Modelo preditivo para classificar engajamento e satisfação no trabalho

PÓS-GRADUAÇÃO EM CIÊNCIA DE DADOS E INTELIGÊNCIA ARTIFICIAL

ALUNO: Alessandre da Silva Martins
ORIENTADORA: Cristina Moreira Nunes

PUCRS online uol edtech

RESUMO

O presente trabalho tem como objetivo desenvolver uma modelagem preditiva de classificação de funcionários engajados e satisfeitos. O modelo usa aprendizado de máquina supervisionado. Dado um novo exemplo de funcionário o qual não conhecemos sua classificação, o modelo é capaz de predizer a classe a que o funcionário pertence, bem como a probabilidade dele pertencer a classe predita. Foi utilizado um conjunto de dados com registros de 311 funcionários. Depois da modelagem de dados, os funcionários foram separados em engajados/satisfeitos e não engajados/insatisfeitos. Foram experimentados aleatoriamente alguns algoritmos e o Random Forest apresentou o melhor resultado para o dataset. Realizada a análise descritiva dos dados criou-se o modelo com 70% do conjunto para treinamento e 30% para teste. A precisão e a acurácia das duas classes alcançaram 100%. Os resultados encontrados indicam que quanto mais satisfação no trabalho mais engajados são os funcionários, sendo os dois atributos fortemente relacionados ao indicador estabelecido como *target*.

Palavras-chaves: modelo preditivo de classificação, aprendizado de máquina supervisionado, probabilidade, engajamento, satisfação.

ABSTRACT

This article has an objective to develop a predictive modeling of engaged and satisfied employees. The model uses supervised machine learning. Given a new employee which is not an example of its classification, the model is able to predict the class to which the employee belongs, such as the probability of him/her belonging to the predicted class. A dataset with records of 311 employees was used. Modeling data, employees were separated into engaged/satisfied and non-engaged/not satisfied. Some algorithms were randomly tested and Random Forest presented the best result for the data set. After performing the descriptive analysis of the data, the model was created with 70% of the set for training and 30% for testing. The precision and accuracy of the two classes reached 100%. The results found indicate the more satisfaction at work, the more engaged the employees are, both attributes being strongly related to the indicator established as target.

1. INTRODUÇÃO

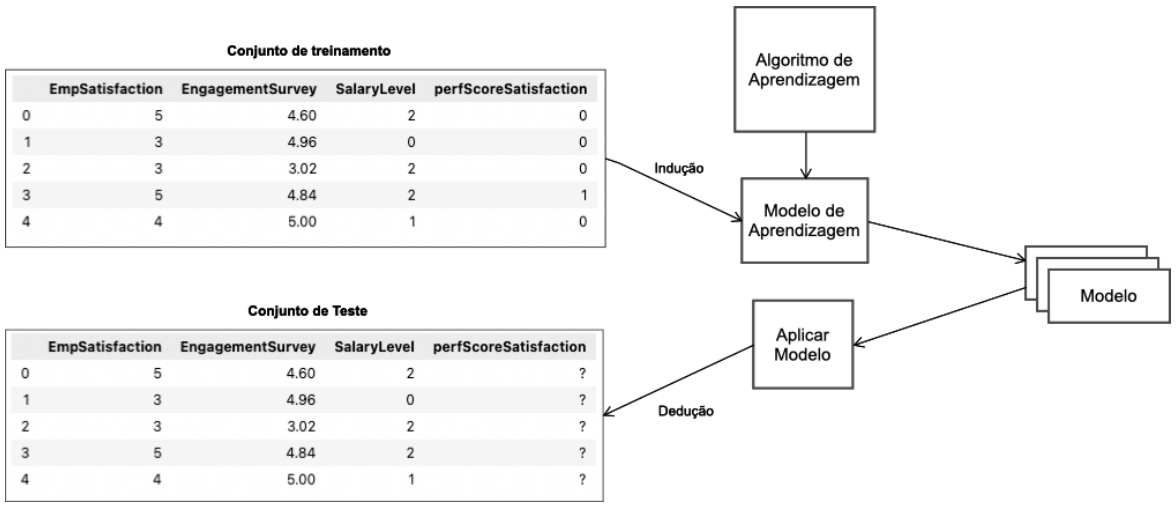
Segundo ALEXANDRE(1983), a “Nova economia” representa a grande mudança que houve no mercado nos anos 1980. As inovações tecnológicas constantes, a internet e mudanças nos modelos de negócio contribuem para acelerar as novas demandas. O mercado exige cada vez mais a capacidade de tomar decisões importantes tão rápido quanto as mudanças. Para agir com confiança precisamos estar atualizados e preparados. A “nova economia” apresenta uma ruptura na forma de organização nas empresas. De organização verticalizada e hierárquica para uma organização horizontal; flexível, incerta, colaborativa, arriscada, imprevisível e diversa.

De acordo com Scanlan(2016), em uma pesquisa de satisfação e engajamento com 600 funcionários, quase a metade afirma que o reconhecimento de seu desempenho é importante para sua satisfação no trabalho. Esta pesquisa mostrou funcionários moderadamente engajados com um índice de 3.8, comparados aos anos anteriores (3.7 em 2014 e 3.6 em 2013). O engajamento dos funcionários pode ou não estar relacionado com a satisfação dos funcionários no trabalho, pois o engajamento está atrelado à conexão e compromisso dos funcionários com seu trabalho e sua organização. Embora os níveis de engajamento indiquem que os funcionários parecem estar moderadamente engajados, esse nível de engajamento pode não ser generalizado em toda a organização, pois em níveis mais baixos de trabalho parecem estar menos engajados.

As empresas possuem profissionais capazes de gerir recursos e lidar com as incertezas e inconstâncias do mercado. A proposta neste trabalho é desenvolver um classificador de profissionais de forma que construam suas carreiras na empresa, e fornecer uma análise dos principais indicadores de engajamento para os profissionais alinhados com a empresa. Propomos também uma forma de calcular a

probabilidade que indica possibilidade do funcionário ser classificado como engajado e satisfeito. Vamos construir uma análise de dados para embasar novos cenários na empresa, e reconhecer a maturidade de um profissional a fim de capacitá-lo para os desafios da empresa. A análise dos perfis profissionais contribui para posicioná-los em favor do negócio da organização.

Figura 1: Diagrama de ideação da Solução



Fonte: https://www.researchgate.net/figure/Figura-15-Abordagem-geral-para-construir-um-modelo-de-classificacao_fig8_276060077.

A Figura 1 aborda como funciona a solução. Com base no conjunto de dados de treinamento, o modelo é capaz de identificar as classes de performance e satisfação dos usuários e ter uma indução generalizada dos exemplos fornecidos. Após a obtenção dessa indução, se aplica o modelo a um conjunto de dados de testes onde não sabemos qual seria a classe de performance e satisfação, o modelo prediz a que classe os funcionários pertencem.

2. METODOLOGIA

Para alcançar o objetivo deste trabalho - de desenvolver uma modelagem preditiva para classificação de funcionários engajados e satisfeitos - se desenvolveu o projeto sobre o conjunto de dados criado por Huebner e Patalano (2021).

Um dos desafios é estabelecer uma métrica no conjunto de dados que classifique os funcionários engajados e satisfeitos, a priori vamos definir como perfScoreSatisfaction. Outro desafio é experimentar algoritmos classificadores dessa métrica quando abaixo ou acima do ideal usando os dados de teste. Selecionamos algumas variáveis dependentes para o perfScoreSatisfaction, incluindo avaliações de desempenho, engajamento, quantidade de atrasos e faltas no trabalho.

No escopo de planejamento estão previstas as fases: 1 - Propor arquitetura de solução; 2 - Modelagem dos dados; 3 - Análise exploração dos dados; 4 - Desenvolvimento do modelo; 5 - Análise dos resultados; 6 - Realização de testes do modelo.

