UNIVERSIDADE ESTADUAL DE CAMPINAS

INSTITUTO DE MATEMÁTICA, ESTATÍSTICA E COMPUTAÇÃO CIENTÍFICA

KAINAN CREMM RAMOS

A MODELAGEM MATEMÁTICA PARA PREVISÃO DE RISCO DE TURNOVER EM EMPRESAS

Campinas
2020

UNIVERSIDADE ESTADUAL DE CAMPINAS
INSTITUTO DE MATEMÁTICA, ESTATÍSTICA E COMPUTAÇÃO CIENTÍFICA

KAINAN CREMM RAMOS

A MODELAGEM MATEMÁTICA PARA PREVISÃO DE RISCO DE TURNOVER EM EMPRESAS

Monografia apresentada ao Instituto de Matemática, Estatística e Computação Científica como parte dos requisitos para a obtenção de créditos na disciplina de Projeto Supervisionado, sob a orientação do Prof. Dr. Laércio Luís Vendite

Campinas
2020

RESUMO

Com a transformação digital e a crescente exponencial do Machine Learning, cada vez mais companhias têm incorporado ciência de dados para tomar decisões de negócios mais assertivas. E essa tendência vem surgindo em diferentes esferas da empresa, uma delas no setor de Recursos Humanos, seja para gerar insights, aumentar a assertividade no recrutamento de colaboradores. E um dos maiores desafios está ligado à saída de funcionários da empresa, o turnover (rotatividade). O presente projeto consiste de predição do risco de turnover, através do modelo de regressão logística múltipla utilizando o Python, apresentando uma alternativa para compreender o fenômeno e trazer soluções de melhoria na retenção de capital humano. O modelo construído mostrou-se eficiente, obtendo uma acurácia de 71,7%, com AUC de 74,9%.

Palavras-chave: Turnover. People Analytics. Regressão Logística. Machine Leaning. Data Science.

1 Introdução

1.1 Turnover

Quando falamos da saída de funcionários em uma empresa, mencionamos o termo “turnover”, que do inglês significa “rotatividade”. Em prática, o turnover é a taxa de rotatividade de colaboradores dentro de uma empresa, um dos principais indicadores da área de Recursos Humanos.

Em termos matemáticos, a taxa de turnover é calculada pela razão entre a quantidade de saída de colaboradores da empresa, pelo número médio de funcionários no dado período.

$$Taxa\ de\ turnover\ (\%) = \frac{\#\ de\ saídas}{\#\ média\ de\ colaboradores}$$

Esse indicador usualmente é calculado mês a mês, de modo a ter um acompanhamento anual acumulando a taxa de cada mês.

$$Taxa\ de\ turnover\ anual = Taxa\ de\ turnover_{Janeiro} + \cdots + Taxa\ de\ turnover_{Dezembro}$$

Sendo que ele é classificado das seguintes formas:

- Turnover involuntário: a iniciativa da saída é feita por parte da empresa, isto é, quando um colaborador é desligado (demitido). E geralmente isso acontece por redução de quadro, por comportamento inadequado, baixa performance, justa causa.
- Turnover voluntário: quando a iniciativa vem do colaborador, ou seja, ele pede a sua demissão, seja por insatisfação com o trabalho, por conta de salário, falta de reconhecimento e oportunidades de crescimento na empresa, dentre diversas outras razões.
- Turnover total: considera-se o número total de saídas sem distinguir entre o voluntário ou involuntário, usualmente calculado pela soma entre as duas parcelas.

Segundo estudos feitos pela Radford Global Technology Survey [3], empresa especializada em estudo voltados para a área de recursos humanos, o turnover total médio no Brasil no ano de 2016 foi de 12,8%, sendo 6,9% voluntário e 5,9% involuntário.

