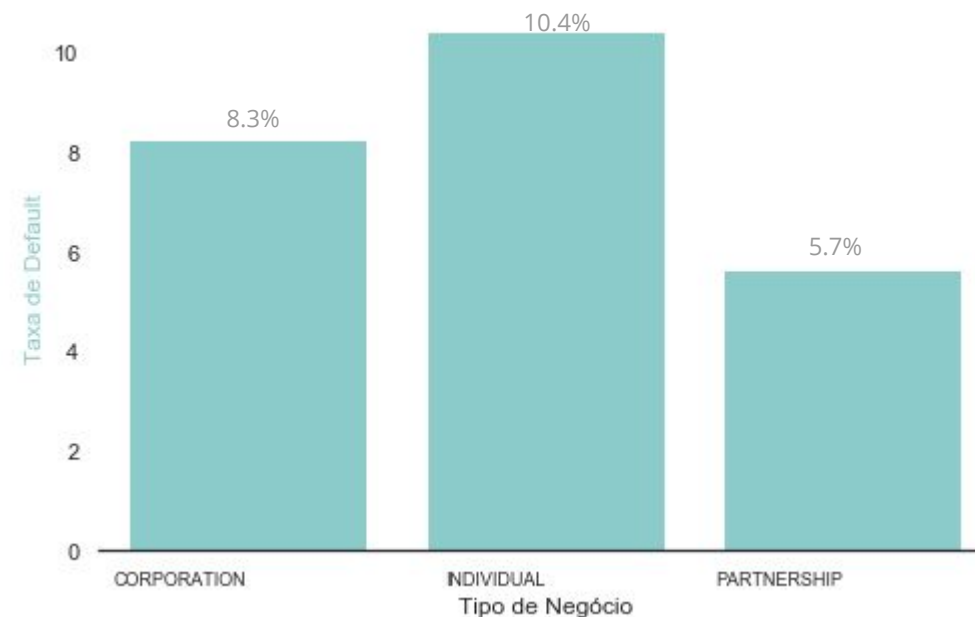
4. Análise Exploratória de Dados

TIPOS DE NEGÓCIO

Volume de Empréstimos por Tipo de Negócio



Taxa de Default por tipo de Negócio



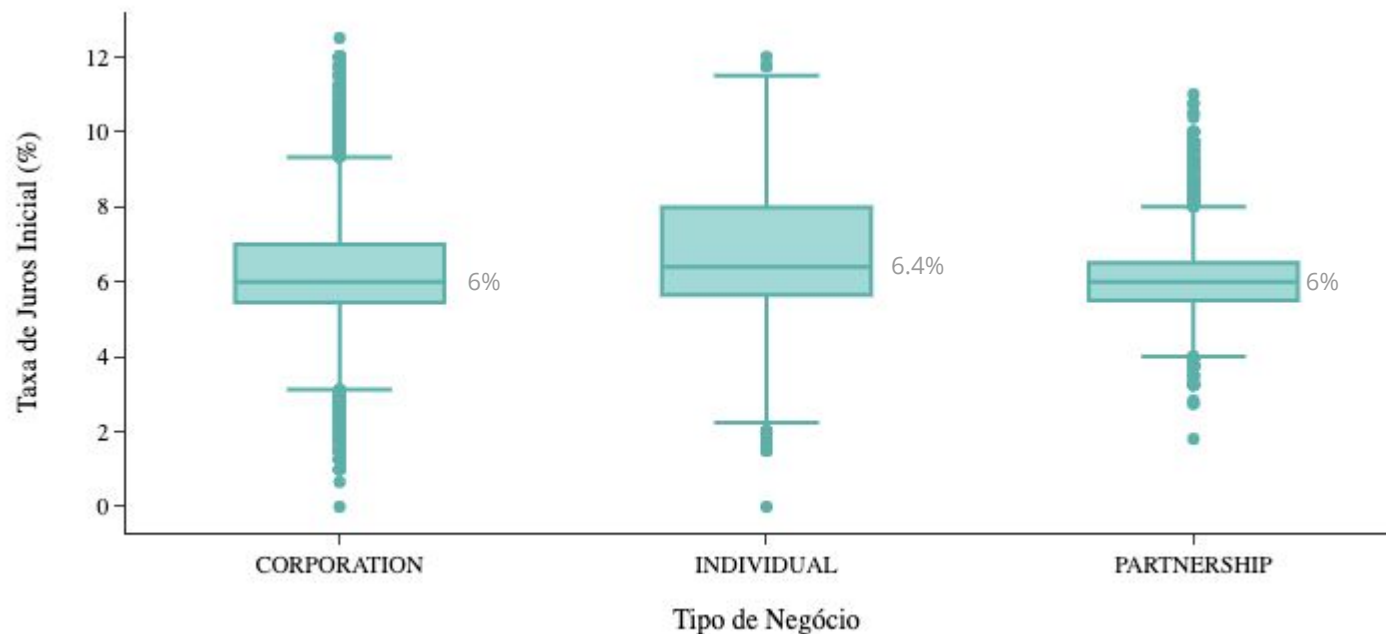
86% dos empréstimos foram de negócios do **tipo Corporação**. Já **empréstimos Individuais representaram 11.9% da base** e a **taxa de default foi a maior** (10.4% dos empréstimos não foram pagos), versus uma taxa de default de 5.7% para Parceria e 8.26% para Corporação.