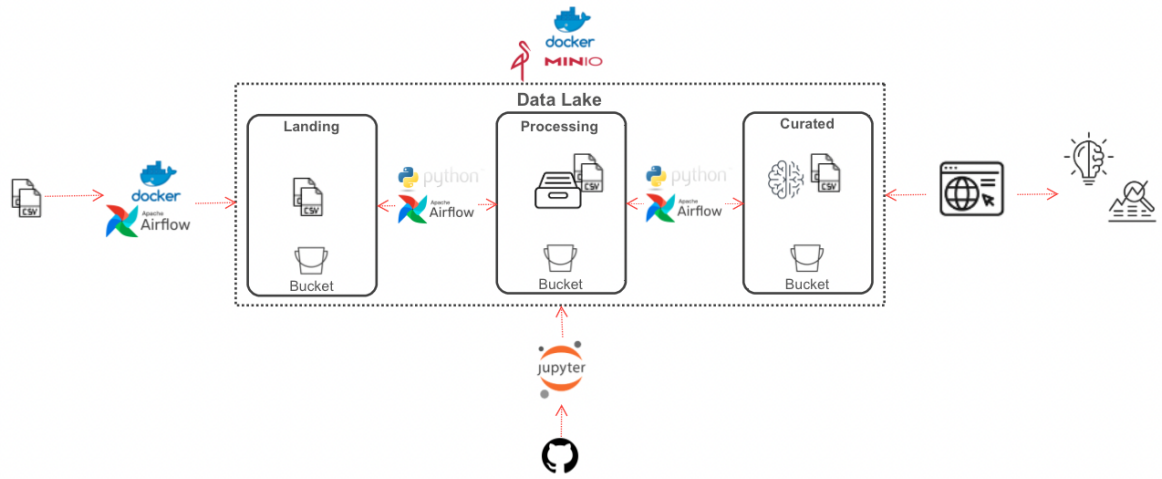
As tecnologias utilizadas foram: MinIO, para trabalhar com DataLake; Apache Airflow, ferramenta para gerar pipelines; Docker, para uso de contêineres com propósito de facilitar o uso das demais tecnologias; Scikit Learn no desenvolvimento do modelo de Machine Learning; Python, como linguagem de programação; e biblioteca Pandas.

Arquitetura da solução

No cenário proposto na Figura 2 visualizamos a arquitetura da solução, onde temos um conjunto de dados semi-estruturados distribuídos em CSV.

Utilizamos o AirFlow em Docker para orquestrar a coleta de dados e selecionar os dados das diversas fontes. O AirFlow vai juntar as fontes e passar para o DataLake.

Figura 2: Arquitetura da Solução



Fonte: Autoria própria.

O DataLake em MinIO realiza a persistência dos dados passando os dados pelos diversos estágios que representam momentos diferentes dos dados: 1 - ‘Landing’ é o estágio dos dados brutos, podem vir das mais variadas fontes e formatos como csv, xls, sql e outros; 2 - ‘Processing’ são os dados processados. Nesse momento eles passaram por padronização, particionamento e consolidação; 3 - ‘Curated’ os dados foram apurados, passaram por um refinamento e são

integrados ao modelo, prontos para consumo. Todo esse pipeline é suportado em docker.

Para ser mais didático, estamos acessando os dados pelo DataLake em um computador local numa estrutura física de arquivos usando MinIO. Criadas landings com os buckets apontando para repositórios locais, dessa forma conseguimos passar o caminho físico para acessar os dados. Em um eventual ambiente de produção podesse criar um servidor específico para rodar MinIO, S3 da AWS ou outros serviços do tipo. Para acessar esses ambientes em cloud usamos protocolos de acesso específicos do fornecedor desse serviço. Na AWS, por exemplo, temos o S3, muito semelhante ao que usamos no MinIO. Dessa forma nos autenticamos no nosso storage para usar os dados contidos nele. Isso se aplica quando formos automatizar o código implementado no Jupyter Notebook, onde criamos um conjunto de tarefas organizadas a ser executadas dentro do servidor Airflow.

Modelagem dos dados

O conjunto de dados HRDataset_v14.csv contém 36 atributos de 311 funcionários. Segundo Huebner e Patalano (2021), são dados reais, os dados sensíveis foram alterados para uso didático para estudos de caso. Dentre informações importantes para o projeto podemos encontrar informações sobre remuneração dos funcionários, data de início, pontuação de desempenho, departamento, satisfação, nível de engajamento, status ativo ou de rescisão.

Na Tabela 1 apresenta-se os tipos de informações encontradas no conjunto de dados.

Tabela 1: Dicionário de dados

Atributo	Descrição	Tipo
EmpID	Identificador exclusivo para cada funcionário	Text
MarriedID	Casado (1 = sim ou 0 = não)	Binary
MaritalStatusID	Código de estado civil correspondente ao campo de texto MaritalDesc	Integer
EmpStatusID	Código de status de funcionário correspondente ao campo de texto EmploymentStatus	Integer
DeptID	Código de identificação do departamento correspondente ao departamento em que o funcionário trabalha	Integer
PerfScoreID	Código de pontuação de desempenho correspondente à pontuação de desempenho mais recente do funcionário	Integer
FromDiversityJobFairID	O funcionário foi originado da feira de empregos Diversity? (1 = sim ou 0 = não)	Binary
Salary	O salário anual do funcionário em dólares americanos.	Float
Termd	Demitido? (1 = sim ou 0 = não)	Binary
PositionID	Código que indica a posição do funcionário	Integer
Position	Descrição da posição que o funcionário ocupa	Text
Zip	Zip code do funcionário	Text

DOB	Data de nascimento do funcionário	Date
Sex	Sexo - M ou F	Text
MaritalDesc	O estado civil (divorciada, solteira, viúva, separada, etc.)	Text
CitizenDesc	Descrição para saber se o funcionário é 'Cidadão' ou 'Não Cidadão Elegível'	Text
HispanicLatino	Campo Sim ou Não para funcionário ser hispânico/latino	Text
DateofHire	Data em que a pessoa foi contratada	Date
DateofTermination	Data em que o funcionário foi rescindido, preenchida se Termd = 1	Date
TermReason	Descrição do motivo pelo qual o funcionário foi demitido	Text
EmploymentStatus	Uma descrição do status do funcionário. Qualquer pessoa que esteja trabalhando em tempo integral = Ativo	Text
ManagerName	O nome do gerente imediato	Text
ManagerID	Identificador exclusivo para cada gerente.	Integer
RecruitmentSource	O nome da fonte de recrutamento onde o funcionário foi recrutado	Text
PerformanceScore	Descrição da Pontuação de Desempenho (Excede as expectativas, Atende às expectativas, Abaixo das expectativas(PIP),	Text

	Excepcional)	
EngagementSurvey	Resultados da última pesquisa de engajamento, gerenciada pelo parceiro externo	Float
EmpSatisfaction	Um índice de satisfação entre 1 e 5, conforme relatado em uma recente pesquisa de satisfação dos funcionários	Integer
SpecialProjectsCount	O número de projetos especiais em que o funcionário trabalhou nos últimos 6 meses	Integer
LastPerformanceReviewDate	A data mais recente da última revisão de desempenho da pessoa.	Date
DaysLateLast30	O número de vezes que o funcionário se atrasou para trabalhar nos últimos 30 dias	Integer
Absences	O número de vezes que o funcionário se ausentou do trabalho.	Integer
TimeInCompany	Tempo de empresa em anos	Integer
SalaryLevel	Nível de salário	Float
PerfScoreSatisfaction	funcionários satisfeitos e performáticos	Binary

Fonte: https://rpubs.com/rhuebner/hrd_cb_v14

Criou-se uma função para calcular o tempo de empresa em anos, o tempo é dado pela diferença da data de contratação do funcionário até a data do término do contrato. Na ausência do término do contrato considera-se a data do último ano do último contratado.

Implementou-se outra função para categorizar o salário por níveis,

estabelecendo os níveis de salário como: Abaixo do percentil inferior de 25% são ‘low’; entre o percentil inferior e superior de 75% são ‘medium’; e os acima do percentil superior são ‘high’.

Criou-se mais uma função para definir a satisfação e performance dos funcionários. O retorno dessa função será *True* quando o atributo 'PerformanceScore' for igual a 'Fully Meets', 'Needs Improvement' e 'EmpSatisfaction' iguais a 3, 4 ou 5 e o funcionário esteja ativo na empresa. Seu resultado é atribuído ao atributo ‘PerfScoreSatisfaction’.

Essas funções e todos os códigos implementados na modelagem dos dados foram introduzidos como dags no Airflow para a fase de automação do processo. A dag passa a ser executada dentro do servidor Airflow.

Análise exploração dos dados

Os Notebooks criados e seus resultados aqui citados, estão disponíveis na plataforma GitHub no link https://github.com/sandremartins/Artigo_PUCRS_CD_IA.