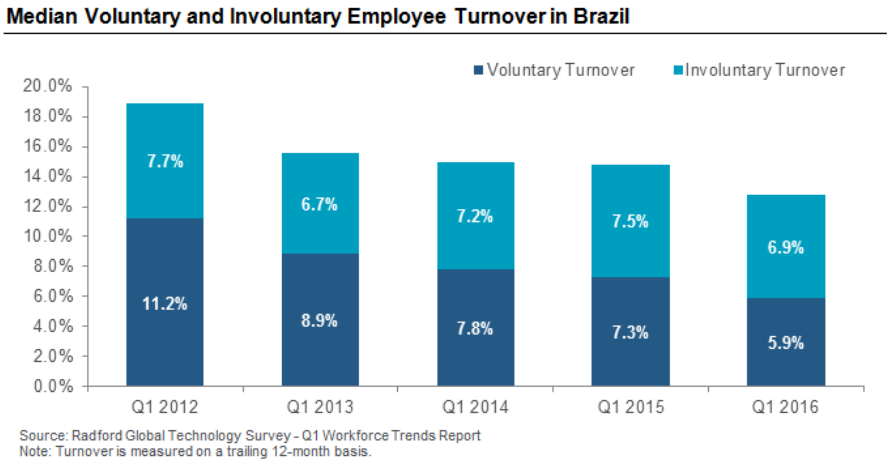

Figura 1: O turnover no setor brasileiro nos anos de 2013 a 2016

E olhando da perspectiva do negócio, quando temos altos índices de turnover, começamos a ter impactos negativos em diversas esferas; seja financeiramente, com todos os custos de recrutamento, seleção e treinamento com a reposição; comprometendo a entrega de resultados, com a diminuição da produtividade do time; a perda de conhecimento e a gestão do mesmo ao longo prazo.

Especialmente quando falamos do turnover voluntário: um estudo realizado pelo Centro Americano de Progresso [10], determinou que o custo médio para uma companhia que perde um empregado com expertise técnica, é de cerca de 213% o custo do salário anual do mesmo.

Além de que, um índice de turnover voluntário alto, começa a trazer indícios de possíveis outros problemas na companhia, tais como baixa remuneração, ambiente tóxico, falta de treinamento de lideranças, etc. E o desafio dentro do RH é o de entender as motivações de saída por trás da alta rotatividade, visto que existem muitas variáveis relacionadas e com alto nível de detalhe, o que gera análises muito complexas.

Por conta disso, cada vez mais empresas estão apostando na utilização de dados e tecnologias de *Big Data*, para tornar o entendimento desse fenômeno e a tomada de decisão mais assertiva. Vemos que companhias como a Nielsen, gigante da área de tecnologia de dados já utilizam modelos matemáticos e estatísticos para prever o risco de saída de um colaborador, tornando possível o entendimento de mudanças internas, que então quando foram implementadas, trouxeram melhorias na rotatividade.

1.2 Modelo de regressão logística múltipla

Diante da infinidade de modelos matemáticos e estatísticos para predição e medição de riscos, a regressão logística múltipla é uma das técnicas utilizadas para tratar de variáveis dependentes que são categóricas, ou seja, dicotômicas, binárias. Permitindo a estimativa da probabilidade de ocorrência de um evento, dado um conjunto de variáveis explanatórias (independentes).

Em termos formais:

Seja Y, variável dependente, tal que:

$$Y_i = \begin{cases} 1, \text{se o evento E ocorre} \\ 0, \text{caso contrário} \end{cases}$$

E o vetor $X_i = \{x_{i1}, x_{i2}, \ldots, x_{in}\}$, das n variáveis independentes (discretas ou contínuas), o modelo de regressão logística é descrito como:

$$P\{Y_i = 1\} = f\,(\beta_0 + \sum_j \beta_j x_{ij} + \varepsilon)$$

Onde:

- β_i são os coeficientes do modelo, calculados via máxima verossimilhança
- x_{ij} representam o estado das variáveis independentes, associadas a i-ésima variável dependente Y
- ε é o erro do modelo
- f é função logística, dada por $f\,(X_i) = \frac{1}{1-e^{-g(X_i)}}$

Podemos ver o comportamento da função na figura 2, cujo comportamento probabilístico tem o formato da letra S, tendo aplicações em diversas áreas: na medicina, em modelos de crescimento de tumores; na biologia, em dinâmicas de população, etc.

