Universidade Federal do Paraná
Setor de Ciências Exatas
Departamento de Estatística
Programa de Especialização em *Data Science* e *Big Data*

Daniel Basso Ribas

Mineração de texto e aprendizado supervisionado na análise de processos sobre financiamento agropecuário

Curitiba
2019

Daniel Basso Ribas

Mineração de texto e aprendizado supervisionado na análise de processos sobre financiamento agropecuário

Monografia apresentada ao Programa de Especialização em *Data Science* e *Big Data* da Universidade Federal do Paraná como requisito parcial para a obtenção do grau de especialista.

Orientador: Prof. Cesar Augusto Taconeli

Curitiba
2019

Mineração de texto e aprendizado supervisionado na análise de processos sobre financiamento agropecuário

Text mining and supervised learning applied in the analysis of rural financing processes

Daniel Basso Ribas

Resumo

As agências e um departamento especializado de uma Instituição Financeira se comunicam para aprovação de operações de financiamento rural; o departamento especializado pode apontar problemas na documentação do financiamento, que deverão ser corrigidos pela agência. Essa ação é denominada ocorrência. A ocorrência por sua vez pode ser contestada pela agência, caso ela entenda que a documentação esteja sim correta. Tomando como base um conjunto de dados gerados a partir dessa comunicação, foram analisados quais fatores podem levar uma ocorrência ter maior probabilidade de ser acatada pelo departamento especializado. A análise dos dados utilizou-se de duas metodologias principais: com os textos das contestações foram criados *word embeddings* através da técnica *word2vec*; a partir deles foi construído um vetor de características para cada contestação que foi alimentado em modelos de *machine learning*. As demais variáveis coletadas e características extraídas manualmente dos textos das contestações foram utilizadas em modelos lineares generalizados com efeitos mistos (GLMM). Os modelos separadamente apresentaram um baixo poder preditivo, com uma acurácia balanceada de 57%. A combinação dos modelos por meio de *ensemble* melhorou a acurácia balanceada para 68%. Embora com uma acurácia não muito alta, o GLMM mostrou que as características extraídas manualmente do texto foram significativas para predição do resultado da contestação e permitiu interpretações válidas em termos das probabilidades. O fato das duas modelagens distintas apresentarem resultados semelhantes serviu como forma de validação dos modelos.

Palavras-chave: processamento de linguagem natural, aprendizado de máquina, modelos lineares generalizados mistos, *ensemble*

Abstract

Agencies and a specialized department of one Financial Institution communicate for approval of rural financing operations; the specialized department may point out problems in the documentation of the financing, which should be corrected by the agency. This action is called an occurrence. The occurrence in turn can be contested by the agency if it understands that the documentation is correct. Based on a set of data generated from this communication, we analysed the factors that may cause an occurrence to be more likely to be complied with by the specialized department. Data analysis was based on two main methodologies: word embeddings were created using the word2vec technique; from them a vector of characteristics was constructed for each contestation that was fed in machine learning models. The other variables collected and characteristics extracted manually from the texts of the contestations were used in generalized linear models with mixed effects (GLMM). The models separately showed a low predictive power, with a balanced accuracy of 57 %. The combination models through ensemble improved the balanced accuracy to 68 %. Although not very accurate, the GLMM showed that the manually extracted characteristics of the text were significant for predicting the outcome of the contestation and allowed for valid interpretations in terms of probabilities. The fact that the two models presented similar results served as a validation of the models.

Keywords: natural language processing, machine learning, generalized linear mixed models, ensemble

1. Introdução

O agronegócio é historicamente fundamental para a economia brasileira, representando mais de 20% do PIB nacional em 2017 [1] e apresentando números expressivos ao longo dos anos [2]. Para fomentar essa atividade, alguns dos principais bancos do país possuem linhas de financiamento especificas, com taxas e prazos atraentes. Uma Instituição Financeira em particular é líder neste segmento, e conta com um departamento especializado para analisar as propostas de novos financiamentos. A rede de agências envia a proposta para esse departamento, que confere toda a documentação e verifica sua aderência às normas internas e externas. Caso algo não esteja conforme os

normativos, o departamento abre uma diligência para a agência de origem, informando os requisitos que não estão de acordo e pedindo regularização. A agência, por sua vez, pode contestar essa diligência, caso julgue que a documentação esteja correta. Por fim, a diligência pode ser acatada ou não pelo departamento que realiza a análise. Toda essa troca de mensagens gera uma rica base de dados, que até o presente estudo não fora analisada com profundidade.

O objetivo deste trabalho foi identificar quais são os fatores associados com que uma contestação seja acatada ou não e ao mesmo tempo prever qual a probabilidade de uma nova contestação ser acatada.

A Seção 2 descreve o conjunto de dados utilizado e suas particularidades; a Seção 3 descreve as ferramentas computacionais utilizadas bem como as técnicas empregadas na análise; a Seção 4 mostra como foi feita a modelagem de dados e os resultados obtidos; por fim na Seção 5 é discutida a relevância dos resultados para o problema apresentado.