4. Análise Exploratória de Dados

TAXA DE JUROS

29

Empréstimos por Tipo de Negócio



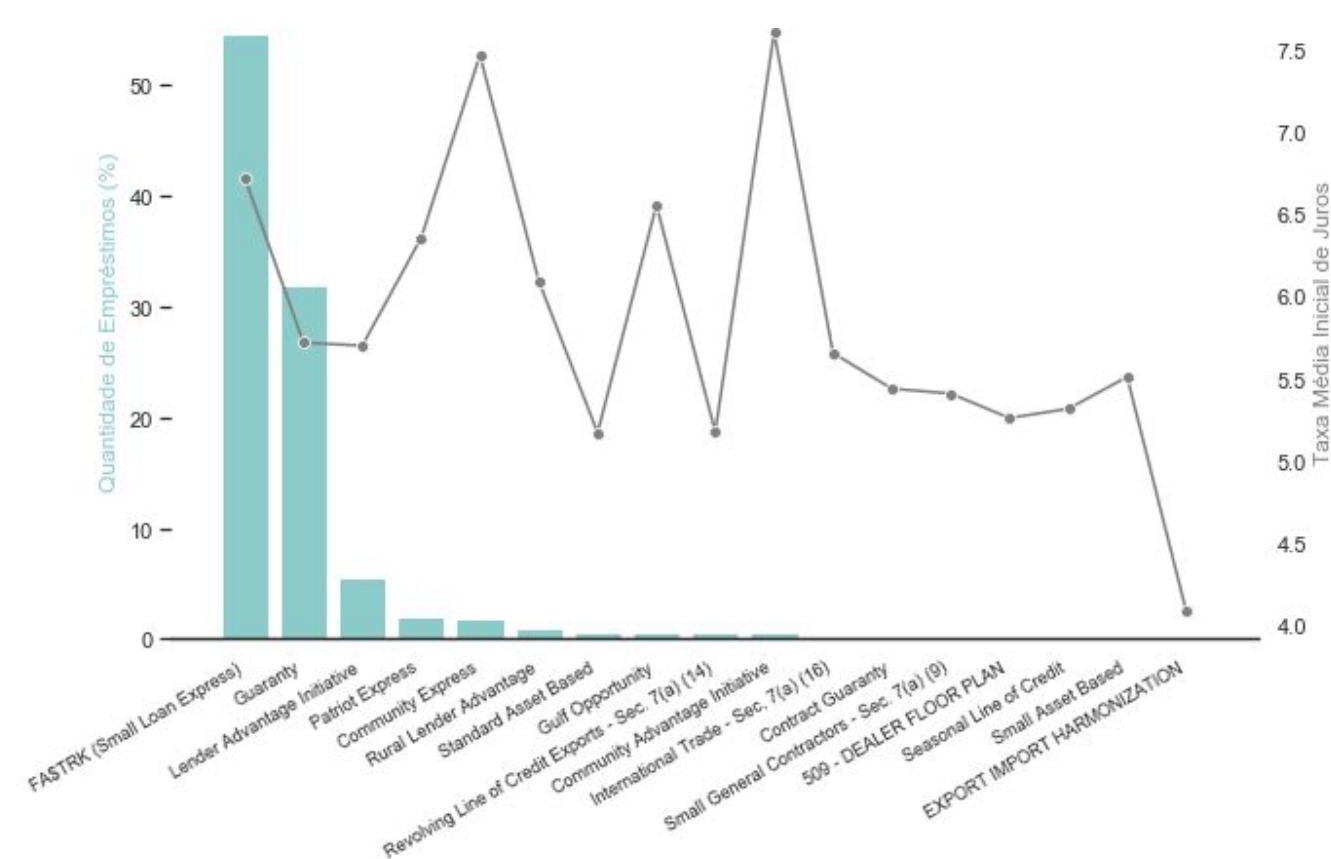
A taxa inicial de juros varia conforme o tipo de negócio. Vemos que a **distribuição dos juros para negócios do tipo Individual é maior do que Corporação (Corporations) e Parceria (Partnership).**

A mediana da taxa de juros foi de 6% para Parceria e Corporação e 6.4% para Individual.

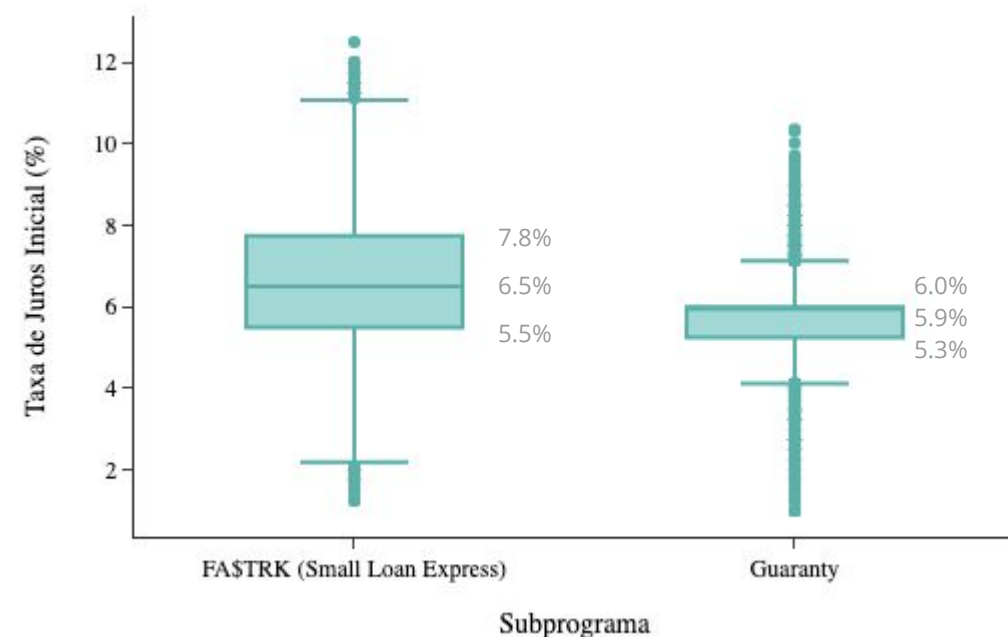
4. Análise Exploratória de Dados

TAXA DE JUROS

Representatividade e Taxa Média Inicial de Juros por Subprograma



Comparação Taxa de Juros Inicial por Subprograma



Os **subprogramas** do tipo **FA\$TRK (Small Loan Express)** [até USD 350k] e **Guaranty** [onde mutuário paga taxa para a SBA] representaram mais de 86% dos empréstimos. Vemos uma grande diferença entre as taxas de juros iniciais, variando entre 4% e 7.6%.