Obtemos alguns resultados dos cálculos de taxas do periodo de Janeiro de 2006 a Julho de 2018: A taxa de PerfScoreSatisfaction é de 54% dos funcionários estão satisfeitos com a empresa e foram avaliados com boa performance; 67% dos profissionais estão ativos na empresa ao longo dos anos 12 anos; 56% se identificaram como do sexo feminino enquanto que 44% se identificaram como do sexo masculino; 44% solteiros e 40% casados; 78% foram considerados de altaperfomance e 12% exelentes; mais de 90% são satisfeitos com o ambiente de trabalho.

Cálculos das médias: Salário médio é 69.020 por ano; A satisfação média 3.8 numa escala de 0 à 4 com desvio de 0.9; O engajamento médio é 4.1 numa escala

de 0 á 5 com desvio de 0.7 e mínima de 1.1; A taxa de 54% com funcionários performáticos e satisfeitos.

Correlation Matrix & Heatmap

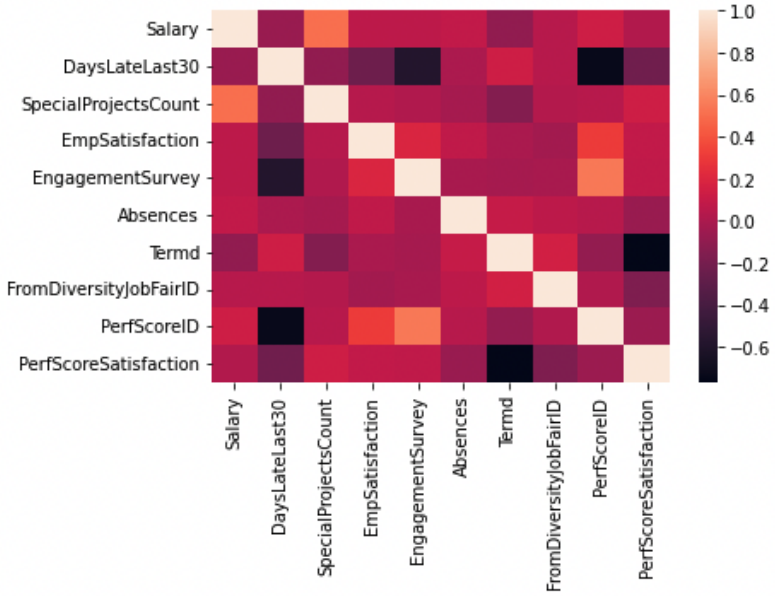
Na Figura 3 evidenciamos os valores mais altos(positivos) e mais baixos(negativos) de algumas correlações utilizando a Tabela 2 para classificação dos valores obtidos.

Tabela 2: Categorização para os valores do coeficiente de correlação de Pearson

Coeficiente de Correlação	Classificação
0 < r \(\leq\) |0,1|	Nula
|0,1| < r \(\leq\) |0,3|	Fraca
|0,3| < r \(\leq\) |0,6|	Moderada
|0,6| < r \(\leq\) |0,9|	Forte
|0,9| < r < |1|	Muito Forte
r = 1	Perfeita

Fonte: https://gpestatistica.netlify.app/blog/correlacao/

Figura 3: Correlation Matrix & Heatmap



Fonte: Autoria própria.

Atributos com uma correlação moderada positiva:

- SpecialProjectsCount com Salary: 0.50
- EngagementSurvey com PerfScoreID: 0.54
- PerfScoreID com EmpSatisfaction: 0.30

Atributos com uma correlação moderada negativa:

- DaysLateLast30 com PerfScoreID: -0.73
- EngagementSurvey com DaysLateLast30: -0.58
- DaysLateLast30 com PerfScoreSatisfaction: -0.22
- PerfScoreSatisfaction com Termd: -0.76

Existem correlações positivas. Entre os atributos SpecialProjectsCount e Salary, faz sentido que funcionários que estão envolvidos em mais projetos tenham maiores ganhos. A correlação de PerfScoreID com EmpSatisfaction é menor mas

representativa para sugerir que os mais satisfeitos são bem avaliados. Outra correlação positiva é entre EngagementSurvey e PerfScoreID, o que indica que os funcionários mais bem avaliados na empresa são mais engajados.

Há também correlações negativas. Entre os atributos DaysLateLast30 com PerfScoreID, EngagementSurvey com DaysLateLast30, PerfScoreSatisfaction com Termd com uma relativa DaysLateLast30 com PerfScoreSatisfaction. Os atrasos nos últimos 30 dias podem ser um indicativo de insatisfação e consequente término de contrato. Vale a pena uma intervenção nesse momento? Podemos assumir que os funcionários que mais atrasaram dos últimos 30 dias estão com o score baixo e consequentemente estão fracamente engajados.

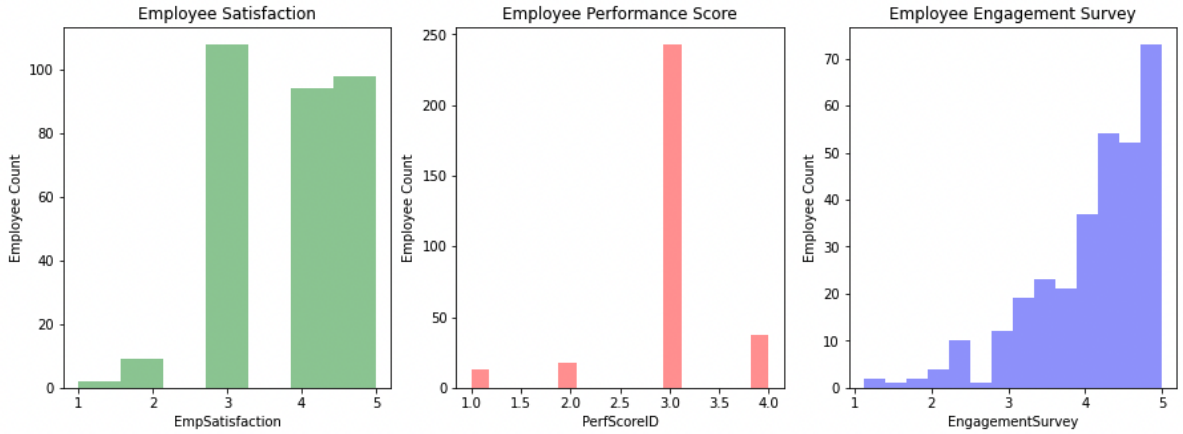
Distribuição de satisfação, performance e engajamento

A avaliação de satisfação, com base na Escala Likert vamos considerar sua forma típica, onde 1: Péssimo, 2: Ruim, 3: Regular, 4: Bom, 5: Ótimo. Na Figura 4 podemos visualizar um pico de funcionários satisfeitos com a empresa. Porém, a maioria avaliou estar muito satisfeito ou muito satisfeito.

A performance tem uma concentração da avaliação dos funcionários com 3.0 bem próximo da média, com poucos funcionários abaixo de 3.

No quesito engajamento a concentração de funcionários é acima de 4. Sendo uma distribuição assimétrica à direita. Existe uma razão para um cenário otimista para a empresa e os funcionários?

Figura 4: Distribuição de satisfação, performance e engajamento



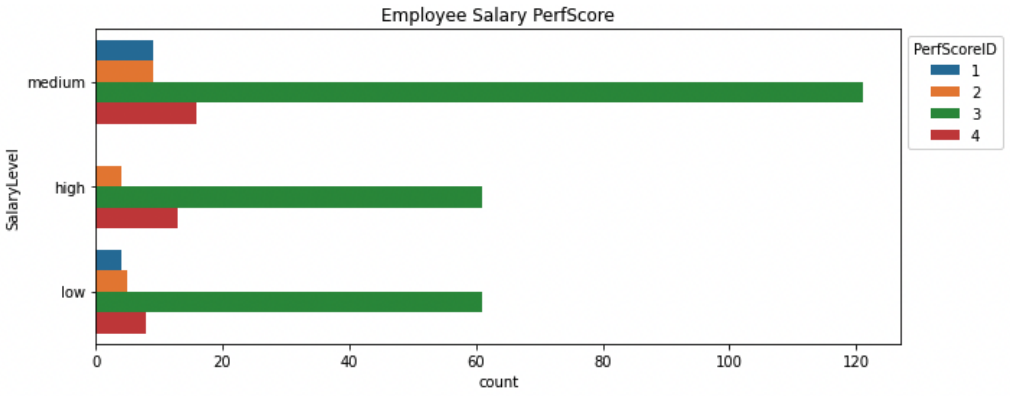
Fonte: Autoria própria.

