Proposta de Projeto Final

Samara Alvarez Alves
19 de dezembro de 2018

1 INTRODUÇÃO

1.1 HISTÓRICO DO ASSUNTO

No mundo corporativo a análise do cumprimento das metas dos funcionários ao final do exercício é muito importante para medir o desempenho dos processos de uma empresa e, com essas informações, colaborar para que alcance seus objetivos [1]. Um indicador importante nessa análise é o KPI, sigla para o termo em inglês Key Performance Indicator, que significa indicador chave de desempenho [2]. Através dos resultados apontados por essa medida é possível quantificar o desempenho da empresa e, assim, permitir que os trabalhadores entendam o quanto cada uma das suas atividades colaboram para o sucesso desses números.

Muitas propostas utilizando técnicas de mineração de dados foram propostas no contexto de avaliação de processos que associam análises de desempenho com os indicadores KPI. No artigo [1], por exemplo, uma abordagem de mineração de dados é aplicada usando regras de associação para desenvolver os KPIs integrados para uma empresa taiwanesa de moldes ópticos. No artigo [2], por sua vez, os autores comparam as performance de modelos de redes neurais, técnicas fuzzy e clustering para prever KPIs na gestão de projetos das empresas.

1.2 ENUNCIAÇÃO DO PROBLEMA

Esse projeto irá explorar dados reais da Companhia de Bebidas das Américas - Ambev. O objetivo será gerar visualizações que buscam resolver o problema de negócio munindo os

[1] https://www.quora.com/When-and-how-should-I-introduce-KPIs-to-the-employees-of-my-tech-startup
[2] https://www.internetinnovation.com.br/blog/saiba-o-que-e-kpi-e-entenda-sua-importancia-para-analises-em-marketing-digital/

responsáveis pelas tomadas de decisão com estratégias que exploram modelos e técnicas de mineração de dados.

Nesse contexto, o presente projeto estuda modelos de predição do percentual de cumprimento da meta ao fim do exercício de 2017 para cada funcionário, além de avaliar os desdobramentos dessas metas por nível hierárquico e entender o impacto e as relações entre o cumprimento delas e as regiões e áreas de atuação da empresa.

1.3 Conjunto de Dados e Inputs

A base de dados foi fornecida pela Ambev para o data challenge 2018 em parceria com a Udacity. O conjunto de dados contém 270.633 linhas e 38 colunas. As variáveis que compõem esse conjunto de dados e as suas categorias e tipos estão elencadas na Tabela 1.1. A última variável da tabela, *'Status Meta'*, atribui o valor 1 caso a meta tenha o seu "Monitoramento Aprovado"na base de dados original e 0 caso contrário. Cabe destacar também que a variável *target*, *'%Pontos Fim Exer'*, indica os percentuais finais cumpridos por meta: 0, 20, 60, 80 e 100%. Na Tabela 1.2 são listadas as variáveis presentes no conjunto inicial de dados que não serão incluídas nas análises pois não é possível corrigir a quantidade de inconsistências presente. Por fim, cabe esclarecer que cada linha da base de dados está relacionada a cada meta do funcionário.

Variáveis	Categoria	Keep	Tipo
'Pais'	nominal	True	object
'Mundo'	nominal	True	object
'Regiao/Area'	interval	True	float64
'Unidade'	interval	True	float64
'Grupo Cargo'	nominal	True	object
'Cargo'	nominal	True	object
'Grade'	interval	True	float64
'Banda'	nominal	True	object
'Area'	nominal	True	object
'Funcionario'	interval	True	float64
'Gestor'	interval	True	float64
'Codigo KPI'	nominal	True	object
'Diretoria'	nominal	True	object
'Areas Diretoria'	nominal	True	object
'Funcoes'	nominal	True	object
'Tipo Meta'	nominal	True	object
'Categoria KPI'	nominal	True	object
'Nome KPI'	interval	True	float64
'Peso KPI'	interval	True	float64
'%Acum Acumulado'	interval	True	float64
'%Ating Fim Exer'	interval	True	float64
'%Pontos Fim Exer'	interval	True	float64
'%Acum Fim Exer'	interval	True	float64
'Status Meta'**	binary	True	int64

Tabela 1.1: Variáveis de estudo.