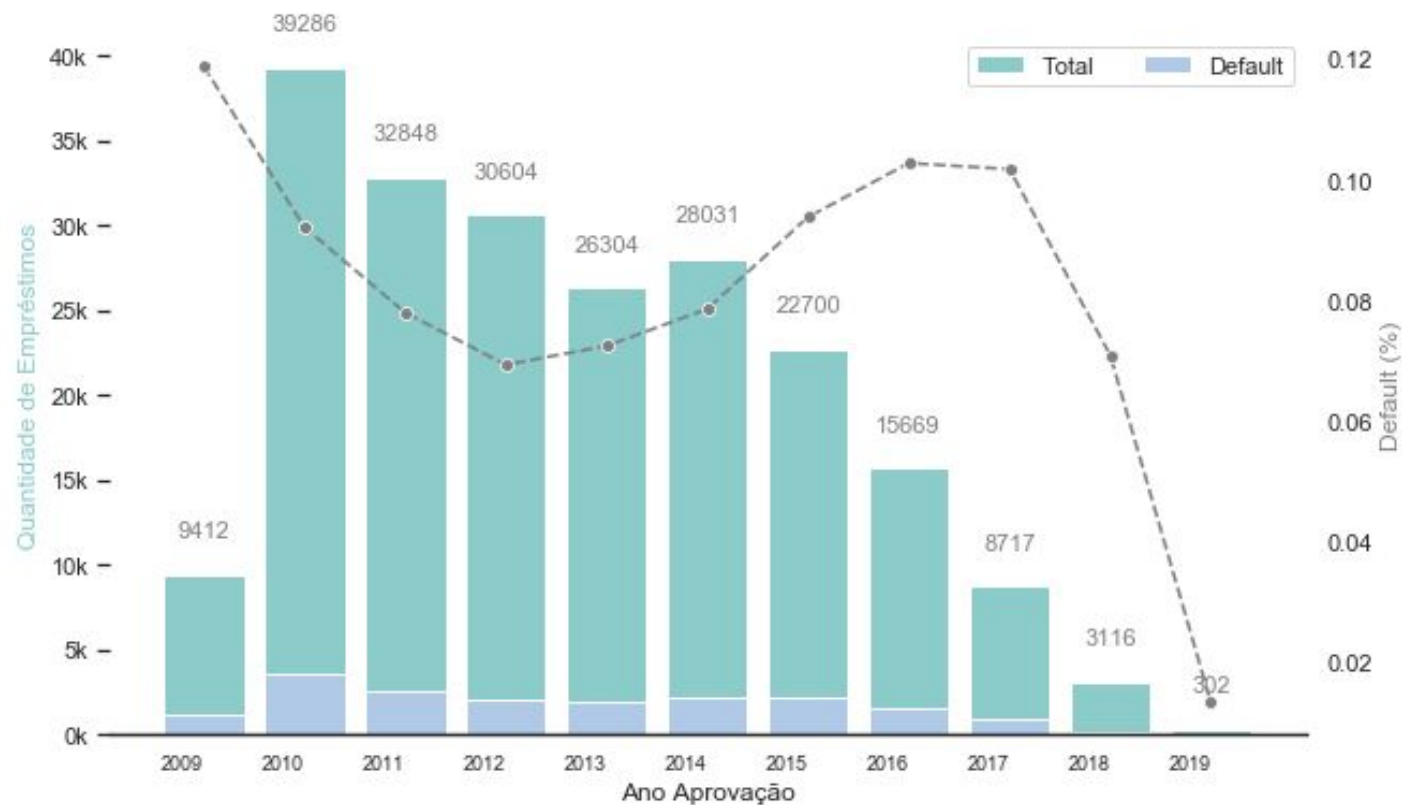
Pelo boxplot, comparando os dois principais subprogramas, vemos que a distribuiçõess das taxas de juros iniciais para subprograma *Small Loan Express* é maior que o *Guaranty*. A mediana de *Small Loan Express* é de 6.5% enquanto que *Guaranty* teve mediana de 5.9%.

4. Análise Exploratória de Dados

TAXA DE JUROS

31

Quantidade de Empréstimos e Default (%) por Ano



A base possui dados de **empréstimos de 01/10/2009 até 30/09/2019**. Podemos ver que o **volume de empréstimos anuais**, considerando anos completos, **caiu ao longo do tempo (de 2010 a 2012)**. As maiores taxas de default foram em 2009, muito provavelmente relacionado a crise de 2008.

Modelagem

O que fizeram os empréstimos darem default?

1

Abordagem adotada

Para o problema proposto, optou-se por criar um modelo preditivo usando um classificador.

Foram ajustadas 2 técnicas de modelagem estatísticas (Regressão Logística e Random Forest).

Usou-se *Filtro*, *Wrapper* e *Embedded* como métodos de seleção de variáveis e hiperparâmetros e cross validation para ambos os modelos.

2

Estratégia

Baseado no melhor modelo otimizado, usar as importâncias das features como forma de interpretar fatores que indicam a maior propensão ao default de um empréstimo.

5. Modelagem

MÉTODOS DE SELEÇÃO

Filtro

Seleciona as propriedades intrínsecas das variáveis, ou seja, a relevância das mesmas. Foi usado o **teste chi-quadrado**. Quanto maior o valor do chi-quadrado indica que a hipótese de independência de duas variáveis é incorreta.

Wrapper

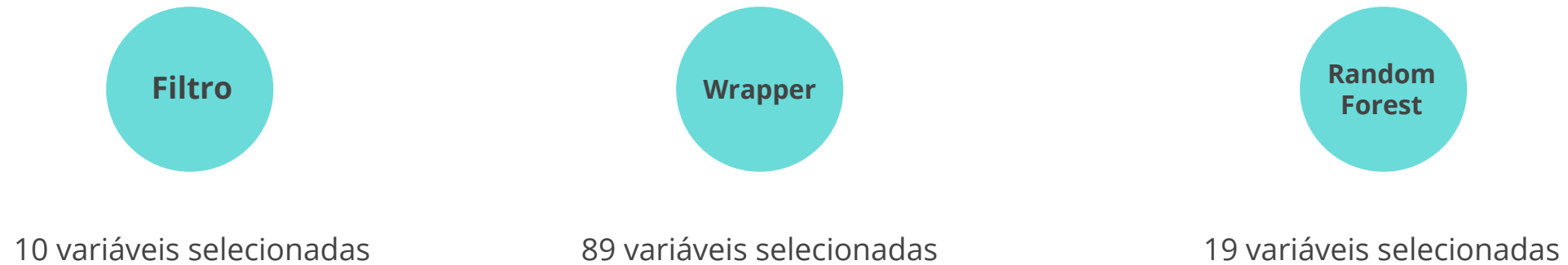
Avaliam um subconjunto de variáveis empregando uma estratégia de busca. Foi usado o **RFE (Recursive Feature Elimination)**. Começa com todas as variáveis do conjunto de treinamento e vai removendo features. Rankeia pela importância, descarta as menos importantes e re-treina o modelo, repetindo o processo.

Random Forest

Estratégia **baseada em árvore**. Variáveis que são selecionadas no topo das árvores são em gerais mais importantes que as selecionadas nos nós finais das árvores. De forma geral, os splits mais ao topo possuem maior ganho de informação.

