Modelos de Regressão para Horas de Treinamento para Candidatos a Mudança de Emprego

Thaís C. Sampaio, João Pedro Campos
Departamento de Engenharia de Teleinformática
Universidade Federal do Ceará, Fortaleza, Brasil
Emails: {thaisc, joaopedroscampos}@alu.ufc.br

***Resumo*—Este trabalho trata do estudo e aplicação de modelos e métodos de regressão linear para predição das horas de treinamento de candidatos a mudança de emprego utilizando preditores de origem categórica pré-processados e tem o objetivo de comparação de diferentes métodos através da medida de performance, das suas características e dos parâmetros utilizados. Os resultados encontrados para esse banco de dados confirmam a suspeita inicial que os modelos de regressão linear seriam insuficientes, dado a baixa correlação entre os pares de preditores e entre os preditores e a saída horas de treinamento do candidato.**

***Palavras-chave*– regressão linear, aprendizado estatístico, pré-processamento.**

I. Introdução

Uma empresa pretende contratar cientistas de dados e deseja, através de análises estatísticas, criar um modelo preditivo para identificar, na etapa de inscrição e antes do treinamento da empresa, as horas de treinamentos completados anteriormente a inscrição dos candidatos. O treinamento faz parte do plano de gerenciamento de pessoas [1] e trás eficiência ao trabalho quando capacitam as pessoas a trabalharem em seus cargos exercendo suas funções [2].

Para isso, a empresa coletou informações de todos os candidatos, como experiências passadas, nível de educação e dados demográficos. Com posse desses dados, deseja-se verificar, também, a relação e a importância de cada preditor, fornecendo parâmetros que podem ser usados para tomada de decisão antes e depois do treinamento fornecido pela empresa e assim reduzir impactos em custo e tempo, além de melhorar a qualidade dos treinamentos.

Para esse problema serão comparados métodos de regressão, pois são utilizados para predição de variáveis quantitativas com valores numéricos como a saída horas de treinamento a ser analisada pela empresa. O melhor método de regressão quanto a performance na predição dependerá das características do banco de dados aumentando a importância da análise estatística e do pré-processamento dos dados [3].

A regressão tem como objetivo principal encontrar a relação entre as entradas que são as variáveis independentes e a saída que é a variável dependente. Essa relação é estimada através dos parâmetros que são calculados em um método supervisionado com a observação da saída real e da saída estimada. A diferença entre as saídas é chamada de erro sendo um problema de otimização com a minimização da função custo ou com a maximização da função de verossimilhança.

Existem vários tipos de regressão e suas performances em determinado conjunto de dados pode ser comparada através do calculo da raiz quadrada do erro médio, em inglês *Root Mean Square Error* (RMSE), e através do calculo do coeficiente de determinação R2. O métodos de regressão podem ser divididos em métodos lineares e não lineares. Dentre os métodos de regressão linear exitem a regressão linear múltipla, a regressão linear com penalização, que pode ser no tipo Lasso ou Ridge, e a regressão linear com redução de dimensionalidade, que pode ser do tipo PLS e PCR. Já um exemplo de método de regressão não linear é a regressão utilizando redes neurais.

Independente do método de regressão o conjunto de dados é dividido em conjunto de treino e em conjunto de teste. O conjunto de treino é usado para treinamento do modelo e estimação dos parâmetros e o conjunto de teste é usado para validação do modelo e avaliação da performance do método. A técnica de *cross-validation* é usada para avaliar os modelos obtidos, utilizando subconjuntos de dados de entrada disponíveis e avaliando-os através do subconjunto complementar desses dados. Após o treinamento, o modelo pode ter sobreajuste, conhecido como *overfitting*, quando tem alta variância e pode ter subajuste, conhecido como *underfitting*, quando tem alto enviesamento, conhecido como *bias*. Assim tem-se o compromisso (*trade-off*) do equilíbrio entre polarização e variância para encontrar a melhor estimativa.

II. Metodologia

O conjunto dados utilizado para predição de horas de treinamento *HR Analytics: Job Change of Data Scientists* está disponível no *Kaggle* [4] e possui dois arquivos, o conjunto de dados de treino e o conjunto de dados de teste que juntos têm $N = 21287$ observações e $D = 12$ preditores que estão caracterizados e descritos na Tabela I e na Tabela II, respectivamente.

