IFPE Campus Paulista

Curso Tecnológico em Análise e Desenvolvimento de Sistemas

Mineração de Dados

Relatório do Projeto da 2ª Unidade

Equipe: Ana Clara Fontes, Heitor Negromonte e Matheus Felipe
Base: Predicting Hiring Decisions in Recruitment Data - https://www.kaggle.com/datasets/rabieelkharoua/predicting-hiring-decisions-in-recruitment-data

1. Compreensão do Negócio

a. Objetivo de Negócio

A partir de dificuldades em selecionar candidatos adequados, resultando em altos custos de contratação, elevado tempo de preenchimento de vagas, e uma taxa considerável de turnover dentro do primeiro ano de emprego, uma empresa visa implementar um projeto de mineração de dados para otimizar o processo de seleção de candidatos e, consequentemente, melhorar a qualidade das contratações.

b. Objetivo de Mineração:

O objetivo da mineração de dados é desenvolver um modelo preditivo que ajude a classificar candidatos em duas categorias: aptos e não aptos para a contratação. Este modelo será baseado em características e dados de recrutamento dos candidatos, tendo foco em classificar corretamente candidatos aptos, sendo crucial para atingir o objetivo de negócio da empresa.

c. Critérios de Sucesso Mineração:

Não foi possível localizar benchmarks específicos para este problema utilizando técnicas semelhantes às deste projeto. Portanto, serão realizados experimentos de linha de base na fase de modelagem, começando com modelos mais simples (KNN, Regressão Logística e Árvore de Decisão) antes de treinar modelos mais complexos (Random Forest e XGBoost). Tendo como base os resultados obtidos primeiramente no conjunto de treinamento para avaliar o desempenho final dos modelos no conjunto de teste.

2. Compreensão dos Dados

Descrição dos dados:

Variável	Descrição	Tipo do dado	Intervalo/Categoria
Age	Idade do candidato.	Inteiro	20 a 50 anos
Gender	Gênero do	Binário	0: Masculino;

	candidato.		1: Feminino
Education Level	Nível mais alto de educação alcançado pelo candidato.	Categórico	1: Bacharelado (Tipo 1); 2: Bacharelado (Tipo 2); 3: Mestrado; 4: Doutorado
Experience Years	Número de anos de experiência profissional.	Inteiro	0 a 15 anos
Previous Companies Worked	Número de empresas anteriores onde o candidato trabalhou.	Inteiro	1 a 5 empresas
Distance From Company	Distância em quilômetros da residência do candidato até a empresa contratante.	Float (contínuo)	1 a 50 quilômetros
Interview Score	Pontuação alcançada pelo candidato no processo de entrevista.	Inteiro	0 a 100
Skill Score	Pontuação de avaliação das habilidades técnicas do candidato.	Inteiro	0 a 100
Personality Score	Pontuação de avaliação dos traços de personalidade do candidato.	Inteiro	0 a 100
Recruitment Strategy	Estratégia adotada pela equipe de contratação para recrutamento.	Categórico	1: Agressiva; 2: Moderada; 3: Conservadora
Hiring Decision (variável alvo)	Resultado da decisão de contratação	Binário	0: Não contratado; 1: Contratado

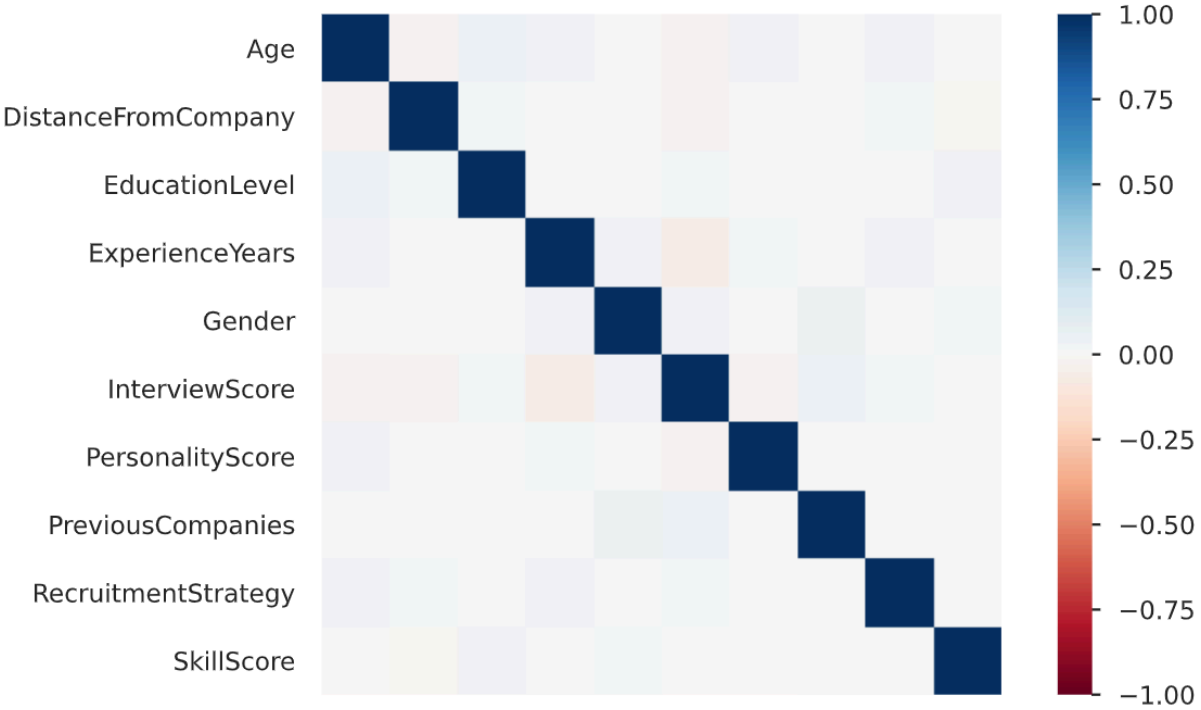
Através da descrição dos dados, é possível perceber que os dados da base possuem, além do tipo binário, valores do tipo categórico e numérico. Isto fará necessário realizar o tratamento desses dados utilizando técnicas de pré-processamento diferentes e adequadas