5. Modelagem

MÉTODOS DE SELEÇÃO



Durante o desenvolvimento do projeto, foram utilizados como critérios de seleção para a modelagem a presença da variável selecionada por ao menos um dos três métodos acima

5. Modelagem

MÉTODOS DE SELEÇÃO

Variáveis que foram selecionadas pelos 3 métodos de seleção

Variáveis que foram selecionadas por pelo menos 2 métodos de seleção

Tabela 3. Variáveis selecionadas por pelo menos dois métodos: Chi-quadrado (chi2), Recursive feature elimination (RFE) e Random forest (RF)

Variáveis	chi2	RFE	RF
sameStateLoan	1	1	1
fklearn_feat__BankState==SD	1	1	1
TermInMonths	1	1	1
shareSBAGuaranteedTotalApproval	0	1	1
fklearn_feat__subpgmdesc==Community Express	1	1	0
fklearn_feat__NaicsSectorDescription==Retail Trade	0	1	1
fklearn_feat__NaicsSectorDescription==Professional, Scientific, and Technical Services	0	1	1
fklearn_feat__NaicsSectorDescription==Other Services (except Public Administration)	0	1	1
fklearn_feat__NaicsSectorDescription==Manufacturing	0	1	1
fklearn_feat__NaicsSectorDescription==Construction	0	1	1
fklearn_feat__NaicsSectorDescription==Accommodation and Food Services	0	1	1
fklearn_feat__DeliveryMethod==PLP	1	1	0
fklearn_feat__DeliveryMethod==COMM EXPRS	1	1	0
fklearn_feat__BusinessType==CORPORATION	0	1	1
fklearn_feat__BankState==OH	0	1	1
fklearn_feat__BankState==AL	1	1	0
daysFromApprovalToFirstDisbursement	0	1	1
SBAGuaranteedApproval	0	1	1
SBADistrictOffice	0	1	1
RevolverStatus	0	1	1
JobsSupported	0	1	1
InitialInterestRate	0	1	1
GrossApproval	0	1	1

A tabela 3 mostra as variáveis que foram selecionadas por ao menos dois dos métodos de seleção de variáveis.

5. Modelagem

MÉTODOS DE SELEÇÃO

Para cada técnica de estatística tradicional, realizamos a modelagem utilizando duas abordagens:

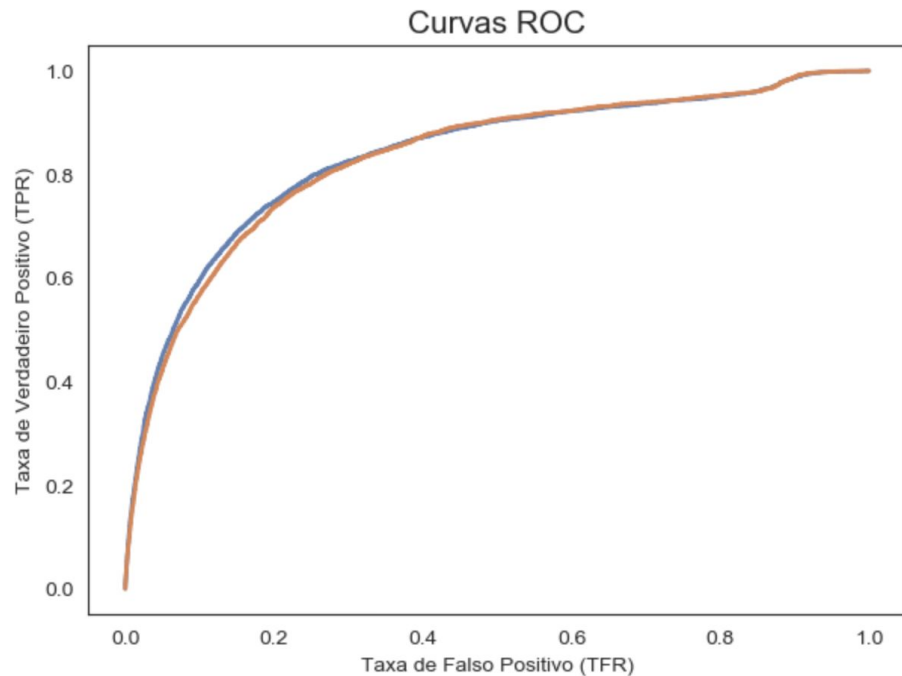
1. Modelo com todas as variáveis: 178 features
2. Modelo usando as variáveis que foram selecionadas por pelo menos um dos métodos de seleção de variáveis: 92 features

Modelagem

- Regressão Logística -

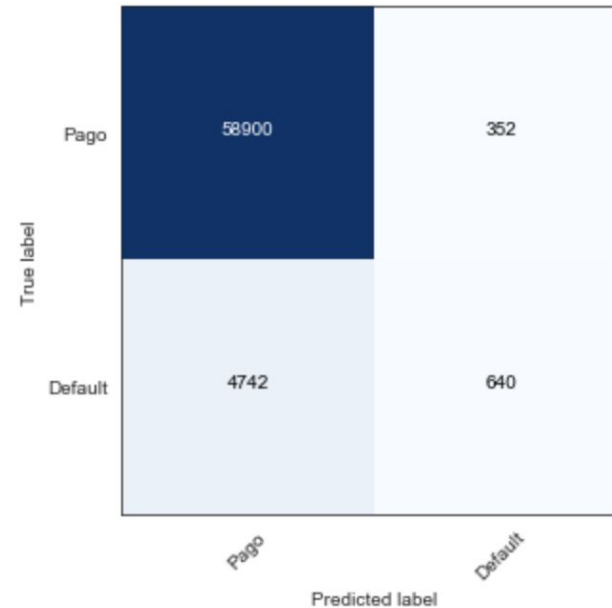
5. Modelagem | Regressão Logística (RL)

Utilizando Hiperparâmetros e Cross Validation (10 folds), os melhores parâmetros para a Regressão Logística com todas as variáveis (RL) foram: C = 100, penalty = L1. Com a seleção de variáveis, os mesmo parâmetros foram selecionados.