2. Conjunto de dados

Os dados foram fornecidos pela Instituição Financeira e compreendem 11340 contestações de ocorrências capturadas no período de 02/01/2018 até 05/04/2019. Destas, 8983 são contestações não acatadas e 2358 acatadas. Dezenove variáveis foram disponibilizadas ao todo, sendo descritas na Tabela 1. De maneira simplificada, as variáveis podem ser divididas em quatro categorias:

1. O texto das contestações das ocorrências;

2. Variáveis construídas a partir do texto da contestação: se o texto cita as Instruções Normativas; se o texto cita o dossiê eletrônico de documentos do cliente; se o texto cita outras siglas referentes aos sistemas do banco; tamanho do texto.

3. Outras variáveis referentes ao processo: localização da agência; equipe que está analisando o financiamento; valor da operação; dias que a operação está em análise, entre outras.

4. A variável resposta: se a contestação foi acatada ou não.

Para as análises, 80% das observações foram separadas para treino (9073) e 20% para teste (2265), mantendo-se a proporção de acatamento igual ao observado no conjunto de dados. Para o *ensemble* apresentado na Seção 4.3, o conjunto de teste atua como se fosse a totalidade dos dados; assim ele foi novamente divido em 80% para treino (1813) e 20% teste (453).

Algumas variáveis incluídas modelo contém informações sensíveis, como por exemplo o valor da operação, os nomes das linhas de crédito e das equipes que analisam as operações e o texto das contestações. Os valores dessas variáveis foram suprimido de eventuais gráficos e tabelas com o objetivo de preservar a informação original. Os textos das contestações em particular também não serão apresentados, apenas os resultados obtidos a partir deles. De todo modo, a apresentação dos resultados na Seção 4 mostra que muitas das variáveis acabaram sendo descartadas na modelagem por não apresentarem significância.

3. Meteriais e Métodos

3.1. Materiais

As linguagens R [3] e Python [4] foram utilizadas para a análise dos dados. O R se mostra uma excelente linguagem para a manipulação dos dados de modo geral, pois possui bibliotecas que permitem a utilização de comandos muito similares ao SQL; além disso, dispões das principais bibliotecas para ajuste de modelos de regressão. O Python ,por sua vez, dispõe de bibliotecas muito úteis para o processamento de linguagem natural. Os principais pacotes empregados estão listados conforme segue:

R:

- tidyverse [5]: conjunto de pacotes que permite a manipulação de dados utilizando uma sintaxe intuitiva derivada do SQL. Também possui uma estrutura de tabelas (denominada tibble) mais eficiente que a estrutura de tabelas padrão do R (data frames);
- ggplot2 [6]: construção de visualizações através da gramática dos gráficos;
- lme4 [7]: ajustes de modelos de regressão com efeitos mistos;
- caret [8]: ajustes de modelos de *machine learning* com validação cruzada.

Python:

- spacy [9]: tratamento dos textos (remoção de alguns nomes próprios, remoção da acentuação e transformação dos textos para letras minúsculas);

Variável	Descrição
ORIGEM_DA_CONTESTACAO	Se a contestação veio de um agência ou superintendência
TEXTO_DA_CONTESTACAO	Texto da contestação
VALOR	Valor da operação de crédito
DESTINO	Se a operação foi expedida ou devolvida
PROCESSO	Caracterizador interno do tipo da operação
ESTUDAR	Caracterizador interno do tipo da operação
LINHA	Linha de crédito rural da operação
CANAL	Canal de contratação da operação
EQUIPE	Equipe responsável pela operação de crédito
UF	Unidade da federação onde a operação foi contratada
SUPER	Superintendência regional de onde a operação foi contratada
PERMANENCIA_TOTAL	Núm. de dias em que a operação esteve no departamento responsável pela análise
PERMANENCIA_DILIGENCIA	Núm. de dias em que a operação esteve aguardando resposta da agência
PERMANENCIA_ANALISE	Núm. de dias em que a operação permaneceu em analise
CITA_IN	Se o texto da contestação cita as instruções normativas do banco
CITA_DOSSIE	Se o texto da contestação cita o dossiê eletrônico de documentos do cliente
CITA_SIGLAS	Se o texto da contestação cita siglas referentes aos sistemas do banco
TAMANHO	Tamanho do texto da contestação
CENOP_ACATOU	Se a contestação da ocorrência foi acatada pelo departamento especializado

Tabela 1: Descrição das variáveis presentes no conjunto de dados disponibilizado pela Instituição Financeira

- gensim [10]: criação de *word embeddings* com *word2vec*;

3.2. Métodos

O texto da contestação foi submetido a técnica *word2vec* [11] para gerar vetores de características, que posteriormente foram inseridos em modelos de aprendizado de máquina como *Support Vector Machine* [12] e *Random Forest* [13]. Essa abordagem foi utilizada pois os vetores gerados por *word2vec* tem uma alta dimensionalidade (no caso em particular, 300 dimensões) e não possuem uma interpretabilidade clara; assim os modelos de *machine learning* são candidatos naturais para a tarefa, pois seu foco está na predição de resultados e apresentam menos complicações no seu ajuste quando se faz a utilização de muitas variáveis preditoras. Nos ajustes com esse tipo de modelos foi utilizada a técnica de validação cruzada [14] (5-*fold*, sem repetições), para se obter uma estimativa de variabilidade em relação aos parâmetros calculados.