Na forma linearizada, o modelo é descrito pela função logit, sob a qual aplicamos os métodos clássicos para determinação dos parâmetros β_i.

$$logit(P\{Y_i = 1\}) = \beta_0 + \sum_j \beta_j x_{ij} + \varepsilon$$

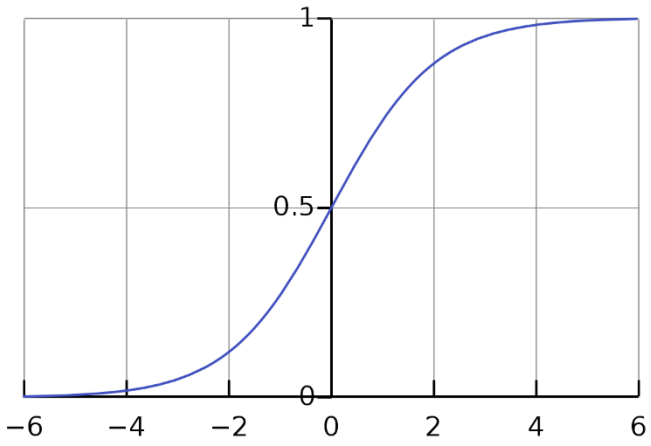

Figura 2: O comportamento da função logística

Ao construir o modelo, por Y se tratar de uma variável categórica, define-se um critério de classificação δ, a partir do qual consideramos a ocorrência do evento (Y=1) ou não; de forma que a mudança nesse parâmetro muda a eficácia do modelo.

$$Y = \begin{cases} 1, \text{ se } \mathrm{P}(\mathrm{Y}=1) \geq \delta \\ 0, \text{ caso contrário} \end{cases}$$

2. Desenvolvimento

2.1 Coleta de dados

Utilizamos a base de dados pública, disponibilizada pela IBM [5], sendo uma database típica do RH de uma empresa, contendo diversos campos, tais como idade dos funcionários, gênero, função, salário, tempo de casa, etc. Dentre um dos campos fundamentais para o nosso estudo, uma indicadora que diz se o funcionário saiu da empresa ou não, a partir do qual conseguimos definir como sendo a nossa variável Y do modelo de regressão logística, afinal, temos interesse em predizer o turnover voluntário.

$$E = \{evento\ que\ ocorre\ quando\ um\ empregado\ sai\ da\ empresa\}$$

$$Y = \begin{cases} 1, \text{ quando E ocorre, isto é, o funcionário sai da empresa} \\ 0, \text{ caso contrário} \end{cases}$$

De um total de 1470 funcionários, temos que 237 saíram da empresa e 1233 continuam. Se calcularmos o índice de turnover: 19,22% de rotatividade, que é considerável.

Ao total temos 30 campos, a maioria deles é numérico, entretanto algumas informações como gênero, departamento, estado civil, são do formato de texto.

Desta forma, para cada campo de índice i designamos o vetor dependente $X_i = \{x_{i1}, x_{i2}, ..., x_{in}\}$ para os n funcionários da empresa, com $i \in (1,30)$.

A tabela 1 mostra as informações que temos para os funcionários da empresa.

Variáveis do problema		
Idade	EnvolvimentoNoTrabalho	TotalDeAnosEmCarteira
Turnover	NivelHierarquico	TreinamentosRealizadosNoAnoAnterior
TurnoverIndicador	SatifacaoNoTrabalho	WorkLifeBalance
SalarioDiario	SalarioMensal	AnosNaEmpresa
DistanciaDeCasa	MonthlyRate	AnosNaPosicaoAtual
Educacao	NumEmpresasTrabalhadas	AnosDesdeUltimaPromocao
CountColaborador	PercentualAumentoSalario	TempoComChefeAtual
ID	AvaliacaoDePerformance	Genero
SatisfacaoAmbienteDeTrabalho	SatisfacaoDeRelacionamento	Departamento
SalarioPorHora	StockOptionLevel	EstadoCivil

Tabela 1: Os campos de informação disponível para os funcionários da empresa

2.2 Análise das variáveis

Com a base de dados em mãos, fomos entender melhor o problema da rotatividade na empresa, através da análise do perfil de colaboradores, traçando medidas para cada um dos campos (média, mediana, desvio padrão), juntamente com as distribuições de probabilidade estimadas já normalizadas (usamos a função *sns.kdeplot* do Python) separando dentre:

- Leavers - colaboradores que saíram da empresa, isto é, Y=1 (cor laranja)
- Ativos - colaboradores que permanecem na empresa, Y=0 (cor azul)

A tabela 2 mostra as distribuições para cada um dos campos numéricos que foram analisados.