a cada tipo de dado para garantir que os modelos possam ser treinados de maneira eficaz e fornecer dados precisos.

Além da observação citada, após obter uma análise mais detalhada através do Pandas Profiling, foi visto que a base em questão não possui dados faltantes ou duplicados e não há alta correlação entre os dados.

Dataset statistics

Number of variables	10
Number of observations	1500
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	117.3 KiB
Average record size in memory	80.1 B



3. Preparação dos Dados

Nesta etapa, foi preciso realizar o tratamento de variáveis categóricas e numéricas presentes na base de dados. As variáveis categóricas foram tratadas utilizando a técnica de One-Hot Encoding, criando novas colunas binárias (0 ou 1) para cada categoria possível. Já para as variáveis numéricas, foi feita a discretização, convertendo-as em categóricas ao dividir o intervalo contínuo em várias categorias (bins). O One-Hot Encoding também foi aplicado após a discretização dessas variáveis.

A base selecionada, sem a variável de saída, estava com 10 colunas. Em consequência da técnica One-Hot Encoding, a base cresceu em sua dimensão para 35 colunas. Mediante isso, decidimos que não seria necessário realizar a redução de dimensionalidade, pois apesar da quantidade de colunas, a base possui 1500 observações, o que implica que a quantidade de dados é suficiente, pois temos além de 10 vezes mais observações do que variáveis.

Dito isso, os resultados foram:

- Quantidade de Atributos: Após o pré-processamento, há 35 atributos.
- Quantidade de Padrões: O conjunto de dados final contém 1500 padrões.
- Distribuição das Classes: Sendo 0 para não contratados e 1 para contratados, ficaram 1035 padrões para a classe 0 e 465 para a classe 1.

4. Modelagem

Devido ao fato de haver muito mais valores para a classe 0 em detrimento da classe 1, ou seja, um certo desbalanceamento de classes, foi decidido que seria útil utilizar a técnica de Validação Cruzada no treinamento dos modelos.

Porém, a função cross_val_score, por padrão, separa os dados em partes (folds) de forma direta e não proporcional. Isso pode ser ajustado através do parâmetro ‘cv’, onde, ao invés de passar o número de folds, podemos passar uma função que separa os folds de outras maneiras. Portanto, para “misturar” as informações e garantir que a proporção de candidatos contratados e não contratados seja a mesma em todos os folds da validação, utilizamos a função StratifiedKFold e passamos o parâmetro shuffle como True.

4.a. Técnicas utilizadas para modelagem:

Utilizamos Regressão Logística, Árvore de Decisão, KNN, Random Forest e XGBoost. Tais técnicas foram selecionadas pois são as mais comuns entre problemas de classificação.

4.b. Configurações experimentais utilizadas

70% dos dados foram utilizados para treinamento e 30% para teste. Hiperparâmetros não foram otimizados, pois os valores obtidos com suas configurações padrão já foram considerados satisfatórios. Porém, ao final da análise, percebemos que poderíamos ter considerado a otimização do modelo KNN. Utilizamos o parâmetro “random_state” na separação de treino e teste e no treinamento dos modelos para que o resultado não fosse alterado no caso de precisar rodar os códigos novamente.

4.c. Métricas utilizadas para a análise:

Matriz de confusão, acurácia, precisão, recall e F1-Score.

4. d. Avaliação dos modelos e métricas:

Se tratando das métricas Precisão e Recall que atuam juntamente aos resultados da matriz de confusão, analisamos o que o valor de cada uma dessas métricas significa no contexto do objetivo de negócio definido previamente. Foi concluído que:

- Recall alto significa que o modelo está identificando quase todos os candidatos que são aptos.
- Recall Baixo significa que o modelo está identificando apenas uma pequena parte dos candidatos que realmente são aptos.
- Precisão Alta significa que quando o modelo classifica um candidato como apto, é muito provável que ele realmente seja apto.
- Precisão Baixa significa que muitos candidatos classificados como aptos não são realmente aptos (FP).

