Modelos de Classificação para Predição de 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 classificação para predição da mudança de emprego de candidatos 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. Para o conjunto de dados escolhido os métodos não lineares apresentam resultados superiores com destaque ao método de classificação que utiliza a Máquina de Vetores de Suporte.**

***Palavras-chave*– classificação, aprendizado estatístico, análise estatística, pré-processamento, mudança de emprego.**

I. Introdução

Uma empresa pretende contratar cientistas de dados e deseja, através de análises estatísticas, criar um modelo preditivo para identificar, após o treinamento realizado na empresa, se um candidato que participou do treinamento aceitará um novo emprego e aceitará trabalhar para a nova empresa ou não aceitará o novo emprego e não continuará após o treinamento. 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 e assim reduzir impactos em custo e tempo, além de melhorar a qualidade dos treinamentos.

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. Porém quando tratamos de uma saída que atribui uma classe, o método de regressão é ineficaz, apesar de poder ser utilizado para classes binárias. Isso acontece porque em um problema de classificação os valores de saída são específicos e a saída de um modelo de regressão é numérica, assim quando usamos um método de regressão em um problema de classificação binária, por exemplo, as estimativas estarão entre 0 e 1 e também podem ter estimativas negativas e acima de 1, ou seja, valores que não correspondem às categorias.

Para esse problema serão comparados métodos de classificação, pois são utilizados para predição de variáveis qualitativas, também chamadas de categóricas. A classificação atribui uma classe ou categoria a uma observação de um conjunto de dados. Para isso o método calcula a probabilidade da observação ser de uma determinada classe, atribuindo àquela com maior probabilidade.

Nenhum método terá performance superior em todas as aplicações e portanto o melhor método de classificação 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].

Os métodos de classificação podem ser divididos em métodos lineares e não lineares. Dentre os métodos de classificação lineares, existe a Classificação Logística e a Análise de Discriminantes Lineares (LDA), além deste exite também a classificação utilizando redes neurais, a Análise de Discriminante Quadrático (QDA), a classificação K-Vizinhos mais Próximos (KNN) e Máquinas de Vetores de Suporte (SVM) que são exemplos de métodos de classificação não lineares.

Independente do método de classificação 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.

II. Metodologia

O conjunto dados utilizado para predição *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 11 preditores.

O conjunto de dados possui variáveis qualitativas, também conhecidas como categóricas, e a saída é binária como podemos ver na Figura 13 que representa a distribuição das classes. A classe 0.0 representa os candidatos a mudança de emprego que escolheram não mudar de emprego ou os candidatos que não escolheram exercer a nova função. Já a classe 1.0 representa os candidatos que aceitaram mudar

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

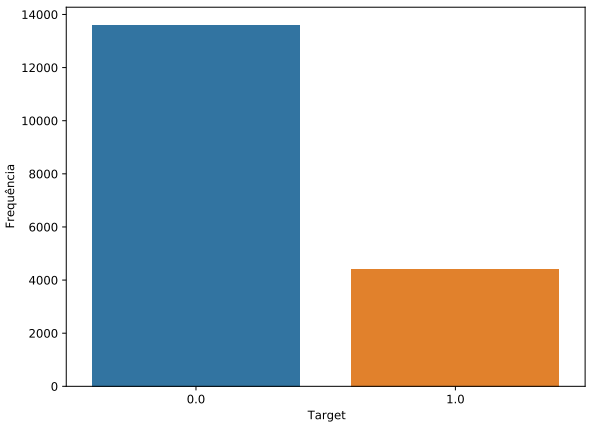

Figura 1. Distribuição das classes

de emprego ou escolheram continuar na empresa após o treinamento, caso do candidato desempregado ou primeiro emprego. Simplificadamente por essas características podemos chamar de um problema de classificação.

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.

A. Classificação linear

1) Regressão Logística: A Regressão Logística, também conhecida como Classificação Logística, é um método de classificação que utiliza a Equação 1 da regressão linear

$$Y = \beta_0 + \beta_1 X \tag{1}$$

e aplica na função logística na Equação 2, pois as saídas do modelo ficam entre 0 e 1 para todas os valores de X.