Relação entre performance e salário

Na Figura 5 podemos observar que a maior parte dos funcionários que saíram da empresa tinha nível de salário ‘médio’. Poucos funcionários com alto salário deixaram a empresa. O que fez 7% dos funcionários com alto salário saírem da empresa?

A maioria, 121 funcionários foram avaliados com performance que atendem totalmente, como engajados e tem salário mediano. 61 funcionários com altos salários foram declarados que atendem totalmente e são engajados. Os mais engajados foram 13 funcionários com salário alto, 16 com salário médio e 8 com salário baixo.

Figura 5: Relação entre performance e salário



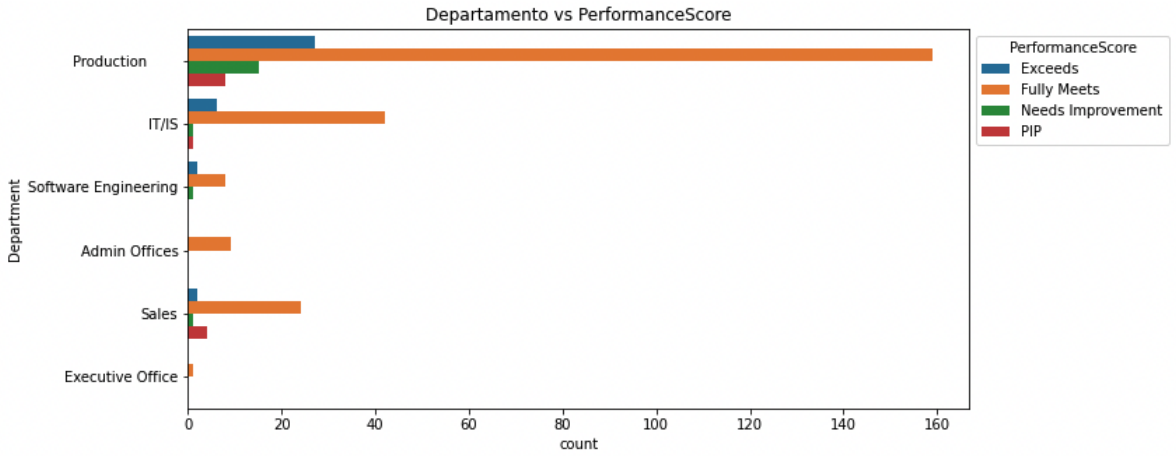
Fonte: Autoria própria.

Importante salientar, dos que tinham salário baixo, 4 funcionários foram avaliados com engajamento abaixo das expectativas. Os funcionários com nível de salário médio são predominantes na empresa, sendo 155 funcionários, representando aproximadamente 50% do quadro efetivo.

Relação de performance por departamento

Conforme a visualização da Figura 6, os departamentos de Produção, Técnico/Suporte e Vendas são os top 3 departamentos com maior quantidade de funcionários satisfeitos e performáticos, sendo o departamento de Produção o mais representativo, tendo uma taxa menor de satisfação e performance.

Figura 6: Relação de performance por departamento



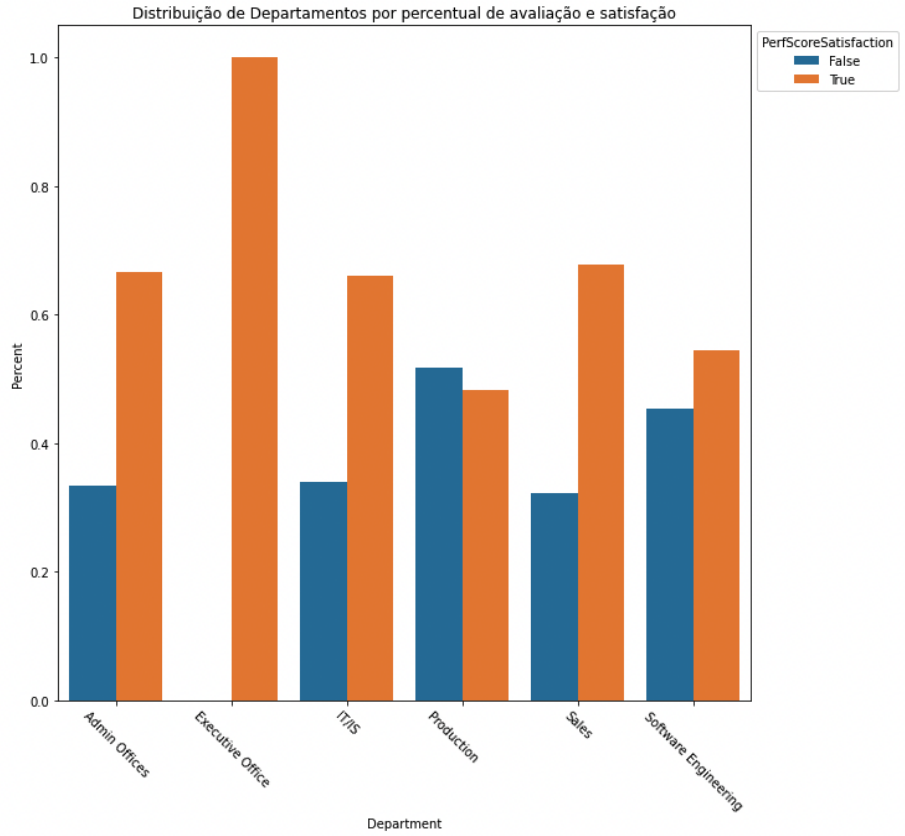
Fonte: Autoria própria.

Relação entre departamento e porcentagem de avaliação e satisfação

O departamento de Produção pode ser o mais representativo, porém tem o menor índice dos mais satisfeitos e performáticos.

Os departamentos de Vendas, IT/IS e Administrativo são muito semelhantes nas avaliações, como podemos observar na Figura 7.

Figura 7: Relação entre departamento e porcentagem de avaliação e satisfação

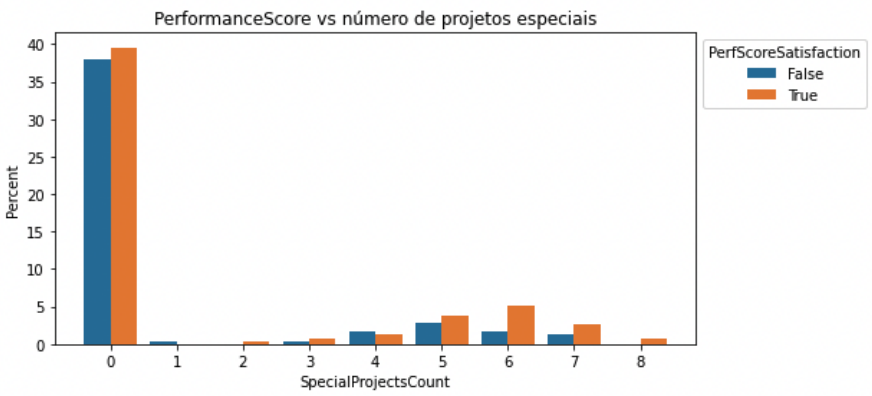


Fonte: Autoria própria.

Relação entre performance e número de projetos especiais

Na Figura 8 podemos ver que 70% dos funcionários não participam de projetos especiais. Dos que participam de projetos, 16 funcionários participam de 6 projetos especiais. Temos leves indícios de crescimento no índice de satisfação e performance em relação à quantidade de projetos especiais.