Seguindo a análise dos dados e pré-processamento desenvolvido no artigo *Analise Estatística no Contexto de Predição de Mudança de Emprego* [5] foram escolhidos 10 preditores que representam as variáveis independentes da entrada X no modelo de regressão para estimar a saída Y. As estatísticas dos dados escolhidos estão na Tabela IV e a correlação entre os pares de preditores e entre os preditores e a saída está na Figura 1.

Após fabricar variáveis numéricas a partir das variáveis categóricas e tratar os elementos faltantes do conjunto de

Tabela I
CARACTERIZAÇÃO DAS COLUNAS DO BANCO DE DADOS

Índice	Coluna	Tipo de dados
0	enrolle_id	Categórico nominal
1	city	Categórico nominal
2	city_development_index	Numérico
3	gender	Categórico nominal
4	relevent_experience	Categórico binário
5	enrolled_university	Categórico nominal
6	education_level	Categórico ordinal
7	major_discipline	Categórico nominal
8	experience	Categórico ordinal
9	company_size	Categórico ordinal
10	company_type	Categórico nominal
11	last_new_job	Categórico ordinal
12	training_hours	Numérico
13	target	Numérico binário

Tabela II
DESCRIÇÃO DOS DADOS NAS COLUNAS

Índice	Descrição
0	Identificação para cada candidato
1	Identificação das cidades dos candidatos
2	Índice de desenvolvimento das cidades dos candidatos
3	Gênero ou sexo do candidato
4	Experiência relevante do candidato
5	Tipo de curso matriculado, se houver
6	Nível de escolaridade do candidato
7	Área de conhecimento principal da formação do candidato
8	Experiência total do candidato em anos
9	Número de funcionários da empresa atual, se tiver empregado
10	Tipo da empresa atual, se tiver empregado
11	Diferença em anos do último emprego
12	Horas de treinamentos completados
13	Classes sobre a decisão sobre aceitar ou não o novo emprego

Tabela III
ESTATÍSTICAS DO CONJUNTO DE DADOS PARA REGRESSÃO

Coluna	Média	Desvio Padrão	Obliquidade
city	43,967851	35,478183	0,407048
city_development_index	0,831290	0,122314	-1,037871
relevent_experience	0,731330	0,443279	-1,043825
enrolled_university	0,456270	0,799995	1,291069
education_level	2,141973	0,689908	-0,076953
major_discipline	4,226338	1,724523	-1,841350
experience	10,279353	6,769513	0,377661
company_size	3,011831	2,677144	0,476299
company_type	1,792682	1,524532	0,748344
last_new_job	2,027356	1,671263	0,786192
training_hours	65,299271	60,087097	1,827681

dados com criação de novas classes e exclusão das linhas com elementos faltantes, é necessário a divisão do conjunto de dados em dados de entrada X e dados de saída Y. O conjunto de dados de entrada serão utilizados para predição saída como na Equação 1.

$$\hat{Y} = f(X) + \epsilon \tag{1}$$

Onde $\hat{Y}$ representa os resultados da predição do modelo e ϵ é o erro irredutível. Depois é necessário fazer a transformação dos dados de entrada X consertando a assimetria dos dados usando a família de transformação Yeo-Johnson. Em seguida é necessário fazer a centralização e escalonamento dos dados X para construir o modelo de regressão sem enviesamento subtraindo a média e dividindo o resultado pelo desvio padrão resultando em média zero e variância unitária.

Com o pré-processamento feito em todo o conjunto de dados de entrada é necessário fazer a divisão em treino e teste e assim dividir os dados em duas partes, uma para treinar o modelo e outra para testar o modelo com dados que ele não conhece.

Para poder avaliar a performance e o desempenho de um método é necessário medir de alguma forma o quão bom é a sua predição. A medida mais comum para essa avaliação é o erro quadrático médio, em inglês *Mean Square Error* (MSE), que pode ser expresso pela Equação 2.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}(x_i)\right)^2 \tag{2}$$

Onde y é a saída original e $\hat{f}(x)$ é a predição feita da saída para cada n-ésima observação. Assim na Equação 2 se a saída predita for próxima a saída observada o MSE será pequeno e nesse caso o melhor método para determinado conjunto de dados é aquele que possui menor MSE. Porém baixo MSE no treinamento não significa baixo MSE no teste. Quando temos um método com pequeno MSE no treino, porém um grande MSE no teste isso é chamado de *overfitting* caracterizando um modelo com grande complexidade que está relacionado a encontrar relações e padrões específicos do banco de dados de treino e que podem ser resultado da aleatoriedade. O caso contrário ao descrito é denominado *underfitting*. Dito isto, a escolha dos melhores parâmetros de $\hat{f}(x)$ acontece através da otimização do erro e do compromisso entre o *bias* e a variância. No geral em métodos mais flexíveis a variância aumentará e o *bias* diminuirá.