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}} \tag{2}$$

Os coeficientes β_0 e β_1 são desconhecidos e devem ser estimados com base no conjunto de treino.

A Classificação Logística é tipicamente utilizada para problemas de classificação entre duas classes e portanto pode ser utilizada nesse conjunto de dados.

2) Análise de Discriminantes Lineares: O classificador LDA resulta da suposição de que as observações dentro de cada classe vem de uma distribuição normal com uma média específica de classe e uma variância comum. Ou seja, $\mathcal{N}(\mu_k, \mathbf{\Sigma})$, onde μ_k é um vetor de média por classe, e $\mathbf{\Sigma}$ é a matriz de covariância que é comum a todas as K classes. O classificador Bayesiano atribui uma observação $X = x$ para a classe para o maior valor de $\delta_k(x)$, que é

$$\delta_k(x) = x^T \mathbf{\Sigma}^{-1} \mu_k - \frac{1}{2} \mu_k^T \mathbf{\Sigma}^{-1} \mu_k + \log \pi_k \tag{3}$$

B. Classificação não linear

1) Análise de Discriminantes Quadráticos: A análise discriminante quadrática (QDA) é uma abordagem alternativa ao LDA. Supõe-se que as observações de cada classe obedecem a uma distribuição gaussiana e estima-se os parâmetros através do teorema de Bayes. No entanto, ao contrário do LDA, o QDA assume que cada classe tem sua própria matriz de covariância Σ_k. Ou seja, $X \sim \mathcal{N}(\mu_k, \Sigma_k)$. Sob esta suposição, o classificador de Bayes atribui uma observação $X = x$ à classe para o maior valor de $\delta_k(x)$, dado por

$$\begin{aligned}\delta_k(x) &= -\frac{1}{2}(x-\mu_k)^T \mathbf{\Sigma}_k^{-1}(x-\mu_k) - \frac{1}{2}\log|\mathbf{\Sigma}_k| + \log \pi_k \\ &= -\frac{1}{2} x^T \mathbf{\Sigma}_k^{-1} x + x^T \mathbf{\Sigma}_k^{-1} \mu_k - \frac{1}{2} \mu_k^T \mathbf{\Sigma}_k^{-1} \mu_k \\ &\quad - \frac{1}{2}\log|\mathbf{\Sigma}_k| + \log \pi_k \end{aligned} \tag{4}$$

Portanto, esse classificador envolve inserir estimativas para $\mathbf{\Sigma}_k, \mu_k$ e π_k em (4.23), e então atribuir uma observação $X = x$ para a classe onde essa quantidade é maior.

A diferença entre os métodos LDA a QDA é o *trade-off* entre o bias e a variância. QDA estima uma matriz de covariância separada para cada classe, para um total de $Kp(p+1)/2$ parâmetros e o modelo LDA há apenas Kp coeficientes lineares a serem estimados. Conseqüentemente, o LDA é um classificador menos flexível do que o QDA

e, portanto, tem uma variância menor. Isso pode levar a um desempenho de previsão aprimorado, porém se a suposição da matriz de covariância comum do LDA for ruim, o LDA pode sofrer alta polarização.

Assim, o LDA tende a ser melhor do que o QDA se houver relativamente poucas observações de treinamento e, portanto, reduzir a variância é crucial. Em contraste, QDA é melhor se o conjunto de treinamento for muito grande, de modo que a variância do classificador não importa crucialmente.

2) K-Vizinhos mais Próximos: Apesar de ser uma abordagem muito simples, o classificador KNN pode frequentemente produzir modelos que são surpreendentemente próximos do classificador Bayes ideal. Em geral para K pequeno o modelo é mais flexível e corresponde a um classificador com baixo viés e alta variância. Já para K grande o modelo é menos flexível e corresponde a um classificador com baixa variância e alto viés. Com K unitário a taxa de erro no treino é nula, porém a taxa de erro no teste pode ser muito alta. Assim a taxa de erro no treino diminui com o aumento da flexibilidade do modelo, mas a taxa de erro de teste atinge um mínimo e depois aumenta. Portanto a escolha do K ótimo é essencial para a performance do modelo e podemos encontrá-lo a partir do ponto de mínimo da taxa de erro de teste.