Outras variáveis mais "simples"(localização da agência, equipe que realizou a análise, tipo de financiamento, etc.), bem como características extraídas manualmente dos textos foram utilizadas em um modelo linear generalizado de efeitos mistos [15]. Isso foi feito pois a regressão permite a interpretação dos resultados em função dos parâmetros de maneira mais simples e objetiva. A inclusão de efeitos aleatórios permite calcular não só a variabilidade entre esses efeitos como também uma possível correlação dos resultados de um mesmo efeito, em relação à resposta. Além disso, algumas variáveis categóricas possuem muitas categorias (como por exemplo UF) e se considerou mais interessante agrupar esses efeitos em um termo do modelo do que em inúmeras categorias. Por fim, foi possível comparar os resultados das duas abordagens de maneira a verificar se as variáveis construídas a partir da contestação possuíam um poder preditivo similar à extração de características do texto de maneira automatizada.

A partir das probabilidades de acatamento calculadas por cada ajuste, é possível construir a *Receiver Operating Characteristic Curve* [16](curva ROC), uma importante ferramenta para visualizar o seu poder preditivo. Com base nos cálculos realizados para construir a essa curva, é computado também o índice de Youden [17], que define um ponto de corte ótimo para predição das categorias da variável resposta. A qualidade preditiva do ajuste, de modo geral, é feita através da área abaixo da curva ROC (AUC).

Outra medida interessante para verificação da capacidade de predição de um modelo é a acurácia balanceada (ACUB). Essa medida é expressa como:

$$\text{ACUB} = \frac{(\text{VP}/\text{P} + \text{VN}/\text{N})}{2},$$

onde P é o número de observações positivas (no caso, sim); N é o número de observações negativas (no caso, não); VP é o número de observações corretamente classificadas como positivas; VN é o número de observações corretamente classificadas como negativas. Essa medida é útil pois representa a acurácia, porém penalizada pelo fato do conjunto de dados possuir a variável resposta desbalanceada, como é o caso apresentado.

A abordagem utilizando curva ROC, matriz de confusão e acurácia balanceada foi realizada todos para os ajustes das seções seguintes, como forma de comparação entre os modelos.

4. Resultados

Antes de se iniciar à modelagem, foi feita uma análise exploratória e gráfica do conjunto de dados, com a finalidade de fornecer uma percepção inicial da sua organização e da relação entre a variável resposta com as demais. Dessa análise já foi possível perceber que muitas das variáveis candidatas a modelagem poderiam ser descartadas, pois as suas distribuições marginais em relação ao acatamento das contestações eram praticamente idênticas. Exemplos dessa situação podem ser observados na Figura 1, que apresenta a relação da resposta com o valor da operação e com o número de dias que a operação ficou em análise.

Assim, a modelagem através de modelos de regressão focou nas variáveis extraídas manualmente do texto. Além dessas, as únicas outras a serem consideradas foram UF e EQUIPE, por motivos que serão apresentados na Seção 4.1.

4.1. Modelagem através de modelos de regressão

Entre a grande variedade de modelos de regressão existentes, a classe dos modelos lineares generalizados de efeitos mistos foi escolhida para essa tarefa (GLMM). Por serem generalizados, permitem a modelagem de variáveis respostas que não possuem distribuição normal; é o caso apresentado, pois uma variável que pode assumir duas possibilidades (sim ou não) possui distribuição de Bernoulli. Além disso, por serem de efeitos mistos, permitem a inclusão ao mesmo tempo de efeitos fixos e aleatórios no modelo. Conforme apresentado na seção 3.2, um dos motivos da utilização dos efeitos aleatórios foi modelar variáveis explicativas com muitas categorias, pois permitiu substituir os $n-1$ parâmetros que deveriam ser estimados caso o efeito considerado fosse fixo, sendo n o número de categorias, por apenas um termo de efeito aleatório, deixando o modelo mais enxuto e fácil de interpretar. Também possível calcular a variância entre eles e uma possível correlação da resposta dentro para um particular efeito.