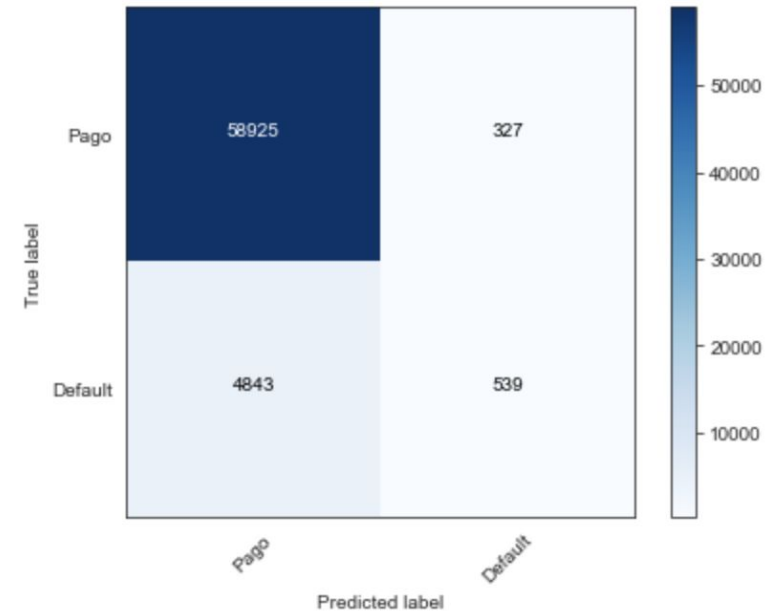


Modelo	AUC ROC
RL sem Feature Selection	0.837
RL com Feature Selection	0.832

Matriz de Confusão: RL sem Feature Selection



Matriz de Confusão: RL com Feature Selection



Modelo	Acurácia	Precision	Recall
RL sem Feature Selection	92.12	64.53	11.89
RL com Feature Selection	92.00	62.24	10.01

A **Regressão Logística (RL) com todas as variáveis (sem feature selection)** apresentou **melhores Precision e Recall** em relação a RL com Feature Selection. Em termos de AUC ROC ambos possuem áreas parecidas (~0.83). A precisão indica que o modelo com todas as variáveis acerta 64.5% dos previstos default. No entanto, o recall de 11.9% indica que o modelo ainda não é capaz de identificar a maior parte dos defaults.

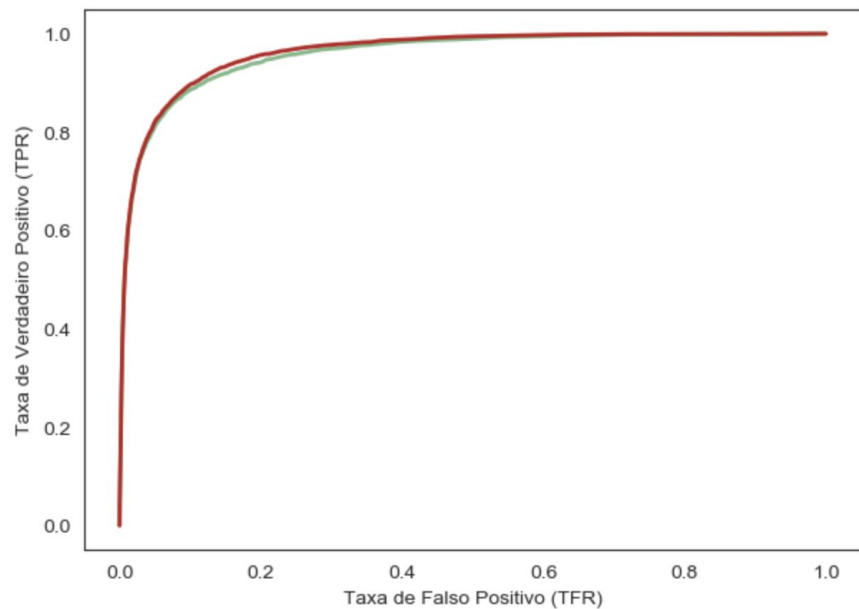
Modelagem

- Random Forest -

5. Modelagem | Random Forest (RF)

Os melhores parâmetros para a Random Forest com todas as variáveis foram: max_depth = 100, min_samples_leaf = 1, min_samples_split = 2 e n_estimators = 1000. Com a seleção de variáveis, o n_estimators foi 600 e demais parâmetros iguais.

Curvas ROC

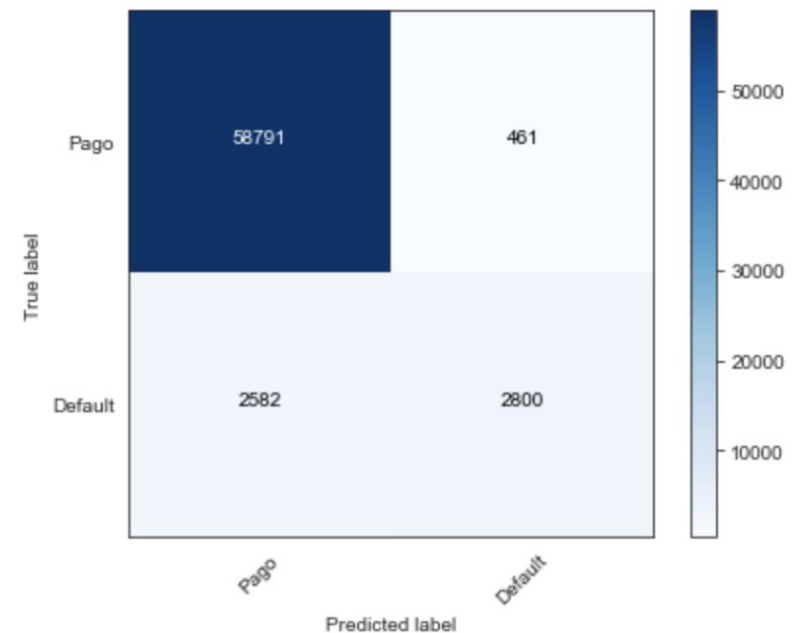
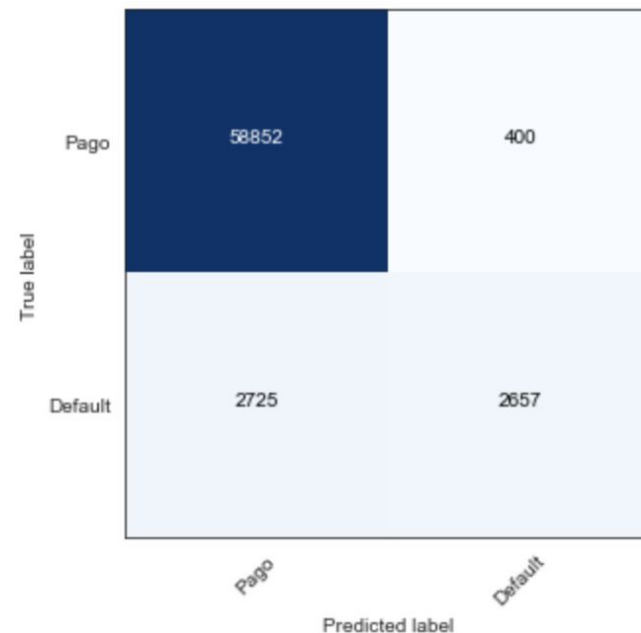


Modelo	AUC ROC
--------	---------

RF sem Feature Selection	0.958
--------------------------	-------

RF com Feature Selection	0.963
--------------------------	-------

Matriz de Confusão: RF sem Feature Selection Matriz de Confusão: RF com Feature Selection



Modelo	Acurácia	Precision	Recall
--------	----------	-----------	--------

RF sem Feature Selection	95.17	86.92	49.37
--------------------------	-------	-------	-------

RF com Feature Selection	95.29	85.86	52.03
--------------------------	-------	-------	-------

A **Random Forest com seleção de variáveis** apresentou uma **melhora** em relação ao modelo com todas as variáveis. A redução do número de variáveis e, portanto, da dimensionalidade dos dados teve um efeito positivo na modelagem.