3) Máquina de Vetores de Suporte: O método de classificação Máquina de Vetores de Suporte (SVM, sigla para expressão em inglês Support Vector Machines) é um método de classificação que é baseado no conceito de separação por hiperplano. Sabendo que um hiperplano é um subespaço que tem $p-1$ dimensões da dimensão dos dados (espaço de p-dimensões), o classificador SVM classifica utilizando-o para separação das classes. Essa separação é feita a partir da maior distância das observações para o hiperplano, assim a observação mais próxima é chamada de margem, pois é a menor distância para o hiperplano. Essa distância é calculada perpendicularmente ao hiperplano.

A utilização desse classificador é para maior robustez para observações individuais e melhor classificação da maioria das observações de treinamento, sendo um método versátil. Embora o classificador SVM geralmente seja bem sucedido ele também pode levar a um sobreajuste quando p é grande.

C. Avaliação do modelo de classificação

A abordagem mais comum para quantificar a performance do modelo é a taxa de acerto. Um bom classificador é àquele que a taxa de acerto do conjunto de teste é maior, ou seja, o erro de teste menor. Podemos representar a taxa de acerto de um modelo pela expressão:

$$\frac{1}{n}\sum_{i=1}^{n} I\left(y_i = \hat{y}_i\right) \tag{5}$$

Onde n é o número de observações, y_i é a variável de saída, dado que temos acesso às respostas por ser modelos supervisionados, e $\hat{y}_i$ é a estimativa do modelo para saída, ou seja, a classe predita. Assim será somado uma unidade quando a classe observada é igual a classe predita para uma determinada observação. Outro método de avaliação é utilizando a matriz de confusão na Tabela III, onde VN é o Verdadeiro Negativo que representa o número de predições corretas para a classe 0, FP é o Falso Positivo que representa o número de predições erradas para a classe 1, FN é o Falso Negativo que representa o número de predições erradas para a classe 0 e o VP é o Verdadeiro Positivo que representa o número de predições corretas para a classe 1. Utilizando os valores da matriz de confusão podemos compor as pontuações a partir do teste do modelo, são exemplos de pontuação de performance do modelo: Pontuação da Precisão, Taxa do Verdadeiro Positivo e Taxa de Acerto. A partir da Taxa do Verdadeiro Positivo e da Taxa do Falso Positivo podemos montar um gráfico comparativo chamado curva ROC em que o melhor modelo é aquele que se aproxima mais do valor unitário da área abaixo da curva.

Tabela III
MATRIZ DE CONFUSÃO

		Predito		
		Não	Sim	Total
Observado	Não	VN	FP	
	Sim	FN	VP	
	Total			N° observações

D. Comparação dos métodos de classificação

A partir do estudo preliminar dos métodos de classificação espera-se que o classificador LDA supere o classificador Regressão Logística quando houver distribuições normais da classe ou quando supomos distribuições normais da classe. Entretanto quando o limite de decisão é altamente não linear, os modelos KNN superam os modelos feitos a partir dos métodos LDA e Regressão Logística, pois no método KNN não é feita suposição quanto ao limite de decisão. A desvantagem do método KNN é a necessidade de que o número de observações seja muito grande e o número de preditores seja pequeno, assim quando N é muito maior que D o KNN tende a reduzir o viés enquanto incorre em muita variância. Caso contrário o método QDA, para limite de decisão não linear, pode ser escolhido no lugar do KNN, pois a suposição do limite de decisão o torna mais performante em conjunto de dados menores.

III. RESULTADOS

A. Pré-processamento

Após realizar o pré-processamento o conjunto de dados foi reduzido para $N = 18014$ observações tratando os elementos faltantes no conjunto de dados pronto para ser usado nas classificações. Também foi feito o balanceamento das classes da saída utilizando o método de subamostragem, assim a saída ficou com o número igual de observações para as duas classes, reduzindo consideravelmente o conjunto de dados. O balanceamento foi feito para trazer um modelo que não fosse viciado para uma das classes assim obteve-se uma nova distribuição da saída na Figura 2.