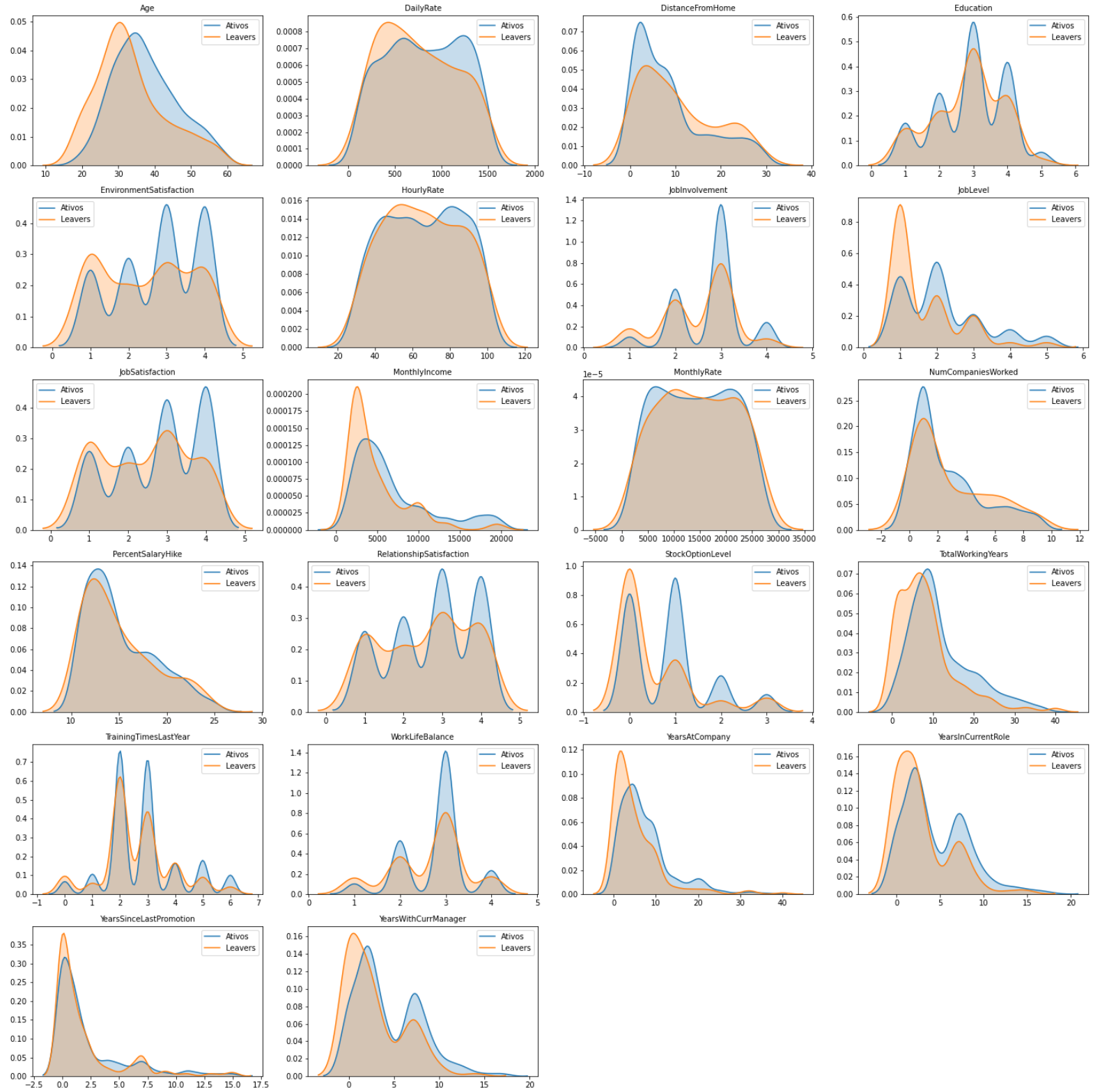

Tabela 2: As distribuições de probabilidade por variáveis do problema

Num panorama geral, podemos observar que:

- Podemos observar que quanto mais novos os funcionários são, temos uma maior chance de risco de turnover; podendo-se inferir que a

rotatividade é maior no início da vida profissional, como estudos feitos por Chahad e Macedo [11].