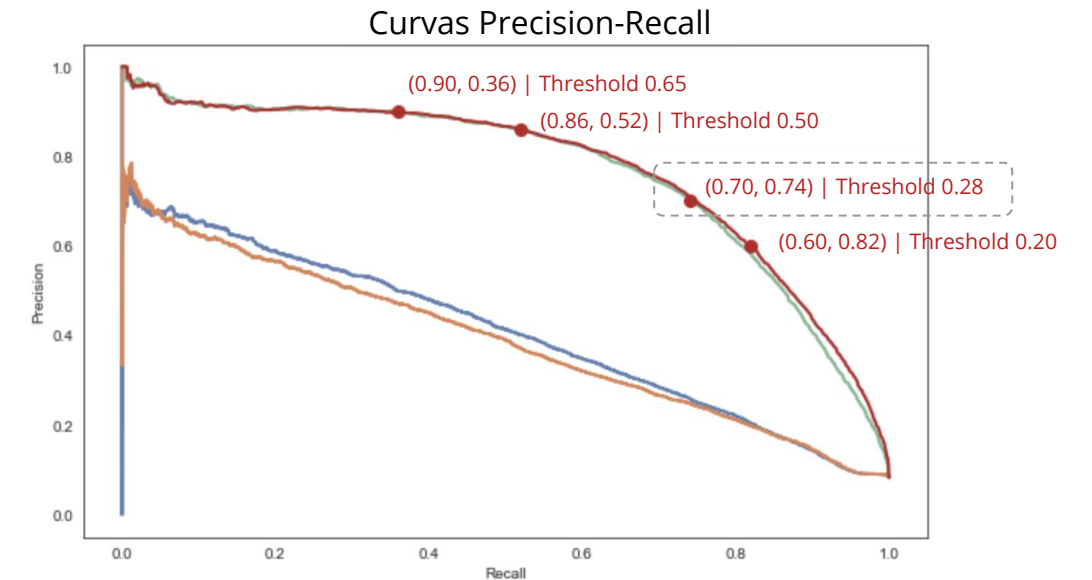
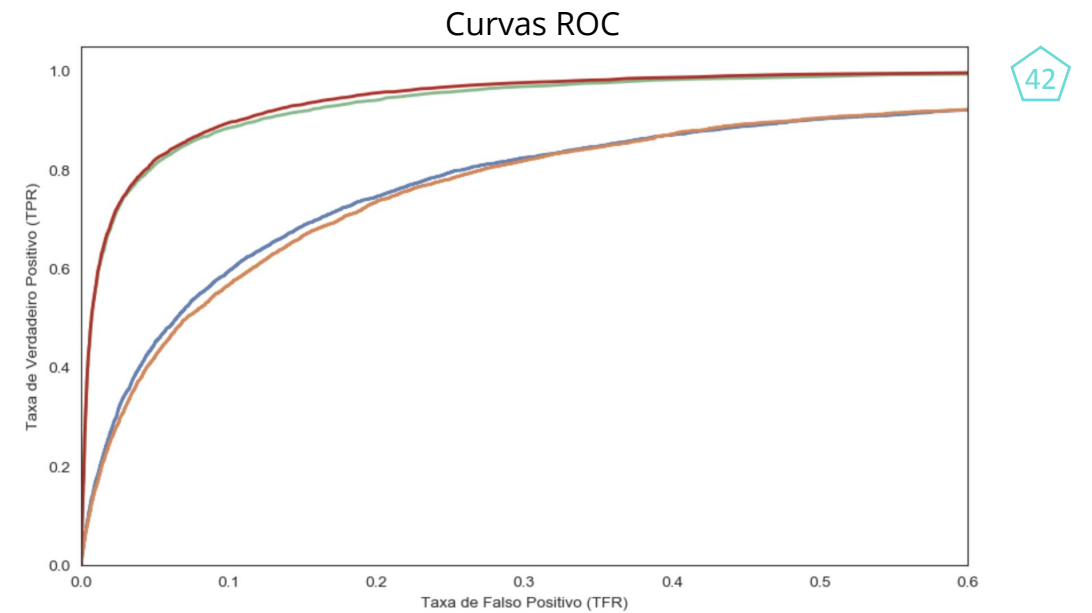
5. Modelagem | Comparativo

Modelo	Acurácia	Precision	Recall	AUC ROC
RL sem Feature Selection	92.12	64.52	11.89	0.837
RL com Feature Selection	92.00	62.24	10.01	0.832
RF sem Feature Selection	95.17	86.92	49.37	0.958
RF com Feature Selection	95.29	85.86	52.03	0.963

A **Random Forest (com Feature Selection)** teve as melhores métricas. Comparando com a melhor **Regressão Logística (sem feature selection)**, houve um **ganho de 23 p.p para a precisão e de 40 p.p para o recall**.

Considerando um threshold de 0.5, dos previstos pelo modelo default, a Random Forest acerta 86% (Precision). E dos empréstimos que realmente foram default, ele consegue identificar 52% (Recall).

Pela curva Precision-Recall é possível melhorar o Recall diminuindo o threshold. No cenário com precision de 0.70, é possível ter um Recall de 0.74, considerando um threshold de 0.28.