Figura 8: Relação entre performance e número de projetos especiais

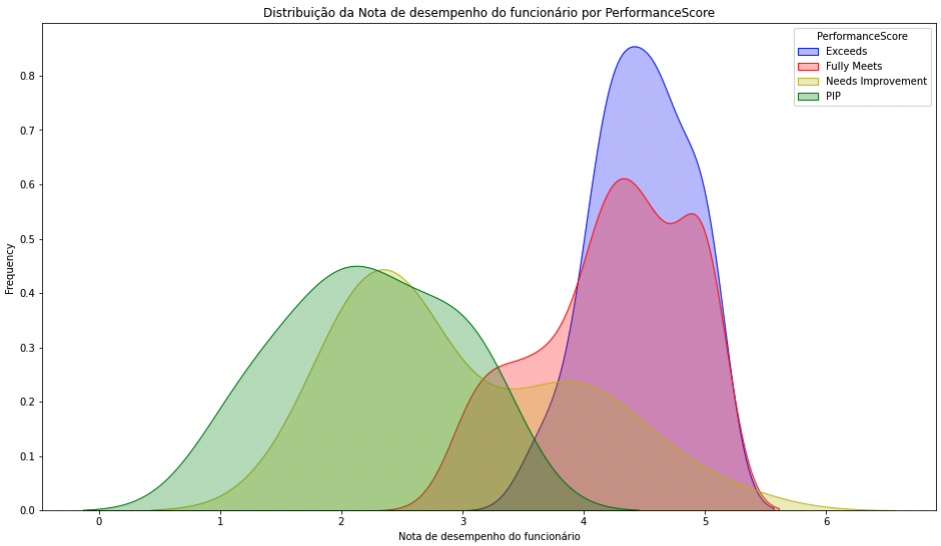


Fonte: Autoria própria.

Relação entre performance e Engajamento

Obtemos uma distribuição de poucos performáticos com os que são pouco engajados e alto performáticos para os que são muito engajados. A Figura 9 mostra que funcionários com baixa performance tendem a ser menos engajados. Funcionários com alta performance tendem a ser mais engajados. O ponto mais elevado de performance para os funcionários que foram mais engajados está acima de 4.

Figura 9: Relação entre performance e Engajamento

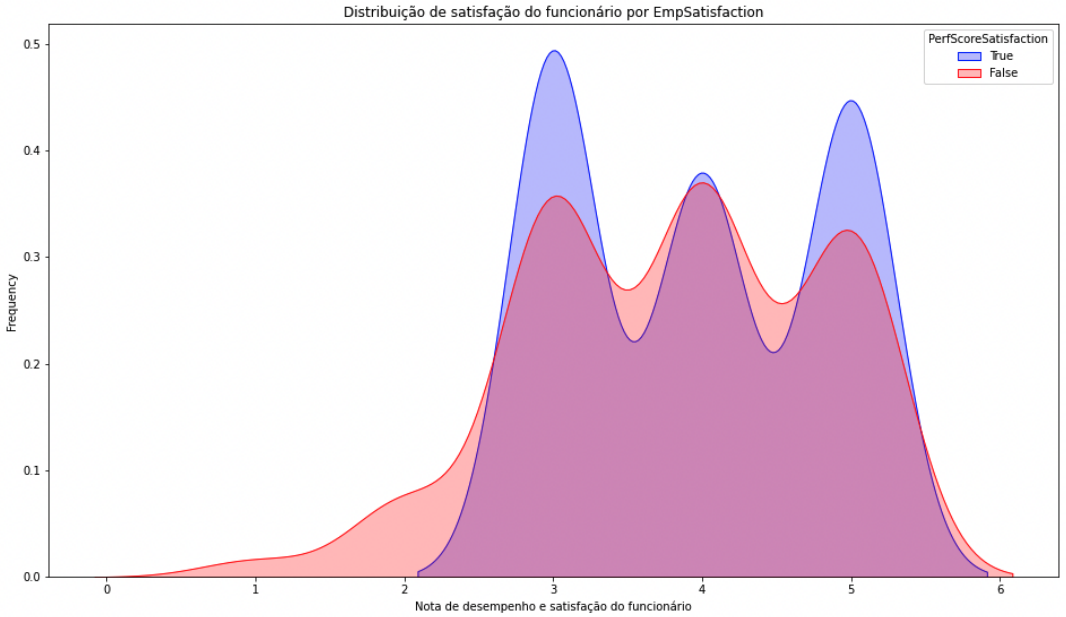


Fonte: Autoria própria.

Relação de performance com satisfação e satisfação

Observamos na Figura 10 uma distribuição multimodal para o conjunto dos que possuem boa performance e são engajados versos os de baixa performance e baixo engajamento. Poucos estão abaixo do PerfScoreSatisfaction 3 e as distribuições são similares. Há uma alta do PerfScoreSatisfaction em 3 e 5, sendo igual ao desempenho quando PerfScoreSatisfaction é 4.

Figura 10: Relação de performance com satisfação e satisfação

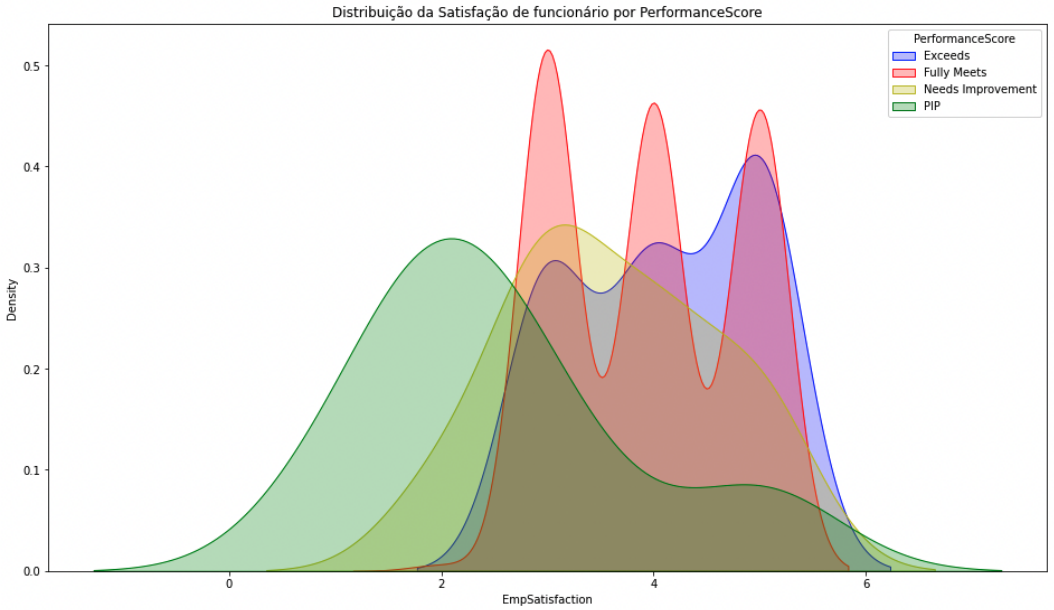


Fonte: Autoria própria.

Relação entre performance e satisfação

Funcionários com o nível de satisfação em 2 ou menos tendem a ter menos performance. Funcionários com o nível de satisfação acima de 3 tendem a ter mais performance. E funcionários com o nível de satisfação em 3 têm maior probabilidade de ser performáticos. Ver Figura 11.

Figura 11: Relação entre performance e satisfação



Fonte: Autoria própria.

Relação entre projetos especiais e performance

Como podemos constatar na Figura 12, há um aumento na satisfação para os funcionários que realizaram mais projetos especiais dentro do grupo de quem foi avaliado como ‘Exceeds’.

O grupo de funcionários com satisfação mediana, tiveram uma pontuação de performance semelhante, apesar do aumento da quantidade de projetos especiais.

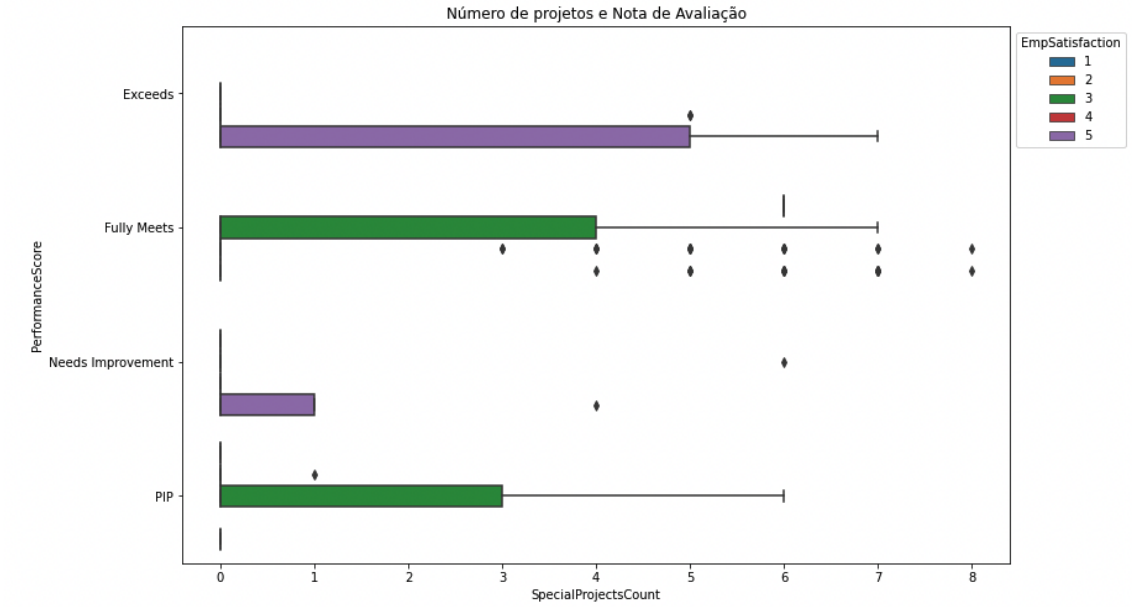
Funcionários que tinham dois projetos e uma avaliação péssima, saíram da empresa. Funcionários com menos de 3 projetos e avaliações baixas não se consideram insatisfeitos.