Diante do fato de que pela inspeção gráfica já se observar que muitas variáveis explicativas não possuíam relação com a resposta, na modelagem era esperado que vários efeitos fossem não significativos. Assim, após muitos ajustes e comparações, apenas UF

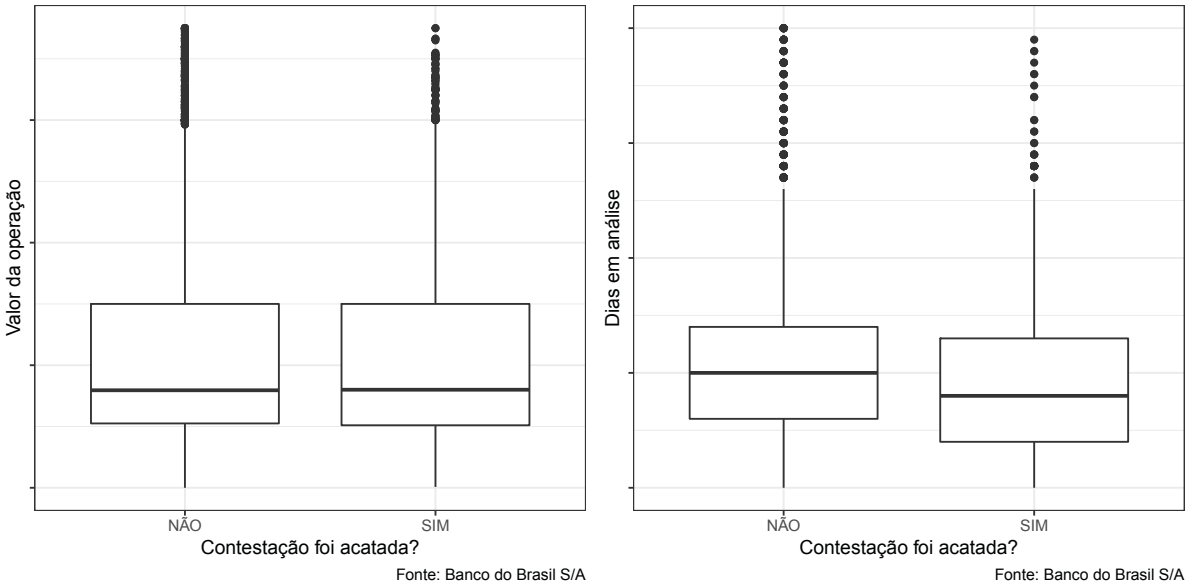

Figura 1: Distribuição do valor da operação e tempo total da análise em função resultado da contestação.

e EQUIPE apresentaram um comportamento suficientemente heterogêneo dentro do seus grupos em relação ao acatamento das contestações para justificar sua inclusão no modelo. Além dessas, todas as variáveis extraídas do texto possuíram significância, o que confirmou que a extração manual de características se baseou em hipóteses coerentes com a realidade.

Assim, considerando $i = \{1, 2, ..., \text{n.º de observações}\}$; $j = \{1, 2, ..., \text{n.º de equipes}\}$ e $k = \{1, 2, ..., \text{n.º de UFs}\}$, o acatamento ou não de uma contestação (y_{ijk}) pode ser descrito como:

$$y_{ijk} \mid \underline{x_i}, u_j, v_k \sim \text{Bernoulli}\left(\pi_{ijk}\right),$$

onde π_{ijk} representa a probabilidade de uma contestação ser acatada, que é expressa por:

$$\pi_{ijk} = \frac{1}{1 + \exp\left(-\eta_{ijk}\right)},$$

e η_{ijk} é o preditor linear, que por sua vez é calculado do seguinte modo:

$$\eta_{ijk} = \underline{x_i'\beta} + u_j + v_k,$$

onde $x_i'\beta$ representa a incorporação dos efeitos fixos; u_j o efeito aleatório para a j-ésima equipe e v_k o efeito aleatório para a k-ésima UF . Tanto u_j quanto v_k são independentes e identicamente distribuídos como $\text{N}\left(0, \sigma^2_{u_j}\right)$ e $\text{N}\left(0, \sigma^2_{v_k}\right)$, respectivamente [18].

Desta forma, o modelo final contou com as variáveis UF e EQUIPE como efeitos aleatórios; e CITA_IN, CITA_DOSSIE, CITA_SIGLAS e TAMANHO como efeitos fixos. Foi também incluído um termo para a interação CITA_IN:CITA_DOSSIE, que embora não tenha apresentado significância estatística neste modelo em específico, apresentou em outros modelos prévios e faz sentido para o problema. As estimativas dos efeitos e sua significância estão apresentados na Tabela 2.

De modo geral, os resultados apresentados são bastante coerentes com as hipóteses levantadas: citar as instruções normativas e o dossiê do cliente aumenta a probabilidade de uma contestação ser acatada, pois é um indício que a argumentação foi melhor embasada. Outro fator interessante está relacionado com o tamanho do texto: textos menores tem uma maior probabilidade de serem acatados. Isso pode estar relacionado com o fato de textos mais curtos serem mais sucintos, apresentando uma maior "confiança"de quem os escreveu sobre a veracidade da contestação. Os efeitos aleatórios para cada EQUIPE e UF foram extraídos também, mas não serão apresentados pelo sigilo da informação.