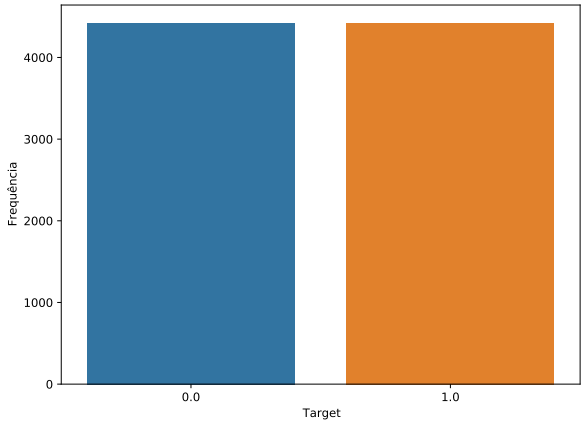

Figura 2. Distribuição das classes após subamostragem

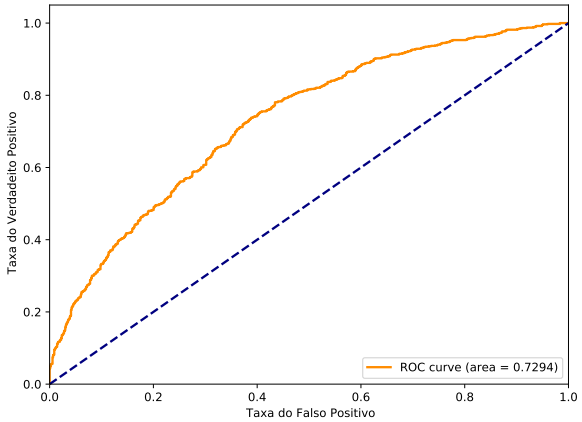

Figura 4. ROC regressão logística subamostragem

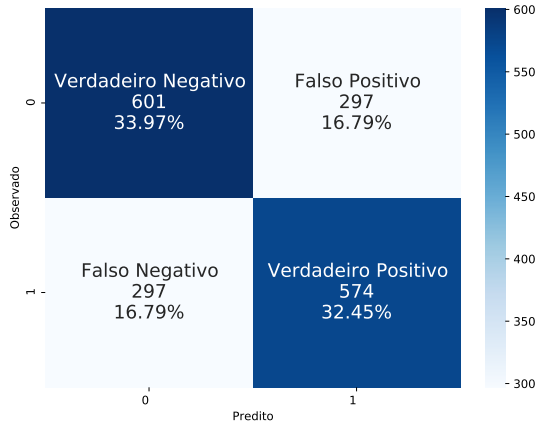

Figura 3. Matriz de confusão logística subamostragem

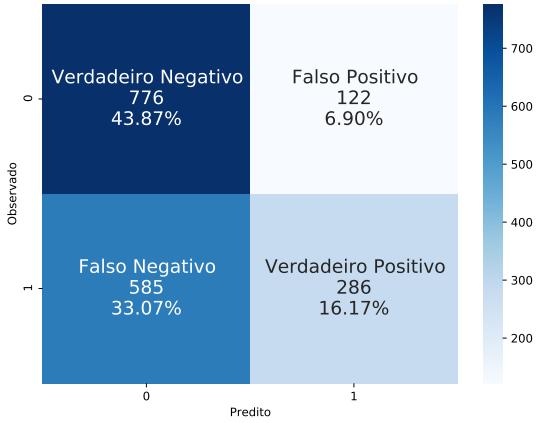

Figura 5. Matriz de confusão LDA subamostragem

B. Classificação linear

1) *Regressão Logística:*
2) *Análise de Discriminantes Lineares:*

C. Classificação não linear

1) *Análise de Discriminantes Quadráticos:*
2) *K-Vizinhos mais Próximos:*
3) *Classificador Máquina de Vetores de Suporte:*

D. Comparação dos modelos produzidos

Na Tabela IV tem as pontuações dos modelos produzidos no presente trabalho, onde P é a Pontuação da Precisão (Precision Score), TVP é a Taxa do Verdadeiro Positivo (Recall Score) e TA é a Taxa de Acerto (Accuracy Score).Os métodos lineares e os métodos que utilizam a suposição de uma distribuição normal tiveram pior resultado que pode ser justificado pela característica fronteira de decisão e distribuição dos dados.