- Com respeito aos campos de satisfação (do ambiente, envolvimento), quanto menor as notas de satisfação, maior é a probabilidade de sair, o que é intuitivo.
- Algumas das variáveis que mais se destacaram, em termos de diferenças de distribuições foram:

  - SalarioMensal – funcionários com menores salários tem uma probabilidade maior de deixar a empresa.

  - StockLevelOption – que diz respeito ao nível de aquisições de ações da companhia que o funciona tem, de forma que quanto maior é o nível de aquisições, menor é a probabilidade de saída, o que pode ser entendido pela relação de pertencimento do mesmo com a empresa;

2.3 Preparação de dados

Antes da aplicação do modelo, transformamos as variáveis categóricas explicativas (do vetor X) tais como gênero, estado civil e departamento em um número, através da função *LabelEncoder*, transformando-as na nossa base de dados.

Normalizamos todas as variáveis, para que não ocorram distorções de escala entre os diferentes campos (visto que temos variáveis com escalas de unidade e outras de milhares), visando uma melhor performance.

Então, com as transformações, separamos a nossa base entre treino e teste (tanto para X como para Y), para utilizarmos o conjunto de treino no ajuste dos coeficientes β_i do modelo; e validarmos a efetividade do modelo com o conjunto teste, realizando as predições. Para isso, utilizamos a função *train_test_split*, sendo que utilizamos diferentes valores para a proporção do tamanho do conjunto de treino, tipicamente com 80% e 60% do percentual da database.

2.4 Aplicação do modelo

Finalmente, exportamos o nosso modelo de regressão linear e realizamos o ajuste do mesmo com o conjunto de teste, a partir do qual, pelo método da

máxima verossimilhança (como descrito em 1.2), calculamos os parâmetros do modelo.

Importante ressaltar que uma das features utilizadas na função *LogisticRegression* onde calculamos o modelo, foi a de class_weight='balanced', cuja importância se dá, pois, a nossa base de dados é desbalanceada: temos 1233 funcionários com variável Y=0 (não saíram da empresa), e apenas 237 com Y=1 (saíram da empresa). Sendo assim, se não aplicássemos uma feature ou uma função de balanceamento nos nossos dados, teríamos um modelo com baixa performance, visto que não haveria uma boa proporção entre as variáveis categóricas da amostra e o modelo tenderia a dar muitos “alarmes falsos”.

Finalizado isso, aplicamos o nosso modelo ajustado para predizer o vetor Y, e comparamos a predição com o conjunto Y de testes.

3. Resultados

O modelo apresentado obteve as seguintes métricas.

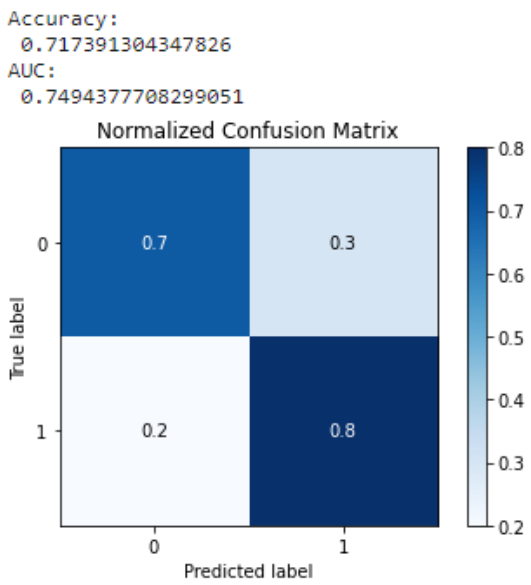

Figura 3: As métricas do modelo logístico apresentado (Jupyter)

- AUC: 74,9%

Contextualizando, o AUC é uma medida de separabilidade, um valor entre 0 e 1 que representa que mostra como seu modelo está performando ao utilizar como base para o cálculo a taxa de falso positivo, a taxa de verdadeiro positivo.

O valor 1 seria o caso ideal (perfeito, em teoria) e 0 o pior caso possível. Vemos que nosso AUC = 74,9%, sendo adequado, dado o threshold utilizado para o modelo, tipicamente de 50%.