Tem outliers de performance ‘Fully meets’ com muitos projetos e satisfação média.

Há um valor pequeno de projetos para os de performance ‘Needs

Improvement', porém com satisfação alta.

Figura 12: Relação entre projetos especiais e performance

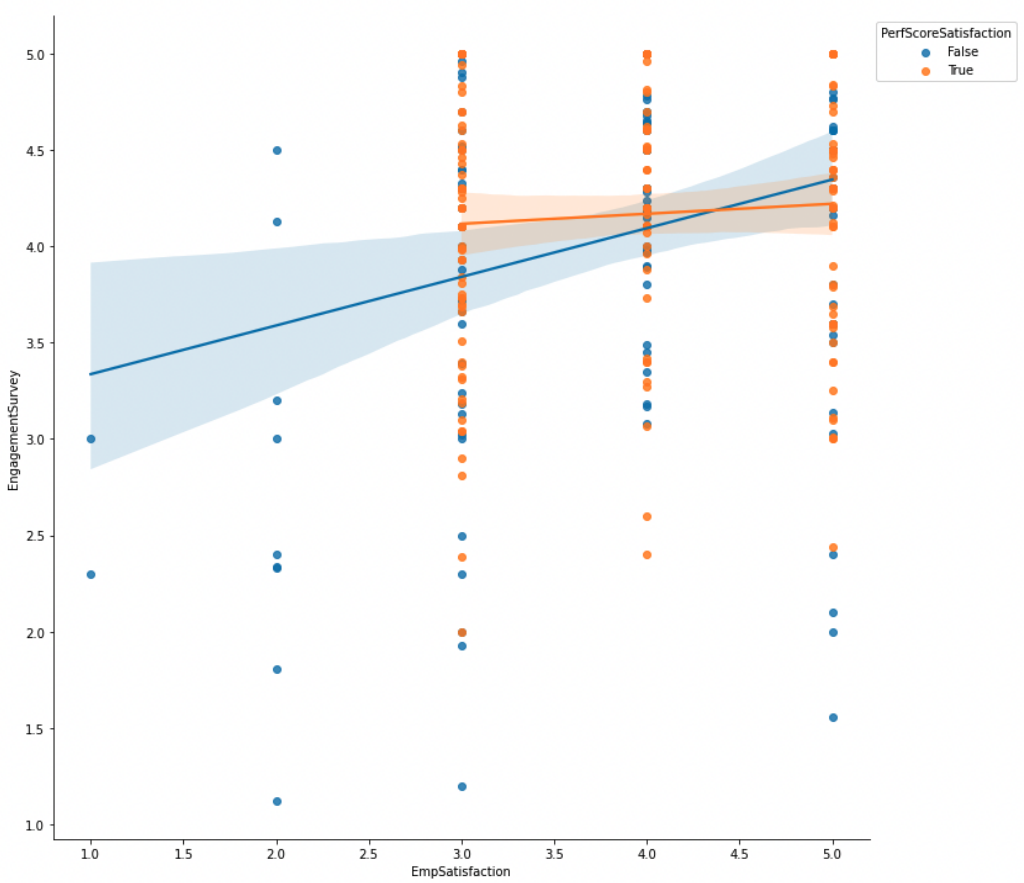


Fonte: Autoria própria.

Relação entre satisfação e engajamento

Temos na Figura 13 a visualização de 2 grupos sutis de funcionários: Cluster 1 é composto por funcionários com baixa performance e satisfação baixa. A satisfação abaixo de 3 e o engajamento inferior a 4 tendem a não satisfazer os requisitos de PerfScoreSatisfaction. O contrário também acontece nesse grupo, tem funcionários engajados e satisfeitos que não se encaixaram em PerfScoreSatisfaction; No Cluster 2 vamos ter os funcionários bons e satisfeitos. A satisfação está acima de 3 e as suas avaliações acima de 3 também.

Figura 13: Relação entre satisfação e engajamento



Fonte: Autoria própria.

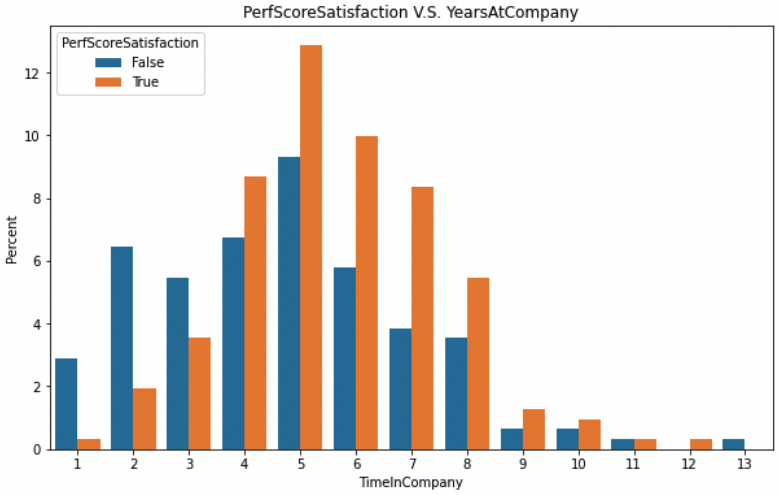
Relação entre PerfScoreSatisfaction e YearsAtCompany

Os funcionários com 5 a 8 anos de empresa são mais engajados e satisfeitos. Na Figura 14 se percebe que os 3 primeiros anos a maioria não consegue ter um grau de satisfação e engajamento satisfatório, entretanto, após 4 anos o quadro inverte e aparecem como maioria os mais engajados e satisfeitos, chegando ao topo com 5 anos. A diferença se mantém até 7 anos, quando começa a se igualar entre os dois grupos.

Os funcionários com 4 anos ou menos de empresa e os funcionários com mais de 9 anos de empresa devem ser analisados à parte para entender o que

houve com esses funcionários.

Figura 14: Relação entre PerfScoreSatisfaction e YearsAtCompany

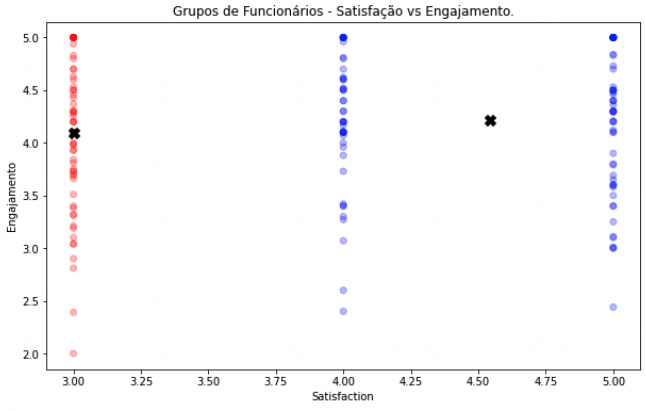


Fonte: Autoria própria.

Classificando os funcionários

Computando os clusters vamos definir um conjunto de dados apenas com o grupo de funcionários com PerfScoreSatisfaction igual a 'True'. Podemos visualizar na Figura 15 como ficou a classificação dos funcionários.

Figura 15: Classificando os funcionários



Fonte: Autoria própria.

Cluster 0 em azul são os funcionários engajados e muito satisfeitos. O Cluster 1 em vermelho são funcionários que atendem as demandas e razoavelmente insatisfeitos e em cor preta os seus respectivos centróides.