Variáveis	Categoria	Keep	Tipo
'Mes'	interval	False	float64
'Prazo'	nominal	False	object
'Regra'	nominal	False	object
'Meta Projeto'	nominal	False	object
'%Ating Metas'	nominal	False	object
'%Pontos Metas'	nominal	False	object
'%Acum Mes'	nominal	False	object
'%Ating Acumulado'	nominal	False	object
'%Pontos Acum'	nominal	False	object

Tabela 1.2: Variáveis não incluídas no estudo.

1.4 PROPOSTA

Este trabalho propõe explorar o cumprimento das metas em vários níveis. Primeiro o conjunto de dados será agrupado por pontos percentuais ao final do exercício por funcionário e por gerente. O objetivo desses agrupamentos é preparar as bases para a construção de modelos de previsão do cumprimento das metas. Para tanto, serão utilizados modelos de classificação nos quais será entendido que a meta foi cumprida se o percentual de pontos por funcionário ou gerente é maior que 70%. Além disso, serão selecionados os gerentes com o maior número de metas cumpridas com o objetivo de estudar a distribuição das suas metas para correlacionar com o comportamento das distribuições das metas dos seus funcionários.

Por fim, serão explorados modelos de clusterização para visualizar a relação das metas pelas categorias: País, Áreas da Diretoria, Diretoria, Tipo de Metas, Região/Área, Bandas e Grades. Espera-se que essas análises permitam uma melhor visualização do conjunto de dados com vista a explorar os modelos de aprendizado de máquinas para permitir prever o cumprimento das metas por funcionário de acordo com cada uma das categorias estudadas.

2 METODOLOGIA

2.1 MODELO DE BENCHMARK

Primeiro será definido um preditor Naive que sempre prediz o cumprimento ou não da meta ao final do exercício, ou seja, ele irá prever que a meta será cumprida com 0 ou 100%. Para isso, será considerado que a meta foi cumprida quando a média dos percentuais de pontos ao final do exercício está acima de 70%. O objetivo desse primeiro modelo é simplesmente exibir como um modelo sem nenhuma inteligência se comportaria.

2.2 DESIGN DO PROJETO

Nessa seção serão discutidas as principais etapas do projeto: análise, limpeza, preprocessamento, agrupamento, aplicação dos modelos de classificação e clusterização e as suas

respectivas métricas de avaliação, representadas pela Figura 2.1. De maneira a coletar informações dos dados antes de aplicar o modelo preditivo a base de dados será dividida em conjuntos de treino e teste, com os percentuais de 80% para treino e 20% para teste. Na base de treino são processadas análises iniciais, tais quais: visualização das classes desequilibradas das colunas e cálculo de estatísticas descritivas básicas para as variáveis do tipo intervalar, Tabela 1.1. Em seguida, na etapa de limpeza da base serão eliminadas as colunas desnecessárias, que não serão úteis na análise e na previsão, presentes na Tabela 1.2.

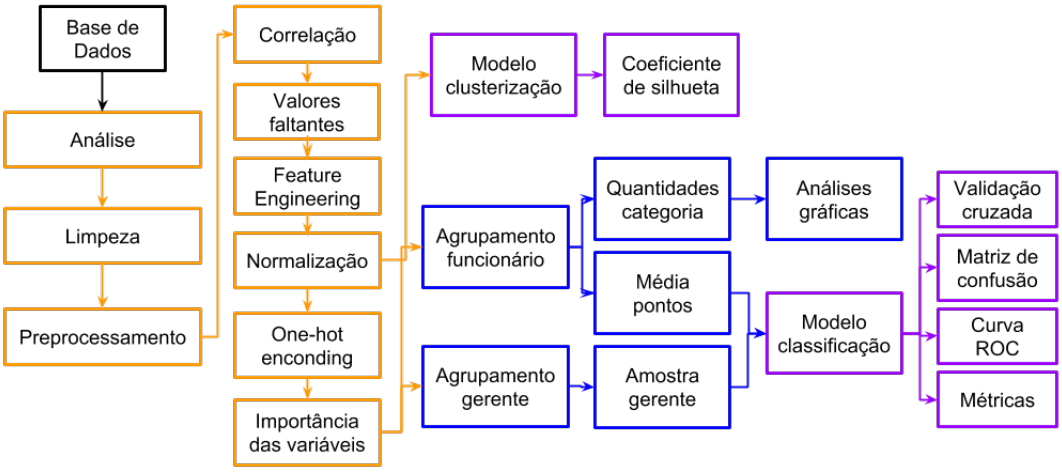

Figura 2.1: Esquema do Projeto.