- Acuracidade: 71,7%

A acuracidade nos mostra, em média, a frequência da assertividade do nosso modelo; sendo que o percentual atingido se mostra pertinente, contudo, iremos analisar nossa matriz de incidência para não tirar conclusões precipitadas.

- Observando a figura 3, vemos que o modelo apresentou 70% de taxa de verdadeiros positivos de 80%, 70% de verdadeiros negativos, 20% de falsos negativos e 30% falsos positivos.

O comportamento da densidade da matriz está pertinente, sendo mais evidente nas diagonais principais. Com respeito aos falsos negativos e positivos, entendemos que o percentual pode ser considerado relevante, entretanto podemos ter diferentes cenários conforme alteramos o threshold δ da probabilidade associada à variável Y;

Em linhas gerais, o modelo mostrou-se eficiente, mesmo sem ter o envolvimento do RH da empresa no processo de desenvolvimento do modelo (afinal, usamos uma base de dados pública, da IBM). Consideramos que a expertise e a vivência do departamento são fundamentais para os inputs do modelo, seja para eliminar ou incluir variáveis, validar hipóteses, etc; acreditamos que em uma aplicação real e com esse alinhamento, poderíamos ter uma performance ainda superior.

4. Conclusão

O projeto consistiu da construção de um modelo para predição do risco de turnover, através do modelo de regressão logística múltipla utilizando Python 3, apresentando uma alternativa para compreender o fenômeno e trazer soluções de melhoria na retenção de capital humano dentro de uma empresa. Em termos de resultado, o modelo mostrou-se eficiente, tendo uma acurácia de 71,7%, com AUC de 74,9%.

Reiteramos a relevância do modelo proposto, na identificação de casos estratégicos de funcionários classificados com risco de turnover, como forma de gerar ações preventivas que contribuam com a permanência do mesmo. Com tudo isso, acreditamos no trabalho conjunto com o departamento de RH, participando das etapas de modelagem, alinhando a estratégia da empresa, para então transformar os insights do modelo em ações que tragam impacto.

5. Referências

[1] Paula, Gilberto. **MODELOS DE REGRESSÃO com apoio computacional**. São Paulo, 2010. pp 199-231.

[2] **Scikit-learn: Machine Learning in Python**, Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011

[3] Pieczynski, Dino. **Employee Turnover Slows in Brazil**. Empower Results, 2016. Disponível em: https://rewards.aon.com/en-us/insights/articles/2016/employee-turnover-slows-in-brazil

[5] **Regressão Logística**. E-Disciplinas USP. Disponível em: https://edisciplinas.usp.br/pluginfile.php/3769787/mod_resource/content/1/09_RegressaoLogistica.pdf

[5] Vulpen, Van Erik. **HR Data Sets for People Analytics**. AIHR, 2020. Disponível em: https://www.analyticsinhr.com/blog/hr-data-sets-people-analytics/

[6] Pissinelli, Glaucia; Duarte, Leonardo e Torezzan, Cristiano. **Modelo de regressão logística múltipla para predição de Turnover**. Simpósio Brasileiro de Pesquisa Operacional, 2017. Disponível em: http://www.sbpo2017.iltc.br/pdf/169536.pdf

[7] Ryder. **IBM HR Analytics Employee Attrition & Performance**. Kaggle, 2018. Disponível em: https://www.kaggle.com/mcminnra/ibm-hr-logistic-regression-89-7-with-rfe

[8] Duarte, Rafael. **Reduzindo Turnover com Ciência de Dados**. Medium, 2020. Disponível em: https://medium.com/@rafaelnduarte/reduzindo-turnover-com-ci%C3%AAncia-de-dados-ce8e98d0a40

[9] Vaz, Arthur. **Dados desbalanceados em problemas de classificação**. Medium, 2019. Disponível em: https://medium.com/data-hackers/como-lidar-com-dados-desbalanceados-em-problemas-de-classifica%C3%A7%C3%A3o-17c4d4357ef9

[10] Boushey, Heather e Glynn, Sarah Jane. **There Are Significant Business Costs to Replacing Employees**. Center for American Progress, 2012. Disponível em: https://www.americanprogress.org/wp-content/uploads/2012/11/CostofTurnover.pdf

[11] MACEDO, R; CHAHAD, J. P. **O FGTS e a Rotatividade da Mão-de-Obra**, FIFE, São Paulo.