Para confirmar as informações apresentadas acima podemos partir da Equação 2 e aplicar o operador linear Esperança decompondo em três termos: a variância de $\hat{f}(x_o)$, *bias* ao quadrado de $\hat{f}(x_o)$ e a variância do erro.

$$E\left\{\left(y_o - \hat{f}(x_o)\right)^2\right\} = Var\left(\hat{f}(x_o)\right) + \left(Bias\left(\hat{f}(x_o)\right)\right)^2 + Var(\epsilon) \tag{3}$$

Portanto podemos evidenciar e confirmar a partir da Equação 3 que para minimizar o valor médio do erro temos que escolher o método que simultaneamente minimize a variância e o *bias*.

A. Regressão linear

A regressão linear simples consiste em uma abordagem para prever uma resposta quantitativa Y dado um preditor X.

$$Y \approx \beta_0 + \beta_1 X \tag{4}$$

Os coeficientes β_0 e β_1 representam a interceptação e a inclinação do modelo linear. Treinando o modelo, podemos obter as estimativas $\hat{\beta}_0$ e $\hat{\beta}_1$, ou seja, a predição de Y dado $X = x$ é

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x, \tag{5}$$

Utilizamos os n dados de entrada e saída para descobrir os coeficientes β_0 e β_1, tal que, para n pares de observação, temos

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

Para $y_i \approx \beta_0 + \beta_1 x_i$, dado $i = 1, \ldots, n$, deseja-se encontrar β_0 e β_1 a fim de obter uma linha resultante mais próxima possível dos n dados. No caso de mais de um preditor, essa abordagem, entretanto, não é satisfatória, pois não leva em consideração a correlação entre os preditores. Assim, expande-se o conceito de regressão linear para acomodar múltiplos preditores, dando a cada preditor um coeficiente de inclinação em um único modelo. Trata-se do conceito de regressão linear múltipla que será discutido em seguida.

1) Regressão linear múltipla: Considerando p preditores distintos, este modelo é dado por

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon, \tag{6}$$

onde ϵ é o erro, dado por um ruído aleatório de média zero. As predições são feitas de forma similar às feitas na regressão simples, da forma

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p. \tag{7}$$

Usando o método dos mínimos quadrados [3], deve-se escolher β_0 e β_1 para minimizar a soma dos resíduos quadrados, que é,

$$\begin{aligned} \mathrm{RSS} &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip} \right)^2 \end{aligned}$$

Pode-se calcular esses coeficientes através de

$$\widehat{\beta} = \left(X^T X \right)^{-1} X^T Y. \tag{8}$$

2) Regressão com penalização: Os modelos de regressão com penalização são utilizados para reduzir o impacto de coeficientes muito grandes que possam instabilizar o modelo de regressão. É similar à Equação 8, porém é adicionado o parâmetro $\lambda \geq 0$ que define a penalidade aplicada ao modelo. Dessa forma, tem-se a seguinte relação:

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 = \mathrm{RSS} + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{9}$$

Diferente dos mínimos quadrados, que gera um conjunto de coeficientes estimados, esse tipo de regressão produz um conjunto de coeficientes $\hat{\beta}_\lambda^R$, para cada valor de λ. Selecionar um bom valor para λ é necessário, e para isso usamos a validação cruzada.

3) Cross-Validation: No caso de uma situação como a mencionada anteriormente, a validação cruzada pode ser útil para estimar o erro de teste associado ao método de aprendizado e assim avaliar seu desempenho. O erro de teste é definido como o erro médio do método de aprendizado estatístico, considerando medidas que não foram usadas durante o treinamento. Para estimar a taxa de erro de teste, utiliza-se um subconjunto das observações de treinamento e aplica-se o método de aprendizado estatístico às observações realizadas.

Alguns métodos de cross-validation são o *validation set sproach*, que divide o conjunto em dados de treino e validação, o *leave-one-out cross-validation* (LOOCV), que utiliza apenas uma única observação (x_1, y_1) para validar o conjunto restante, e o método *k-fold*, utilizado neste estudo.

Nesse último método, dividimos os dados de entrada em k subconjuntos de dados (ou folds) os quais são treinados, e o restante do conjunto de dados $(k-1)$ é utilizado para o teste do modelo. Esse processo é repetido k vezes, com um subconjunto diferente reservado para avaliação (e excluído do treinamento) a cada vez. Neste trabalho foi utilizado $k = 10$.