Tabela IV
TABELA COMPARATIVA

	Regressão Logística	LDA	QDA	SVM	KNN
P	65.90	70.10	71.29	68.03	69.46
TVP	65.90	60.03	67.85	84.27	74.17
TA	66.42	32.84	70.72	72.75	71.23

IV. Conclusão

Os resultados são favoráveis a utilização de métodos não lineares como podemos ver na Tabela IV que mostra melhores pontuações para os modelos de classificação produzidos utilizando QDA, SVM e KNN.

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.

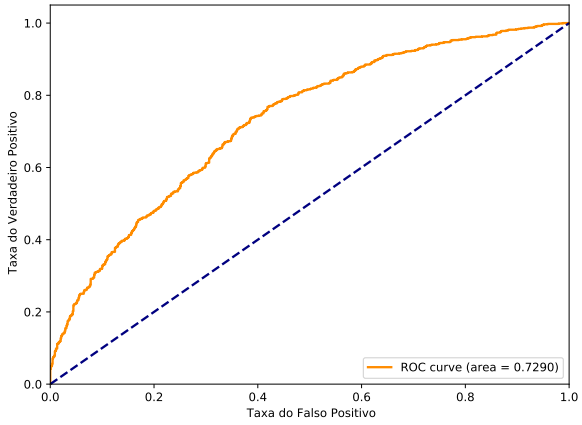

Figura 6. ROC LDA subamostragem

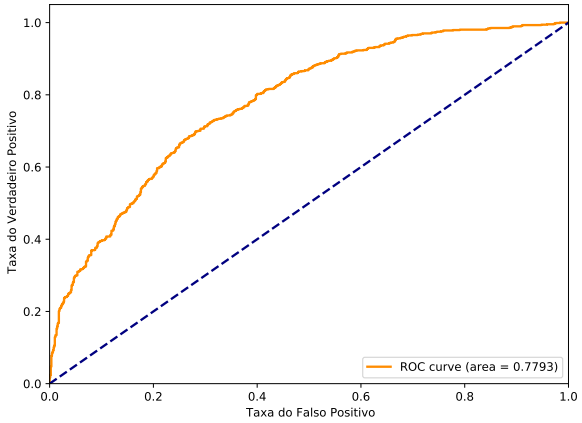

Figura 8. ROC QDA subamostragem

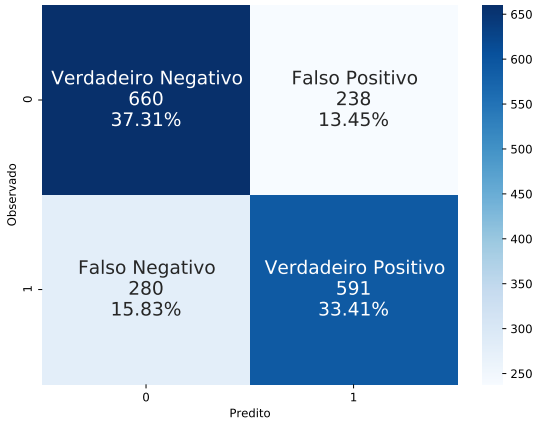

Figura 7. Matriz de confusão QDA subamostragem

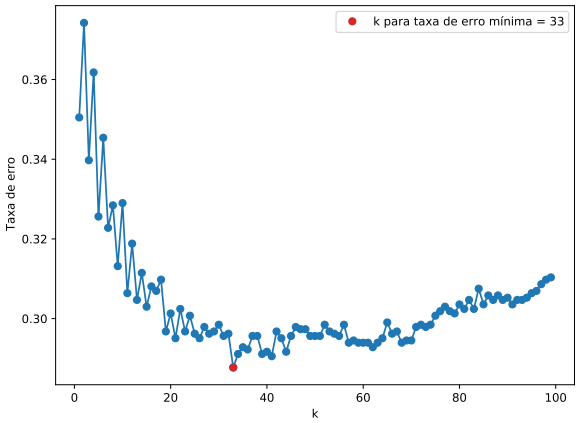

Figura 9. K ótimo KNN subamostragem

[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.