5. Modelagem | Interpretação

43

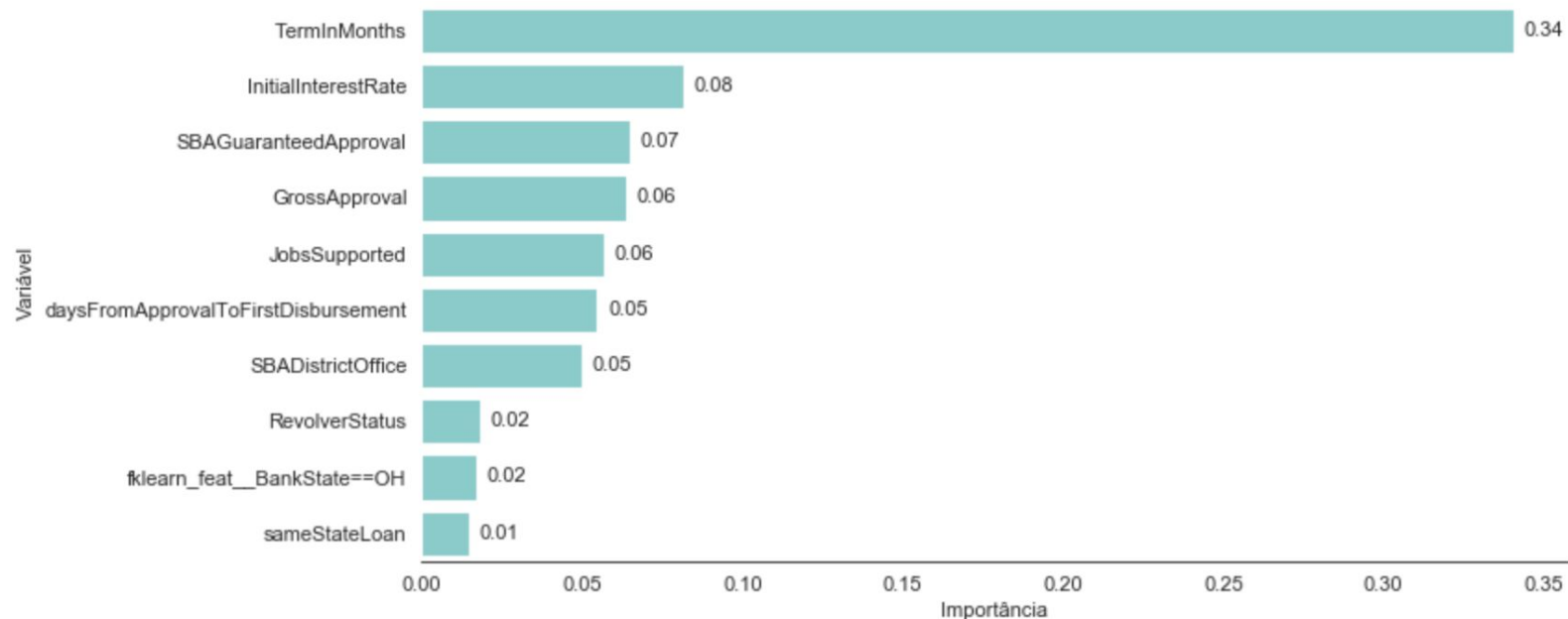
Dessa forma, **o melhor modelo** para predição de default foi a **Random Forest utilizando o método de Feature Selection**. O critério de seleção adotado foi utilizar as variáveis que apareceram ao menos uma vez por um método de feature selection (filtro, wrapper ou random forest).

Variáveis Temporais: Podemos ver que a duração do empréstimo (*TermInMonths*) é a principal variável, seguido da diferença de dias da aprovação até o primeiro pagamento (*daysFromApprovalToFirstDisbursement*)

Valores emprestados O valor garantido pelo SBA (*SBAGuaranteedApproval*) e total do empréstimo (*GrossApproval*) também foram identificadas como importantes. Além do tipo do empréstimo (*RevolverStatus*)

Variáveis geográficas Se o empréstimo é realizado no mesmo estado do banco prestador (*sameStateLoan*), o estado do banco de OH (Ohio) também foi identificado como relevante.

Principais Variáveis Random Forest com seleção de variáveis (Top 10)

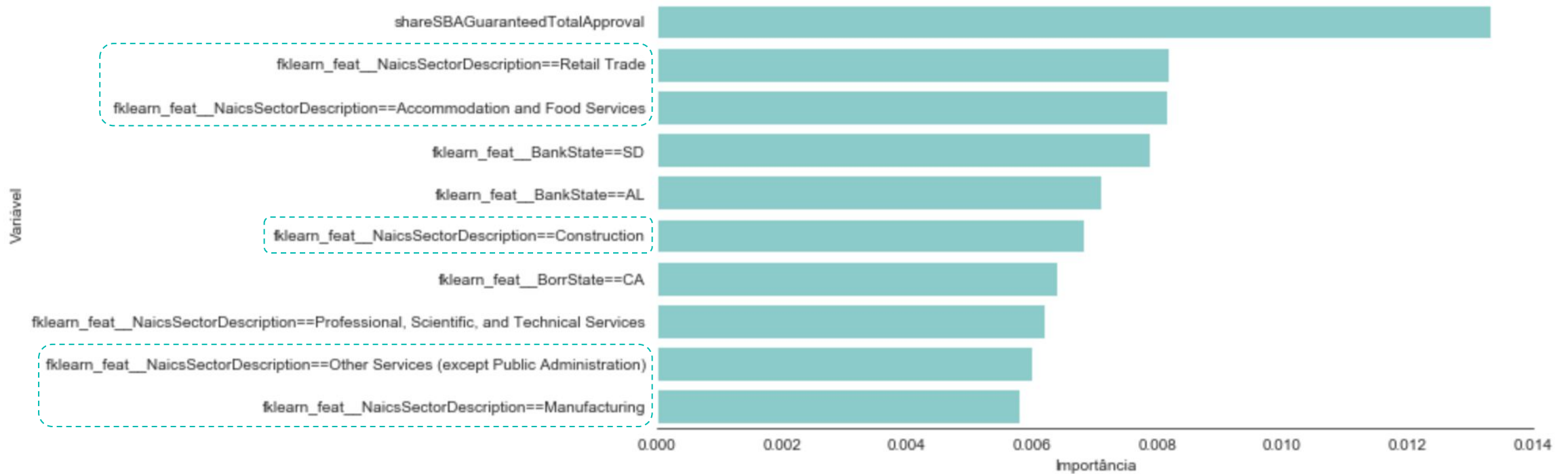


Interessante observar que a **quantidade de empregos gerados é fator que contribui para o default**. Possivelmente isso tem relação com o custo de contratação de novos empregados, o que faz o pagamento do empréstimo ser comprometido de alguma forma. Em termos de duração, quanto menor a duração maior a chance do empréstimo não ser pago. E como o valor do empréstimo é proporcional à duração, a variável *GrossApproval* também foi selecionada.