Efeitos fixos	Efeito	IC(95%)	P-valor
Intercepto	-1,59	(-1,73; -1,43)	0
CITA_DOSSIE	0,39	(0,24; 0,53)	0
CITA_IN	0,45	(0,32; 0,58)	0
CITA_SIGLAS	-0,50	(-0,81; -0,20)	0,001
TAMANHO	-0,10	(-0,16; -0,04)	0
CITA_DOSSIE: CITA_IN	-0,22	(-0,51; 0,06)	0,133
Efeitos aleatórios			
EQUIPE	0,49	(0,39; 0,62)	-
UF	0,10	(0,01; 0,22)	-

Tabela 2: Estimativa dos efeitos do modelo de regressão final, seu respectivo intervalo de confiança de 95% e significância estatística (P-valor). Os efeitos fixos representam um acréscimo ao intercepto do modelo, enquanto efeitos aleatórios são expressos em função de um desvio padrão em torno desse mesmo intercepto.

A curva ROC específica para esse ajuste pode ser observada na Figura 3.A; ela apresenta para esse modelo um índice de Youden de 0,19. Isso significa que observações com probabilidade de acatamento calculada pelo modelo superior a 0,19 vão ser preditas como acatadas e observações com probabilidade menor a 0,19 vão ser preditas como não acatadas. Com isso foi criada uma matriz de confusão, conforme descrito na Tabela 3, que confronta as categorias da resposta reais (observadas) com aquelas preditas pelo modelo e consequentemente permite o cálculo de sua acurácia balanceada.

Para esse ajuste, a acurácia balanceada foi de 58% e a área abaixo da curva ROC (AUC) foi 0,61, apresentando uma fraca capacidade de predição.

Em um primeiro momento esse resultado pode causar certo desconforto, uma vez que fica abaixo do que

GLMM

		Predito	
		Sim	Não
Real	Sim	935	158
	Não	861	313

Ponto de corte: 0,19

Acurácia balanceada: 58%

Tabela 3: Matriz de confusão para o modelo de regressão ajustado. O ponto de corte ótimo foi escolhido através do índice de Youden.

CITA_IN	CITA_DOSSIE	CITA_SIGLAS	TAMANHO	UF	EQUIPE	Probabilidade
0	0	0	540	SP	EQUIPE20	0,05
1	0	1	250	SP	EQUIPE20	0,10
1	1	0	100	SP	EQUIPE20	0,25

Tabela 4: Exemplo de probabilidades de acatamento de uma contestação para cenários hipotéticos.

se espera em um bom ajuste. Porém, se analisarmos as predições em termos das probabilidades, é possível extrair resultados práticos e interessantes. A Tabela 4 apresenta alguns diferentes cenários em que uma contestação possui até 5 vezes mais chance de ser acatada do que outra. O que ocorre é que por natureza uma contestação tem baixa probabilidade de ser acatada de modo geral, mesmo nos casos em que a probabilidade é mais alta, o que torna a predição do resultado final muito difícil de realizar. Além disso, foi possível identificar alguns fatores que aumentam a probabilidade da contestação ser acatada e em que magnitude. Por fim, a extração dos efeitos aleatórios por categoria, principalmente os de equipe (que foram mais significativos), servem para indicar em que equipes e estados o percentual está acima e abaixo da média, podendo nortear um futuro plano de ação.

4.2. Modelagem através de *machine learning* e *word2vec*

A análise descrita nessa seção focou exclusivamente nos textos das contestações, pois são uma variável mais complexa e se considerou que a abordagem pelo modelo de regressão, em um primeiro momento, não seria suficiente para extrair toda a informação contida nela.

O tratamento preliminar dos textos focou na limpeza de situações indesejáveis: retirada de nomes próprios de pessoas, retirada de acentuação, conversão de todos os caracteres do texto para letras minúsculas. Por fim foram construídos bigramas (duas palavras que para os efeitos de análise serão consideradas como uma) a partir da frequência com que apareciam juntas. Isso permitiu a construção de vetores de características para cada palavra, através da técnica *word2vec*. Após um período de experimentação com os parâmetros de ajuste da técnica, considerou-se que um vetor com 300 características, um tamanho de janela de 15 palavras e uma frequência mínima das palavras entre os textos igual a 2 produziram resultados satisfatórios e consistentes.

Uma maneira de conferir se os vetores construídos estão coerentes é verificar quais são os vetores mais próximos de uma palavra em particular, criando uma medida de similaridade. Isso indica palavras que aparecem em um mesmo contexto, segundo o ajuste da técnica. Como exemplo, a Tabela 5 demostra quais são as palavras mais similares em contexto às palavras "dossie","in"(abreviação de instrução normativa) e "casado". Analisando esse e outros cenários, concluiu-se que os vetores eram verossímeis e poderiam ser usados.

Para obter o vetor de características de uma contestação, foi feito o calculo da média dos vetores das palavras presentes no seu texto.