4) Regressão com redução de dimensionalidade (PLS e PCR): Para comparar a performance do modelo com o conjunto de dados de validação, utiliza-se *partial least squares* (PSL) e *principal components regression* (PCR). A PCR consiste na análise de componente principal no conjunto de dados, que reduz as variáveis preditoras às chamadas de componentes principais (PCs), que são uma combinação linear dos dados originais. Esses componentes principais são usados para construir o modelo de regressão linear. Através da $cross - validation$, é escolhido o número de PCs. A PCR é adequada quando o conjunto de dados contém preditores altamente correlacionados. Além da PCR, pode-se usar a *partial least squares* (PLS). Essa técnica também constrói um conjunto de combinações lineares das entradas para a regressão, mas, ao contrário da PCR, ela também usa y (a saída) para essa construção.

III. RESULTADOS

A. Pré-processamento

Após realizar o pré-processamento o conjunto de dados foi reduzido para $N = 20032$ observações tratando os elementos faltantes no conjunto de dados e obtendo dados transformados e divididos prontos para serem usados nas regressões.

Primeiramente é importante observar na matriz de correlação como mapa de calor na Figura 1 que não há correlação linear entre os pares de preditores e não há correlação linear entre os preditores e a saída. O maior valor de correlação encontrado entre pares de preditores foi entre os preditores *major discipline* e *education level* com correlação de 0,64. Outra característica importante para análise dos resultados é a caracterização, Tabela I, das entradas e da saída escolhidas, Tabela IV. As variáveis escolhidas para entrada são na grande maioria de origem de dados categóricos e a variável escolhida para saída, horas de treinamento, é originalmente numérica com média 65,30 e desvio padrão 60,09.

Tabela IV
ESTATÍSTICAS DO CONJUNTO DE DADOS PARA REGRESSÃO

Método	RMSE	R2
Regressão Linear Múltipla	60.72	-0.0006498
Regressão Linear Múltipla com Penalização	59.50	-0.0006648
PCR	60.00	0.0005148
PLS	60.03	-0.0003430

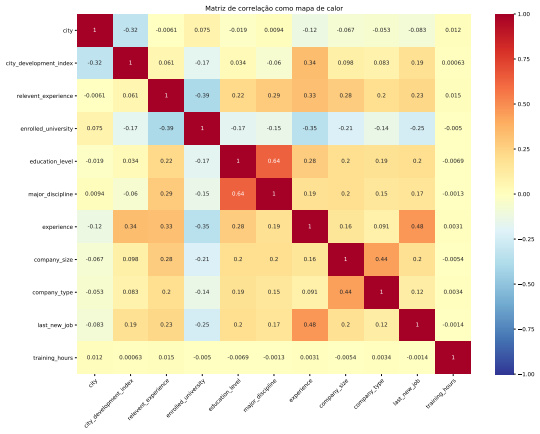

Figura 1. Matrizes de correlação cruzada dos preditores.

B. Regressão Linear Simples

O conjuntos de dados foi dividido de forma que 80% foi utilizado para treinar o modelo e 20% para teste. Com isso foi obtido uma $RMSE = 60,72$ e um $R^2 = -0.0006498$, Tabela IV.

O valor do coeficiente de determinação R^2 indica o quanto o modelo justifica as variações das saídas, ou seja, o quanto o nosso modelo está explicando os dados ou o quanto o modelo pode prever corretamente, logo o modelo não prevê as horas de treinamento dos candidatos corretamente.

Em seguida, foi utilizado uma validação cruzada com $10 - folds$ para uma nova estimação do modelo. Nesse caso, a menor RMSE foi de $56,85$, embora não seja uma diferença considerável esse modelo supera o anterior o que mostra a eficiência da validação cruzada.

Comparando os resultados de RMSE às características dos dados de saída, média e desvio padrão, percebemos que os valores estão muito próximos, outra evidência que o modelo não é eficiente para esse conjunto de dados.

C. Regressão linear com penalização

Os modelos de regressão com penalização servem para controlar coeficientes muito elevados que possam gerar instabilidade no modelo. Para estimar o parâmetro de restrição (λ) foi utilizada, novamente, a validação cruzada com $10 - folds$ que pode ser observada na Figura 2. Foi calculado os valores de RMSE e R2 para cada λ entre 0 e 1000 e observou-se pelos resultados que a diminuição da variância com o aumento de λ do modelo não melhora o modelo de forma significativa e não foi possível encontrar o valor ótimo para λ, pois o aumento de λ resultava no mesmo padrão de valores de RMSE e R2.