5. Modelagem | Interpretação

Extrapolando para as próximas 10 features da Random Forest, é possível verificar quais tipos de indústrias tem maior influência no default.

Random Forest com seleção de variáveis (Ranking de 11 a 20)



Interessante observar que alguns setores têm maior importância na variável alvo: **Acomodação e alimentação, comércio (Retail Trade), construção e manufatura**. O share garantido pelo SBA (*shareSBAGuaranteedTotalApproval*) também é importante, o que é esperado pois ela anda junto com o valor total aprovado.

Qual o impacto para os negócios?

6. Conclusão | Impacto ao negócio

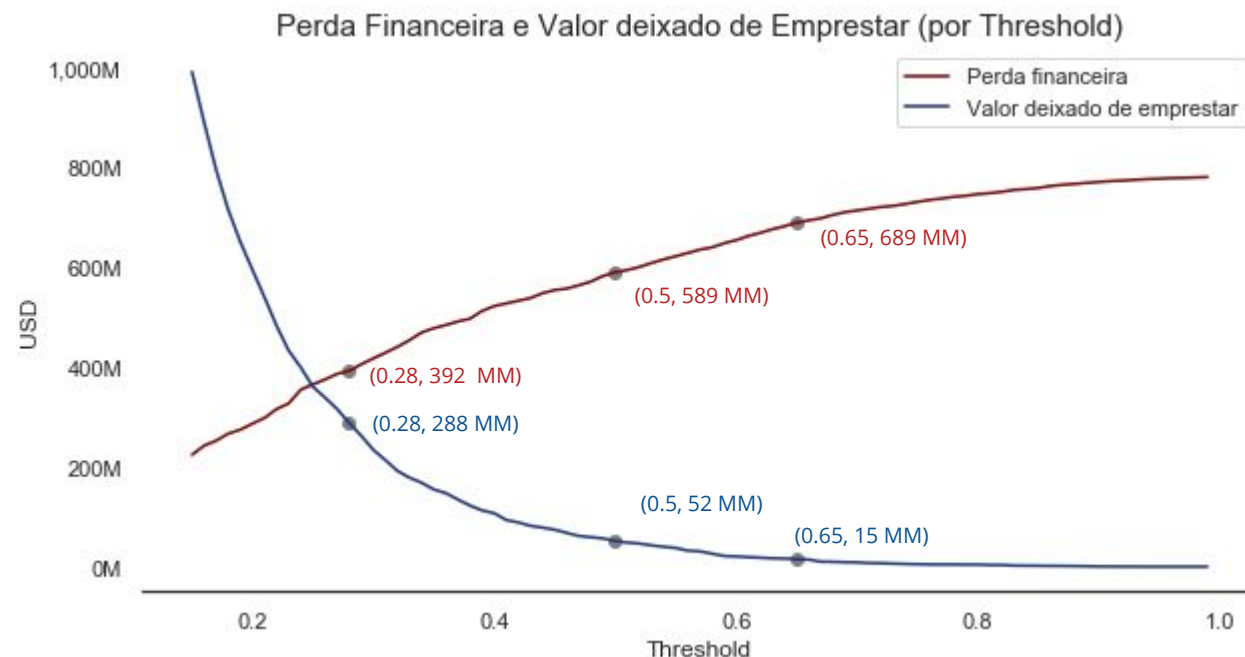
Em termos de negócio, é importante analisar qual será o impacto financeiro da decisão de negócio. Temos duas principais perspectivas para avaliar:

1. Quanto, em milhões (USD), serão as perdas financeiras? Ou seja, falso negativos dado pelo modelo.

2. Quanto, em milhões (USD), serão deixados de emprestar? Ou seja, falso positivos dados pelo modelo.

No período de teste analisado o **SBA garantiu USD 15.19B** de um total emprestado de USD 20.37B.

O gráfico ao lado mostra qual seria a **perda financeira** (falsos negativos) e o **valor deixado de emprestar** (falsos positivos) para diferentes thresholds. Baseados nos thresholds analisados anteriormente na curva PR (Precision-Recall), destacamos qual seria o impacto financeiro para os thresholds 0.28, 0.5 e 0.65.



Se adotarmos o threshold de 0.28, a perda financeira da SBA seria de USD 392 MM (1.89% do valor total garantido), porém ela deixaria de emprestar USD 288 MM. Ao aumentar o threshold para 0.5, o SBA aumentaria sua perda financeira para USD 589 MM (3.9% do valor total garantido). Em compensação o valor deixado de emprestar cairia para USD 52 MM, isso representa um impacto total de 2763 empresas (4.3% dos empréstimos) que seriam beneficiadas pelo empréstimo e honrariam seus compromissos.

6. Conclusões

- O modelo **Random Forest com Feature Selection** apresentou melhorias significativas em relação à Regressão Logística, com um **ganho de 23 p.p para a precisão e de 40 p.p para o recall**.
- O **método de seleção de variáveis** trouxe uma **melhoria de 2.6 p.p para o Recall** da Random Forest.
- A **duração do empréstimo** bem como a **taxa de juros inicial** são variáveis importantes para **prever o default**.
- A quantidade de empregos gerados também tem peso. Possivelmente há alguma relação com o custo de contratação de novos empregados, o que faz o pagamento do empréstimo ser comprometido de alguma forma.
- Determinados setores têm maior importância na variável alvo: Acomodação e alimentação, comércio (Retail Trade), construção e manufatura. Franquia também aparece, indicativo que Franquias possuem melhor estruturação financeira.

6. Conclusões

- Num contexto de empréstimos, a combinação entre Precision e Recall é importante. A precisão de 86% indica que, dos previstos default, o modelo acerta 86% das vezes. Já o Recall de 52% indica que, dos empréstimos que deram default, quanto o modelo identificou como sendo default. Ou seja, o modelo não conseguiu classificar corretamente 48% dos empréstimos que deram default. Essa métrica ainda está longe do ideal.
- Assim, avaliou-se a variação do threshold. Variando de 0.5 para 0.28, temos que a precisão cai para 70% (-16 p.p) mas em compensação o recall aumenta para 74% (+22 p.p). Em termos financeiros, o valor de **perda financeira evitada para o SBA seria de USD 197 MM** (de USD 589 MM para USD 392 MM). O principal trade-off dessa decisão seria que 2763 empresas teriam o empréstimo negado indevidamente (~4.3% do total de empréstimos).

LABDATA FIA – Laboratório de Análise de Dados

49



Unidade Pinheiros



Unidade Paulista