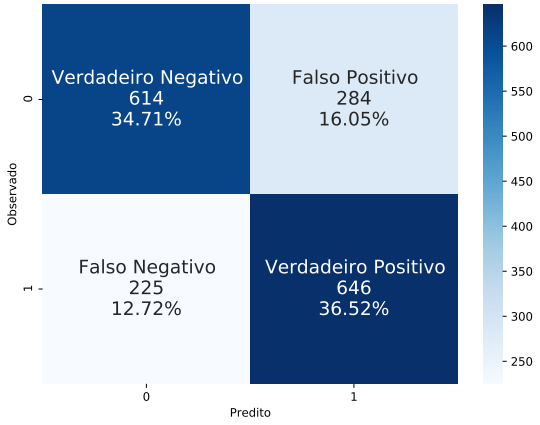

Figura 10. Matriz de confusão KNN subamostragem

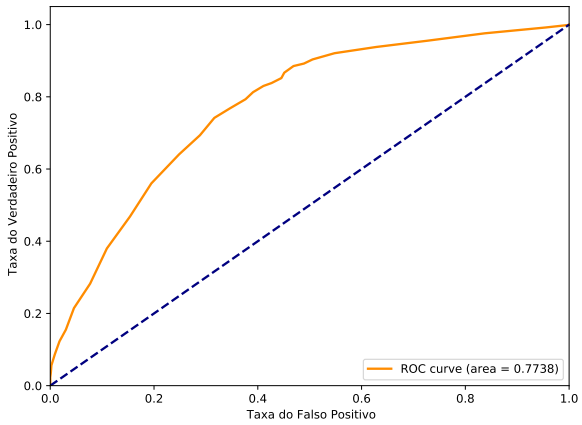

Figura 11. ROC KNN subamostragem

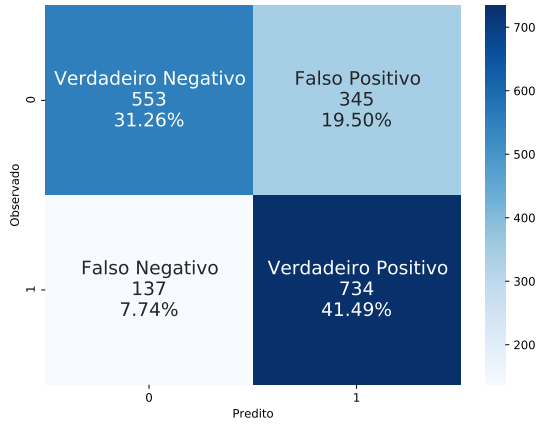

Figura 12. Matriz de confusão SVM subamostragem

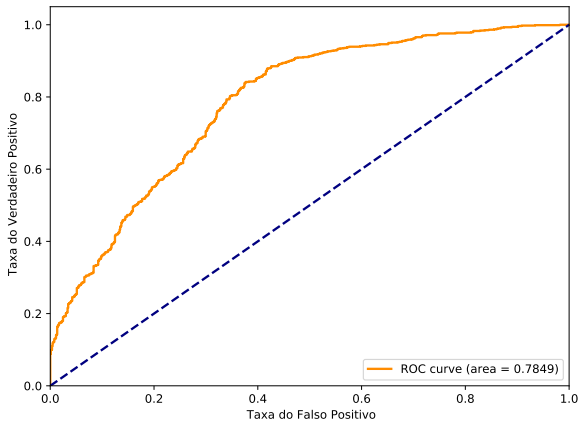

Figura 13. ROC SVM subamostragem