Depois desse procedimento, os vetores produzidos podem ser usados como entrada em qualquer método preditivo. Para esse fim foram escolhidos os modelos de *Random Forest*, *Support Vector Machine* (SVM) e *Multi Layer Perceptrons* (MLP). O MLP acabou sendo descartado após um curto período de testes pois estava produzindo resultados similares aos demais, mas era mais complexo de ser ajustado.

Os desempenhos do SVM e do *Random Forest* podem ser verificados na Tabela 6. Percebeu-se que am-

Palavra	Comparada com	Similaridade
dossie	del	0,86
	dossie_da_operacao	0,82
	dossie_eletronico	0,80
	deoc	0,77
in	in_item	0,86
	in_citada	0,81
	norma	0,74
	instrucao	0,73
casado	regime_de_separacao	0,91
	regime_de_comunhao	0,91
	esposo	0,90
	comunhao_parcial	0,89

Tabela 5: Medidas de similaridade produzidas pela técnica *word2vec* para as palavras, "dossie", "in"e "casado".

SVM

		Predito	
		Sim	Não
Real	Sim	1187	235
	Não	607	236

Ponto de corte: 0,2

Acurácia balanceada: 57%

Random Forest

		Predito	
		Sim	Não
Real	Sim	1027	187
	Não	767	284

Ponto de corte: 0,23

Acurácia balanceada: 58%

Tabela 6: Matrizes de confusão para o modelos SVM e *Random forest*. O ponto de corte ótimo foi escolhido através do índice de Youden.

bos os modelos apresentaram resultado muito similar ao modelo de regressão, com uma acurácia ajustada próxima aos 58%. A comparação entre os modelos através das curvas ROC na Figura 3 também apresenta essa similaridade.

O fato do desempenho dos modelos de *machine learning* se aproximarem do modelo de regressão foi algo inesperado, pois a natureza das duas construções é bastante distinta. Porém, foi visto como algo positivo: representa mais um fator que mostra que a extração manual de características do texto foi coerente e ao mesmo tempo serve como uma evidência de que com o conjunto de dados apresentado pode ser difícil obter um ajuste mais satisfatório.

Ao contrário da regressão, a abordagem utilizada nessa seção não permite uma análise mais profunda dos fatores que levam ao resultado obtido, pois os vetores de característica construídos não carregam nenhum significado interpretável sobre os textos. E também, embora exista a possibilidade de extrair a influência das variáveis em um modelo de *machine learning*, eles não são o método mais indicado para esse fim.

4.3. Combinação de modelos

Embora produzam desempenhos preditivos muito parecidos no que diz respeito a categorização, verificou-se que os modelos apresentados nas Seções 4.1 e 4.2 nem sempre atribuem probabilidades similares para o mesmo caso, como apresenta a Figura 2.

A partir dessa constatação surgiu a ideia de se criar uma terceira modelagem, tomando como variáveis explicativas as predições anteriores. Esse procedimento é conhecido como *stacking* [19], uma das várias técnicas de *ensemble* (combinação de modelos). Além de tentar melhorar o poder preditivo da análise de dados como um todo, dependendo de como é feita a combinação, é possível também mensurar quais dos modelos estão contribuindo de maneira mais significativa para o resultado final.

	Efeito	IC(95%)	P-valor
Intercepto	-3,2	(-3,7; -2,6)	0
pred_rf	3,9	(2,5; 5,3)	0
pred_svm	1,0	(-1,9; 4,0)	0,5
pred_glmm	3,4	(2,1; 4,6)	0

Tabela 7: Efeitos das predições dos modelos anteriores no ajuste do *ensemble* (GLM) com seu respectivo intervalo de confiança de 95% e significância estatística (P-valor).

Do ponto de vista conceitual, a ideia de combinar as modelagens anteriores também é coerente, pois cada uma abordou características distintas do problema apresentado.

Desse modo, as predições da regressão, do *Random Forest* e do SVM foram utilizadas em um modelo linear generalizado muito similar com o apresentado na Seção 4.1, porém contando apenas com efeitos fixos. As estimativas dos efeitos e sua significância estão apresentados na Tabela 7.

A primeira constatação é de que as predições de probabilidade do SVM não apresentaram significância estatística nesse *ensemble*. Comparativamente com os outros modelos, o *SVM* apresentou predições de probabilidade muito pouco variantes, quase sempre próxi-

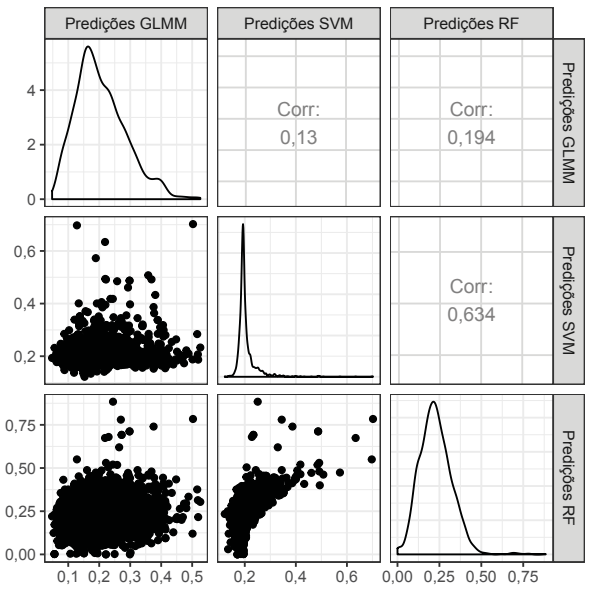

Figura 2: Matriz de gráficos de dispersão onde são confrontadas as probabilidades calculadas pelo modelos apresentados nas Seções 4.1 e 4.2.

mas a 20%, que é justamente a proporção de contestações acatadas sobre o total. Isso claramente contribuiu para a falta de significância desse modelo.