Comparando os resultados da performance do modelo de regressão linear com penalização e os resultados comentados anteriormente, não houve mudanças significativas em RMSE e R2, assim o modelo não prevê as horas de treinamento dos candidatos corretamente e os resultados de RMSE estão tão altos quanto o desvio padrão da saída evidenciando que o modelo de regressão linear com penalização também não é eficiente para esse conjunto de dados.

D. Modelo com redução de dimensionalidade

Para determinar a quantidade de componentes a serem utilizados foi analisado a curva de variância cumulativa em função dos números de PCs na Figura 3. A curva tem o ponto de 6 PCs correspondente a aproximadamente 80% da variância do banco de dados e escolhemos esse número de dimensões para o PCR e PLS.

Comparando os resultados da performance do modelo com redução de dimensionalidade, PCR e PLS, e os resultados comentados anteriormente, não houve mudanças significativas em RMSE e R2, assim o modelo não prevê as horas de treinamento dos candidatos corretamente e os resultados de RMSE estão tão altos quanto o desvio padrão da saída evidenciando que o modelo de regressão com redução de dimensionalidade, PCR e PLS, também não é eficiente para esse conjunto de dados.

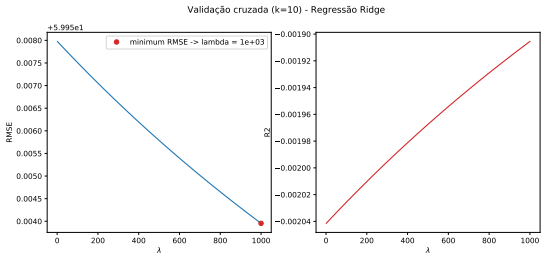

Figura 2. Validação cruzada na Regressão Ridge

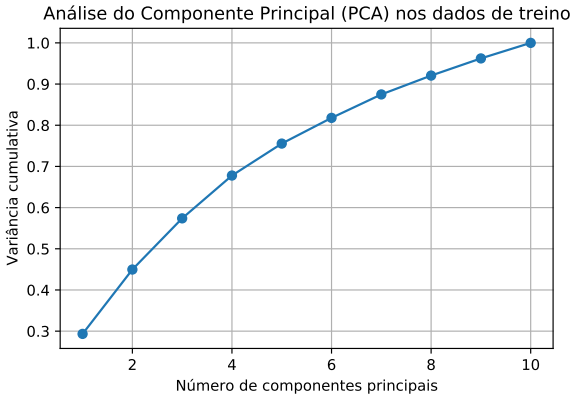

Figura 3. Análise da variância cumulativa nos dados de treino versus número de componentes principais

IV. Conclusão

O banco de dados escolhido possui baixa covariância entre seus dados, o que foi comprovado através dos diversos métodos de treinamento aplicados durante este trabalho. Na Tabela IV, por exemplo, percebe-se que todos os valores para R2 obtidos são próximos a zero, mas, para o modelo prever corretamente, o valor de R2 deve se aproximar de 1. Isso acontece porque os dados são dados de natureza de classificação, onde a maioria dos preditores são categóricos, e assim, não possuem dados contínuos ou grandezas que possa ser aplicada regressão. O impacto disso seria o *underfitting*, devido ao conjunto de dados ter alta variância. Conclui-se que é necessário estudar outros métodos de regressão que produzam modelos não lineares e que possam corresponder as características do nosso banco de dados, ou seja, métodos de regressão que atendam a banco de dados com saída com alta variância e sem correlação linear entre preditores e saída. Sugere-se estudos futuros de métodos de classificação para esse banco de dados, por possuir muitas variáveis categóricas e as outras características citadas anteriormente.

Referências

[1] J. A. N. S. Francisco Rodrigo P. Cavalcanti, *Fundamentos de gestão de projetos: gestão de riscos*. Atlas, 2016.

[2] J. N. Ram Charan, Stephen Drotter, *Pipeline de liderança: o desenvolvimento de líderes como diferencial competitivo*. Elsevier, 2012.

[3] G. J. D. W. Trevor Hastie, Robert Tibshirani, *An Introduction to Statistical Learning: With Applications in R*. Springer, 2013.

[4] Möbius. (2020, dec) Hr analytics: Job change of data scientists. [Online]. Available: https://www.kaggle.com/arashnic/hr-analytics-job-change-of-data-scientists

[5] T. C. S. João Pedro S. Campos, "Analise estatística no contexto de predição de mudança de emprego," 2021.