Desenvolvimento do modelo

Nesse momento carregamos o conjunto de dados que está pronto para trabalhar com a modelagem da machine learning, estamos prontos para gerar o modelo preditivo que vai estimar se um funcionário será engajado no seu trabalho e se estará satisfeito com a empresa. Esse é um dos problemas que propomos a resolver.

Iniciamos a etapa de pré processamento. Diante do problema que propomos resolver nesse projeto, trataremos com aprendizado supervisionado, pois criamos a informação se o funcionário é performático e satisfeito. O atributo que define isso é o PerfScoreSatisfaction definido para ‘True’ ou ‘False’. Usaremos essa classe para realizar uma classificação do tipo supervisionada passando esse atributo para o modelo de treinamento como *target*.

O atributo PerfScoreSatisfaction possui valores qualitativos classificados como categoria nominal, o que nos conduz a trabalhar com classificação supervisionada, temos um atributo que servirá de *target* para o modelo.

Pré Processamento de dados

Convertendo os atributos categóricos em dados numéricos, aplica-se uma técnica simples de transformação dos dados, transformando dados categóricos em dados binários por duas razões: é recomendável para muitos dos modelos de classificação que usamos em ciências de dados, onde se usa valores numéricos ou

binários como entrada; não sendo preparados para trabalhar com valores categóricos; os atributos categóricos tem pouca cardinalidade, tendo poucos registros únicos.

Uma vez que temos vários atributos em diferentes escalas, aplicou-se MimMaxScaler do sklearn para garantir a mesma escala para os atributos. Isso é importante no processo de convergência do algoritmo. Os cálculos internos que o algoritmo faz para gerar o modelo na mesma escala ajuda o modelo a ter uma performance melhor, a convergir com mais velocidade.

Apesar do conjunto de dados ter a classe target perfScoreSatisfaction balanceada, o conjunto de dados também tem outras classes desbalanceadas. Respeitando a distribuição de classes com uma amostra que reúne classes mais ou menos homogêneas entre si, usamos o parâmetro ‘stratify’ no momento que se particiona os dados para treino e teste.

Como temos à disposição vários algoritmos de classificação, selecionamos aleatoriamente alguns modelos para entender qual seria mais interessante a ser usado com esse conjunto de dados.

3. RESULTADOS

Treinamento e análise dos resultados

Pensando na satisfação e performance do funcionário, preparou testes para atributos relevantes para outros algoritmos. Se estabelece qual a métrica que vou levar em consideração para decidir se o algoritmo é eficiente.

Função do modelo base

Começando por criar um modelo baseline, para que os modelos a serem experimentados cheguem próximos do resultado do baseline. Se o modelo testado não tiver os resultados próximos da baseline, não faz sentido investir tempo nesse modelo. Além disso, a função da baseline é importante para conseguir explicar o resultado do modelo de forma intuitiva para pessoas que não entendem machine learning. Por exemplo, gerei um modelo de classificação de engajamento e satisfação que é superior à média, a média chegamos no resultado n, enquanto que nosso modelo atingiu n vezes 10. Os interessados conseguirão entender que o modelo ganha de algo simples que já é usado, como a média.

Figura 16: Resultados da função do modelo base

```
Base Model AUC = 0.5

              precision    recall  f1-score   support

           0       0.00      0.00      0.00        29
           1       0.54      1.00      0.70        34

    accuracy                           0.54        63
   macro avg       0.27      0.50      0.35        63
weighted avg       0.29      0.54      0.38        63

Test Result:
------------
Confusion Matrix:
 [[ 0 29]
 [ 0 34]]

accuracy score: 0.54
```

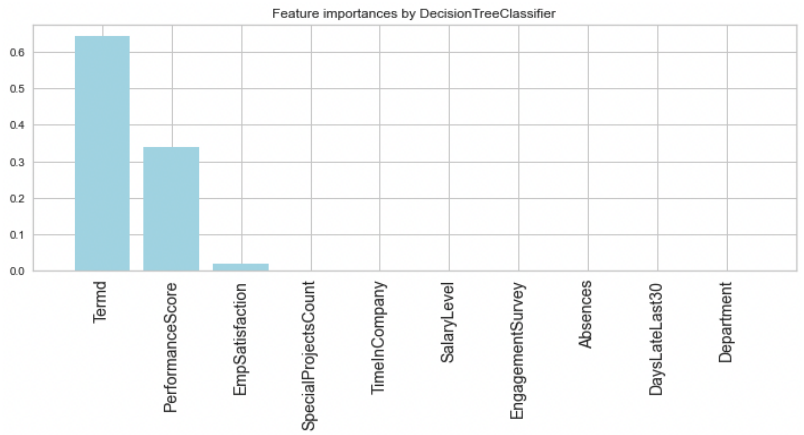
Fonte: Autoria própria.

Na Figura 16 os resultados mostram que o baseline não acertou muito ou só acertou somente os funcionários satisfeitos e engajados. A matriz de confusão dá mais detalhes, confirmando que o modelo baseline realmente não está tão bem. Para a classe 1 o baseline tem uma precisão de 54% que é exatamente a probabilidade de funcionários engajados e satisfeitos na empresa. A classe 0, não acertou nada. A baseline só acerta a classe 1. O Base Model AUC deu 54%, pois só acertou uma das duas classes.

Algoritmo de Árvore de Decisão

Tendo em vista que a quantidade de atributos influência nos processos e complexidades dos modelos, o scikit-learn têm em alguns algoritmos um parâmetro chamado feature_importances_. Iremos usar esse parâmetro para classificar a relevância de um atributo para os modelos. Treinamos os modelos e verificamos quais atributos são mais importantes para o modelo criado, que por sua vez retorna uma listagem de atributos e sua relevância, ver Figura 17. Em ambiente de produção é necessário mais experimentos com outros algoritmos para selecionar as features importantes.

Figura 17: Gráfico exemplo de features importantes



Fonte: Autoria própria.

No modelo de árvore de decisão os atributos mais relevantes são Termd e PerfomanceScore com uma indicação pequena para EmpSatisfaction.

Figura 18: Resultados do Algoritmo de Árvore de Decisão

Decision Tree AUC = 0.9655172413793103

	precision	recall	f1-score	support
0	1.00	0.93	0.96	29
1	0.94	1.00	0.97	34
accuracy			0.97	63
macro avg	0.97	0.97	0.97	63
weighted avg	0.97	0.97	0.97	63

Confusion Matrix:
[[27 2]
[0 34]]

accuracy score: 0.97

Fonte: Autoria própria.

Na Figura 18, a árvore de decisão tem acurácia maior, Decision Tree AUC = 0.96 e aumentou a precisão das duas classes comparado com outros modelos.

Algoritmo de Regressão Logística

No modelo de regressão logística as features mais relevantes são DaysLateLast30, Absences e Department.

Figura 19: Resultados do Algoritmo de Regressão Logística

Logistic Regression AUC = 0.539553752535497

	precision	recall	f1-score	support
0	0.67	0.14	0.23	29
1	0.56	0.94	0.70	34
accuracy			0.57	63
macro avg	0.61	0.54	0.47	63
weighted avg	0.61	0.57	0.48	63

Confusion Matrix:
[[4 25]
[2 32]]

accuracy score: 0.57

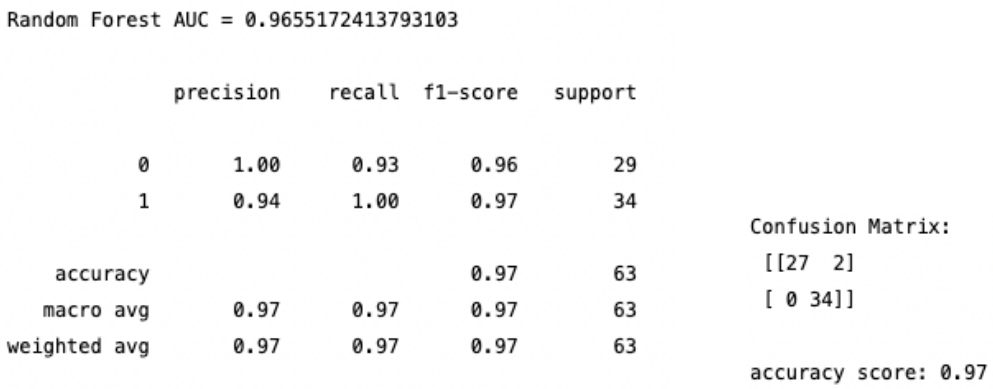
Fonte: Autoria própria.