Após tal interpretação, definimos que o ideal seria encontrar um equilíbrio adequado entre precisão e recall. Isso pode ser avaliado com a métrica F1-Score, que leva em conta as duas métricas em questão. Portanto, além de avaliar a acurácia de cada modelo, terão melhor desempenho aqueles que atingirem bons valores na métrica F1 Score e mostrarem ter poucos erros em sua matriz de confusão. Devido aos conjuntos terem sido estratificados na validação cruzada, cada fold obteve um resultado. Logo, foi tirada a média de cada métrica. Abaixo serão apresentados os resultados obtidos em cada modelo.

ÁRVORE DE DECISÃO

	ACURÁCIA MÉDIA	**PRECISÃO MÉDIA**	**RECALL MÉDIO**	**F1-SCORE MÉDIO**	**MATRIZ DE CONFUSÃO**
TREINAMENTO	89.33%	83.47%	83.64%	83.42%	[[657 57] [55 281]]
TESTE	84%	70%	78%	74%	[[278 43] [28 101]]

Interpretação: A árvore de decisão tem um desempenho razoável, com uma leve queda de performance entre o treinamento e o teste, indicando possível overfitting. No teste, a precisão é um pouco baixa, o que significa que há alguns falsos positivos. O recall de 78% indica que o modelo consegue identificar a maioria dos candidatos da classe positiva.

REGRESSÃO LOGÍSTICA

	ACURÁCIA MÉDIA	**PRECISÃO MÉDIA**	**RECALL MÉDIO**	**F1-SCORE MÉDIO**	**MATRIZ DE CONFUSÃO**
TREINAMENTO	84.29%	77.79%	71.43%	74.32%	[[645 69] [96 240]]

TESTE	88%	80%	76%	78%	[[297 24] [31 98]]

Interpretação: A regressão logística apresenta um bom equilíbrio entre precisão e recall no conjunto de teste, com ambos sendo razoavelmente altos. O modelo não parece estar sofrendo de overfitting, pois a performance é consistente entre os conjuntos de treinamento e teste.

KNN

	ACURÁCIA MÉDIA	**PRECISÃO MÉDIA**	**RECALL MÉDIO**	**F1-SCORE MÉDIO**	**MATRIZ DE CONFUSÃO**
TREINAMENTO	67.90%	50.10%	36.61%	42.17%	[[590 124] [213 123]]
TESTE	68%	43%	39%	41%	[[256 65] [79 50]]

Interpretação: O KNN apresenta um desempenho inferior em comparação com outros modelos. A acurácia é baixa, assim como a precisão e o recall. O modelo está errando bastante ao classificar, resultando em um F1-Score baixo. Pode ser um indicativo de que o KNN não é adequado para este problema ou que os parâmetros precisam ser ajustados.

RANDOM FOREST

	ACURÁCIA MÉDIA	**PRECISÃO MÉDIA**	**RECALL MÉDIO**	**F1-SCORE MÉDIO**	**MATRIZ DE CONFUSÃO**
TREINAMENTO	91.71%	90.51%	83.02%	86.52%	[[684 30] [57 279]]
TESTE	93%	93%	81%	86%	[[313 8] [25 104]]

Interpretação: O Random Forest apresenta um desempenho robusto tanto em precisão quanto em recall no conjunto de teste. O modelo mantém alta acurácia e não parece estar sofrendo de overfitting, sendo um dos melhores modelos em termos de equilíbrio entre precisão e recall.

XGBOOST

	ACURÁCIA MÉDIA	**PRECISÃO MÉDIA**	**RECALL MÉDIO**	**F1-SCORE MÉDIO**	**MATRIZ DE CONFUSÃO**
TREINAMENTO	92%	90.35%	84.22%	87.09%	[[683 31] [53 283]]

TESTE	93%	91%	83%	87%	[[311 10] [22 107]]

Interpretação: O XGBoost também apresenta um excelente desempenho, com valores de precisão e recall altos no conjunto de teste. Ele manteve uma alta acurácia e, assim como o Random Forest, não parece sofrer de overfitting. É o melhor modelo em termos de F1-Score, indicando um bom equilíbrio entre precisão e recall.

5. Avaliação

A árvore de decisão e a regressão logística têm desempenhos moderados. A regressão logística, apesar de simples, apresenta um bom equilíbrio, enquanto a árvore de decisão mostra um leve overfitting. Já o KNN não se saiu bem, sugerindo que pode não ser o melhor para este problema. Enquanto isso, os modelos Random Forest e XGBoost são os modelos com melhor desempenho geral, apresentando alta acurácia, precisão, recall e F1-score. Eles seriam ideais para o objetivo do negócio, onde é importante um bom equilíbrio entre a identificação correta de candidatos aptos e a minimização de falsos positivos. Um desses modelos que poderia seguir para a etapa de implantação, seria o Random Forest, pois, se tratando de classificar corretamente mais candidatos aptos, foi o que mais se destacou.