Conforme descreve a Tabela 8, o *ensemble* demonstrou uma acurácia balanceada de 68%, uma claro acréscimo aos modelos anteriores. Isso também fica evidenciado pela sua curva ROC (Figura 3.D) que apresenta uma AUC de 0,71. Esse resultado foi considerado satisfatório, uma vez que na prática está sendo construído indiretamente a partir dos componentes de apenas três variáveis: TEXTO_DA_CONTESTACAO, UF e EQUIPE.

Ensemble

		Predito	
		Sim	Não
Real	Sim	267	36
	Não	91	58

Ponto de corte: 0,22

Acurácia balanceada: 68%

Tabela 8: Matriz de confusão para o modelo *ensemble* ajustado (GLM). O ponto de corte ótimo foi escolhido através do índice de Youden.

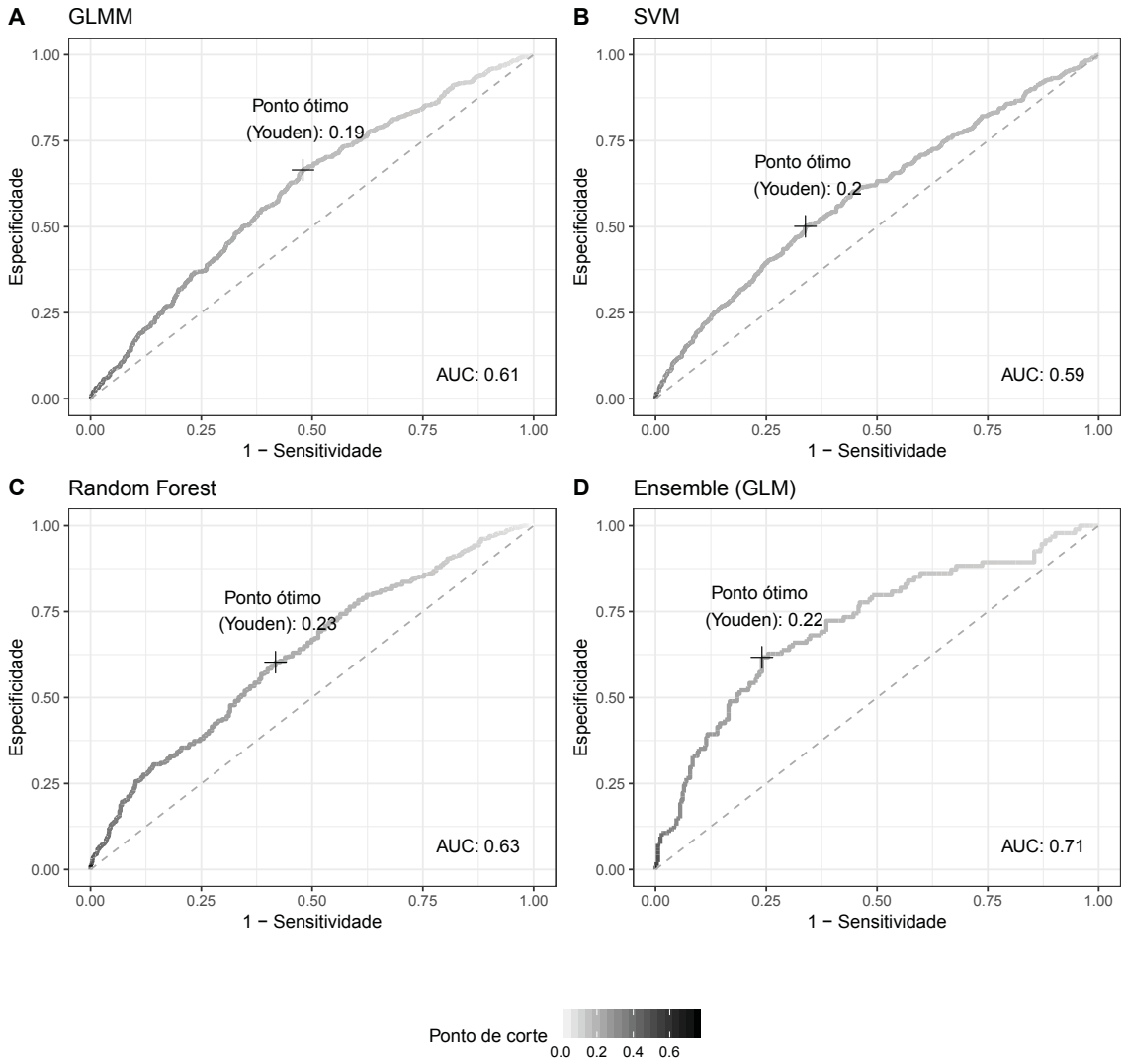

Figura 3: Curvas ROC para os quatro modelos ajustados para os dados. O modelo D (*ensemble*) é a combinação dos três anteriores.