A etapa de preprocessamento será divididas em subetapas. Primeiro, serão avaliadas as correlações das colunas com a presença de dados faltantes, que serão posteriormente preenchidos com valores 0; com a moda ou a mediana de cada coluna; ou com valores -1 para penalizar a inconsistência da base. A escolha desse preenchimento levará em consideração a análise da distribuição desses valores em cada categoria.

Em seguida, serão avaliadas as distribuições das colunas de variáveis intervalares, por meio de histogramas, com o objetivo de verificar se um processo de normalização dos atributos se faz necessário. Nessa etapa, serão desenvolvidas análises visuais também das demais categorias de acordo com a exploração de dados mais apropriada. Para a escolha dos melhores recursos visuais utilizaremos informações do site *data to viz* [3]. Ao final dessa etapa serão avaliadas as possibilidade de *Feature Engineering* das colunas [4]. Nas duas próximas etapas serão realizados procedimentos de *one-hot encoding* e análise das importâncias das *features* utilizando o algoritmo *random florest*.

O objetivo da próxima etapa é agrupar o conjunto de dados de duas maneiras: por gerente e por funcionário. No agrupamento por gerente, será selecionada uma amostra dos gerentes com o maior número de metas cumpridas de maneira a correlacionar com o comportamento das metas dos seus funcionários. Já no agrupamento por funcionário, serão calculadas as

[3] https://www.data-to-viz.com/

[4] https://towardsdatascience.com/understanding-feature-engineering-part-1-continuous-numeric-data-da4e47099a7b

médias de pontos obtidos ao final do exercício e a média ponderada dos KPI, pelos seus respectivos pesos. Além desses cálculos, serão adicionados agrupamentos pelos atributos país, mundo, região/área, unidade, grupo cargo, grade, banda, diretoria, área da diretoria e tipo meta para obter outros tipos de visualização da base de dados.

Na última etapa serão explorados modelos *baselines* de classificação e de clusterização. Os modelos de classificação terão como objetivo prever o percentual de meta por cada agrupamento da etapa anterior para prever o percentual de pontos ao final do exercício, enquanto que os modelos de clusterização endereçarão as relações entre as os atributos e o cumprimento da meta dos funcionários. Cabe esclarecer que a meta será considerada cumprida caso o percentual médio de pontos seja superior a 70%. Os modelos de clusterização a serem testados serão: o *K-means* e o modelo de mistura gaussiana. Já os modelos de classificação serão: *random florest*, *Adaboost*, *SVM*, *XGboost* e regressao logistica. Por fim, os melhores candidatos aos modelos terão seus parâmetros otimizados.

ALTERAR: *EXPLICAR OS MODELOS descrever cada um dos modelos que você usará ao seu leitor (prós, contras e porque ele potencialmente resolve o problema) https://scikit-learn.org/stable/tutorial/machine¡ear*

2.3 Avaliação Métrica

De maneira a avaliar e comparar o modelo naive e os modelos baseline serão utilizados a acurácia, a precisão, a sensibilidade (recall) e o f-score, expresso na fórmula abaixo. As classes para os cálculos dessas métricas são: cumprimento da meta (caso a média de pontos percentuais na categoria é maior que 70%) ou não cumprimento (caso contrário). Por fim, para os melhores candidatos aos modelos será estudada a curva ROC.

$$fscore = \frac{precision \cdot recall}{precision + recall}$$

Referências

[1] *S.-H. Liao and P.-Y. Hsiao, "Mining business knowledge for developing integrated key performance indicators on an optical mould firm,"* International Journal of Computer Integrated Manufacturing, *vol. 26, no. 8, pp. 703–719, 2013.*

[2] *S. S. Fanaei, O. Moselhi, S. T. Alkass, and Z. Zangenehmadar, "Application of machine learning in predicting key performance indicators for construction projects,"* methods, *vol. 5, no. 09, 2018.*