A acurácia foi semelhante a do baseline, o Logistic Regression AUC = 0.57. Para a classe 0 o modelo de regressão logística acerta 67% enquanto que para classe 1 acerta 56%.

Algoritmo de Floresta Aleatória

No modelo de florestas aleatórias as features mais relevantes são Termd, PerformanceScore, EngagementSurvey e TimeInCompany.

Figura 20: Resultados do Algoritmo de Floresta Aleatória



Fonte: Autoria própria.

O modelo de Floresta aleatória alcançou 97% de acurácia para a classe 1 e 100 a classe 0. Iremos usar esse modelo levando em consideração os resultados na Figura 20, a assertividade desse modelo desenvolvido está satisfatória.

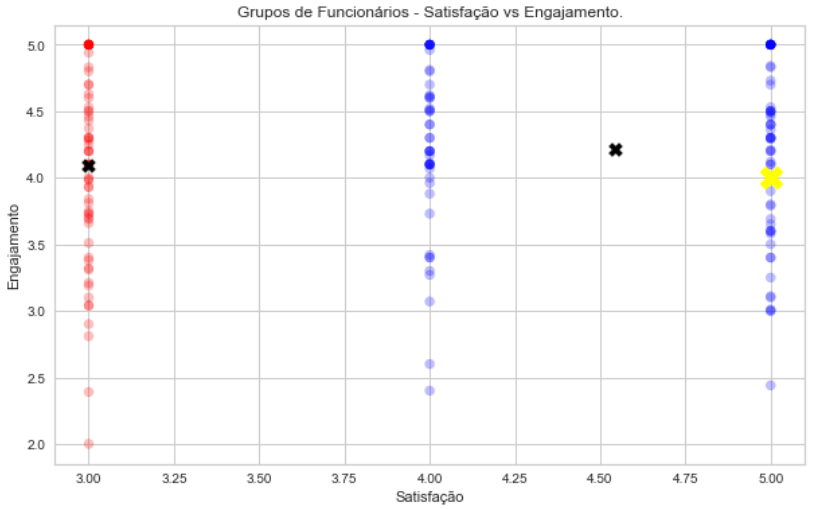
Realização de testes do modelo

Após testarmos os modelos, seguimos com o Random Forest na escolha do

modelo para um ambiente de produção futuro. Foram realizados inúmeros testes dos quais vamos exemplificar um de cada classe.

Atribuímos um perfil que seria considerado um funcionário engajado e satisfeito com os seguintes valores: daysLateLast30 = 0, absences = 0, empSatisfaction = 5, engagementSurvey = 4, performanceScore =4, salaryLevel = 4, timeInCompany = 5, specialProjectsCount = 6. A predição do modelo para a amostra de teste é de funcionário satisfeito e engajado com probabilidade de: 74.00%. Na Figura 21 temos a sinalização dele em amarelo.

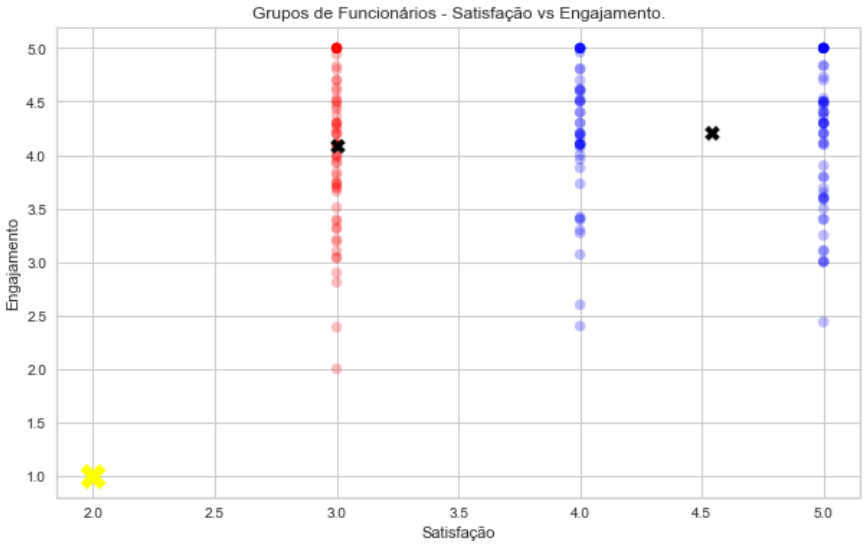
Figura 21: Gráfico da amostra de funcionário satisfeito e engajado



Fonte: Autoria própria.

Com outra amostra de perfil mais provável como insatisfeito e desinteressado: daysLateLast30 = 10, absences = 15, empSatisfaction = 2, engagementSurvey = 1, performanceScore = 2, salaryLevel = 1, timeInCompany = 3, specialProjectsCount = 0. A predição do modelo para a amostra de teste é de funcionário não satisfeito e não engajado com probabilidade de: 69.00%. Como sinalizado na Figura 22 em amarelo.

Figura 22: Gráfico da amostra de funcionário insatisfeito e não engajado



Fonte: Autoria própria

4. CONCLUSÃO E TRABALHOS FUTUROS

O modelo proposto tem uma boa assertividade e pode ser implementado futuramente em um ambiente de produção. A sugestão do autor é desenvolver um aplicativo para disponibilizar uma interface de consumo para o modelo, ou seja, pegar o resultado final e disponibilizar para o usuário final.

A análise realizada atende os requisitos para o modelo proposto e constrói um Modelo para responder às dúvidas e gerar soluções de recursos humanos. No caso de uma pesquisa em uma empresa, é importante ter um planejamento que responda a algumas questões, como por exemplo as respostas dos atributos do conjunto de dados: Quantas vezes o funcionário atrazou? Quantas vezes o funcionário faltou? Como é a satisfação do funcionário diante de suas condições de trabalho? Ele é considerado um funcionário engajado? O funcionário atende às expectativas da empresa? Tem salário compatível com o mercado? Está sobrecarregado de atividades relacionadas ao trabalho?

Os próximos passos são identificar lacunas para entendimento de alguns fatores que podem contribuir com o resultado: Quais fatores influenciam para um funcionário avaliar mal a empresa? Como reter pessoas engajadas? Podemos nos antecipar e saber se um determinado colaborador vai sair da empresa?

REFERÊNCIAS

ALEXANDER, Charles P. The New Economy. **Time**, 1983. Disponível em: <http://content.time.com/time/subscriber/article/0,33009,926013,00.html>. Acesso em: 9 jul. 2022.

HUEBNER, Rich; PATALANO, Carla. **(9) How do you get started on your first HR analytics project? | LinkedIn**. How do you get started on your first HR analytics project? Disponível em: <https://www.linkedin.com/pulse/steps-get-started-your-first-hr-analytics-project-rich-huebner-phd/>. Acesso em: 28 jul. 2022.

INC, MinIO. **MinIO | Code and downloads to create high performance object storage**. MinIO. Disponível em: <https://min.io>. Acesso em: 17 ago. 2022.
SCANLAN, Katya. Employee Job Satisfaction and Engagement: Revitalizing a Changing Workforce. p. 68, 2016.

Docker Desktop. Docker Documentation. Disponível em: <https://docs.docker.com/desktop/>. Acesso em: 17 ago. 2022.

Human Resources Data Set. Disponível em: <https://www.kaggle.com/datasets/rhuebner/human-resources-data-set>. Acesso em: 4 jul. 2022.

pandas - Python Data Analysis Library. Disponível em: <https://pandas.pydata.org/>. Acesso em: 17 ago. 2022.

Quick Start — Airflow Documentation. Disponível em: <https://airflow.apache.org/docs/apache-airflow/stable/start/index.html>. Acesso em: 17 ago. 2022.

RPubs - HR Dataset Codebook v14. Disponível em: <https://rpubs.com/rhuebner/hrd_cb_v14>. Acesso em: 4 jul. 2022.

scikit-learn: machine learning in Python — scikit-learn 1.1.2 documentation. Disponível em: <https://scikit-learn.org/stable/>. Acesso em: 17 ago. 2022.

Welcome to Python.org. Python.org. Disponível em: <https://www.python.org/>. Acesso em: 17 ago. 2022.