5. Conclusão

Tendo como base as contestações recebidas pelo departamento especializado em analisar processos rurais da Instituição Financeira estudada, este trabalho teve como objetivo primário identificar os fatores associados ao seu acatamento. Em termos de predição das categorias, os resultados apresentados na Seção 4 evidenciaram a essência complexa e orgânica do problema, que não conseguiu ser ajustado satisfatoriamente nem com a análise textual tão quanto com as variáveis "tradicionais". De qualquer modo, o ajuste do modelo com efeitos mistos produziu alguns *insights* de características do texto que contribuem para a aceitação da contestação, além de fornecer interpretações interessantes em função das probabilidades estimadas. Esse modelo permitiu identificar o quanto certas características do texto, a equipe que analisa a operação e a UF de origem da operação influenciam na probabilidade de acatamento de uma contestação. Também permitiu verificar que em alguns cenários os processos possuem chance muito maior de serem acatados do que outros, ainda que a probabilidade de acatamento seja quase predominantemente inferior a 50%, o que dificulta a classificação.

Esses resultados podem facilmente produzir ações práticas: as equipes com maior incidência de contestações acatadas podem ser melhor orientadas; contestações com probabilidade de acatamento maior que um determinado valor podem receber prioridade; as agências que abrem as contestações podem ser orientadas a escreverem textos mais curtos para facilitar a sua interpretação. Cabe ao comitê estratégico do departamento especializado na análise das operações de crédito decidir quais ações são relevantes diante do apresentado.

Os modelos de *machine learning* se revelaram menos úteis, pois produziram resultados similares à regressão porém sem a interpretabilidade dos parâmetros. Ainda assim, serviram para mostrar que a regressão estava coerente e de certa forma validaram seu ajuste.

O *ensemble* se mostrou interessante pois melhorou a acurácia balanceada da predição para um nível mais aceitável. Ele pode ser usado na prática para fazer a predição da probabilidade de acatamento ao invés de se usar apenas o GLMM. Porém isso não significa que as interpretações fornecidas pelo GLMM deixam de ser válidas.

Por fim, a análise apresentada foi considerada satisfatória pois cumpriu a proposta de entender o fenômeno estudado e ao mesmo tempo pode servir para fornecer subsídios em futuras tomadas de decisão sobre o assunto. O problema em questão se mostrou interessante, pois apresenta uma situação real bastante complexa.

Referências

[1] Agropecuária puxa o PIB de 2017. `http://www.agricultura.gov.br/noticias/agropecuaria-puxa-o-pib-de-2017`, Jun 2019. [Online; accessed 6. Jun. 2019].

[2] Agropecuária Brasileira em Números. `http://www.agricultura.gov.br/assuntos/politica-agricola/agropecuaria-brasileira-em-numeros`, Jun 2019. [Online; accessed 6. Jun. 2019].

[3] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2019.

[4] Welcome to Python.org. `https://www.python.org`, Jun 2019. [Online; accessed 2. Jun. 2019].

[5] Hadley Wickham. *tidyverse: Easily Install and Load the 'Tidyverse'*, 2017. R package version 1.2.1.

[6] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016.

[7] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015.

[8] Max Kuhn. Contributions from Jed Wing, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, Brenton Kenkel, the R Core Team, Michael Benesty, Reynald Lescarbeau, Andrew Ziem, Luca Scrucca, Yuan Tang, Can Candan, and Tyler Hunt. *caret: Classification and Regression Training*, 2019. R package version 6.0-82.

[9] spaCy · Industrial-strength Natural Language Processing in Python. `https://spacy.io`, Jun 2019. [Online; accessed 2. Jun. 2019].

[10] Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA. `http://is.muni.cz/publication/884893/en`.

[11] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[12] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[13] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, Oct 2001.

[14] Ron Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. pages 1137–1143. Morgan Kaufmann, 1995.

[15] Norman E Breslow and David G Clayton. Approximate inference in generalized linear mixed models. *Journal*

of the American statistical Association, 88(421):9–25, 1993.

[16] Tom Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874, 2006.

[17] William J Youden. Index for rating diagnostic tests. *Cancer*, 3(1):32–35, 1950.

[18] Julian J Faraway. *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models*, pages 195–196. Chapman and Hall/CRC, 2016.

[19] David H Wolpert. Stacked generalization. *Neural networks*, 5(2):241–259, 1992.