Desempenho do Support Vector Machine utilizando PCA

André G. Monteiro - Leonardo Aranha - Luara M. Marino - Renato K. Yamashiro

25 de abril de 2022

Resumo

Em alguns cenários, o cientista de dados se depara com grande quantidade de variáveis e a sua análise pode ser cansativa e demorada. O PCA pode ser uma ferramenta poderosa, aumentando a distância entre as variáveis e reduzindo a sua dimensionalidade através das técnicas de autovalores e autovetores da álgebra linear.

1 Introdução

Para o estudo, será utilizado o dataset "house prices", com 80 variáveis. Elas representam as características dos imóveis no momento onde é realizada a venda.

O objetivo final deste documento é a comparação de desempenho entre o modelo utilizando PCA (Principal Component Analysis) e outra não utilizando PCA.

2 Metodologia

2.1 PCA

A Análise de Componentes Principais (ACP) ou Principal Component Analysis (PCA) é uma técnica matemática que utiliza autovetores e autovalores, na qual se faz a mudança de base dos dados, o que resulta na redução da dimensionalidade (número de componentes principais) e distânciamento entre os vetores.

A redução da dimensionalidade se faz pela transformação dos dados para um novo sistema de coordenadas de forma que a maior variância por qualquer projeção dos dados fica ao longo da primeira coordenada (componente 0) e a segunda maior variância fica ao longo da segunda coordenada (componete 1), assim por diante.

2.2 SVM

A máquina de suporte (SVM) ou support-vector machine, é um conceito da ciência da computação para um conjunto de métodos de aprendizado supervisionado. Pode ser utilizado para regressão ou

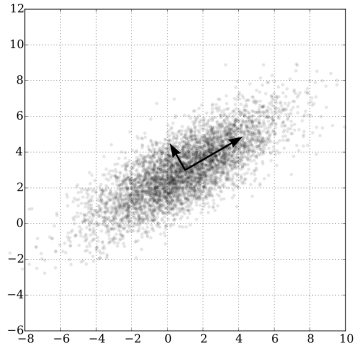

Figura 1: Exemplo de funcionamento do PCA.

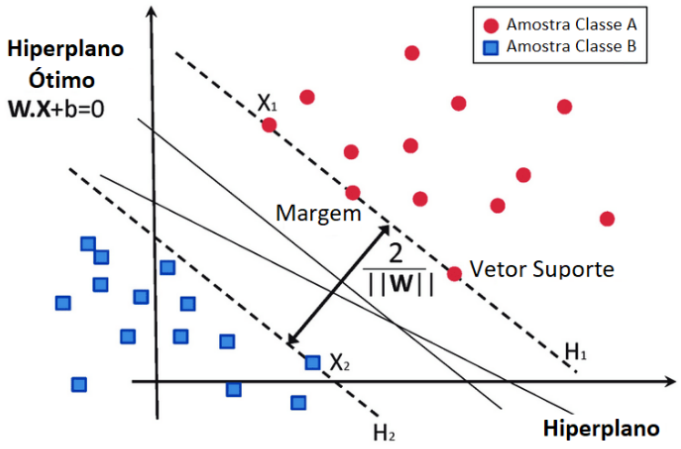

Figura 2: Exemplo de funcionamento de uma SVM.

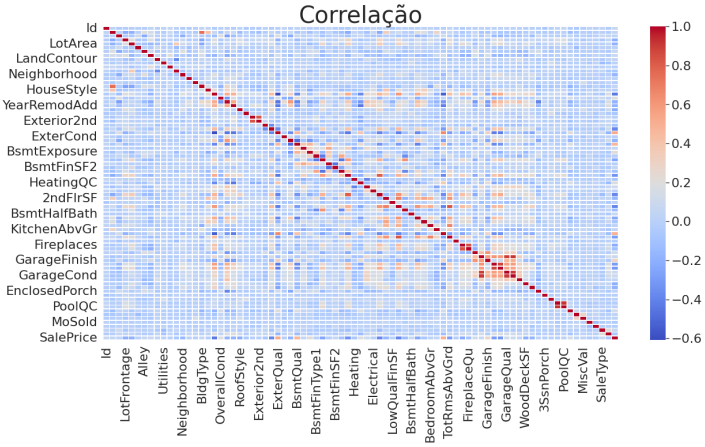

Figura 3: Matriz de correlação.

classificação. As variáveis são uma representadas no hiperplano, onde o SVM tenta encontrar uma linha de separação entre as classes, como podemos ver na Figura 2.

O dataset utilizado possui 80 variáveis, e para otimizar o desempenho da SVM, é necessário a análise e filtragem dos dados, o que demanda tempo para o cientista de dados. A Figura 3, mostra a matriz de correlação, e como podemos observar, a análise de cada variável pode ser custosa. Como o PCA diminui a correlação entre as variáveis e o método de SVM utiliza hiperplano e as distância entre os labels para a classificação, será comparada o desempenho entre os modelos como descrito anteriormente.

3 Resultados

3.1 Modelo SVM sem PCA

O modelo seguinte possui o seguinte pipeline:

1. Smote para solucionar o desbalanceamento de dados;
2. Standard Scaler para a normalização dos dados;
3. Support Vector Machine para múltiplas classes.

	precision	recall	f1-score	support
High Price	1.00	0.82	0.90	38
Low Price	0.66	0.61	0.63	41
Mid Price	0.92	0.95	0.93	266
accuracy			0.90	345
macro avg	0.86	0.79	0.82	345
weighted avg	0.90	0.90	0.89	345

Figura 4: Metrica para SVM sem PCA.

	precision	recall	f1-score	support
High Price	0.94	0.89	0.92	38
Low Price	0.60	0.78	0.68	41
Mid Price	0.95	0.91	0.93	266
accuracy			0.90	345
macro avg	0.83	0.86	0.84	345
weighted avg	0.91	0.90	0.90	345

Figura 5: Metrica para SVM com PCA.

O modelo teve a acurácia de 90 porcento, onde as métricas de precisão, recall e f1-score foi favorável para as casas de preços altos e preços médios e não favorável para preços baixos, como observado na Figura 4.

3.2 Modelo SVM com PCA

O modelo seguinte possui o seguinte pipeline:

1. Smote para solucionar o desbalanceamento de dados;
2. Standard Scaler para a normalização dos dados;
3. PCA com 20 componentes principais;
4. Support Vector Machine para múltiplas classes.

Utilizando PCA, o modelo obteve uma leve otimização nas métricas recall e f1-score para Low Price como mostra a Figura 5.

4 Conclusão

Utilizando 20 componentes principais, a sua representatividade foi equivalente a 67 porcento dos dados totais, apesar de sacrificar as métricas para as casas de preços altos, o resutado final foi próximo ao do modelo sem PCA, porém em termos de dimensionalidade, reduzimos para 25 porcento dos dados totais.

No entanto, ao observarmos a Figura 7, é claro que o modelo ainda teve dificuldade na diferenciação entre as casas de preço baixo e médio, já que os respectivos pontos são próximos entre si, o que pode indicar que as características entre essas duas classes são próximas.

O modelo teve também o overfitting pois ao prever os preços dos dados novos, foi classificado para todas as linhas como casas de preço médio como mostra a Figura 8.

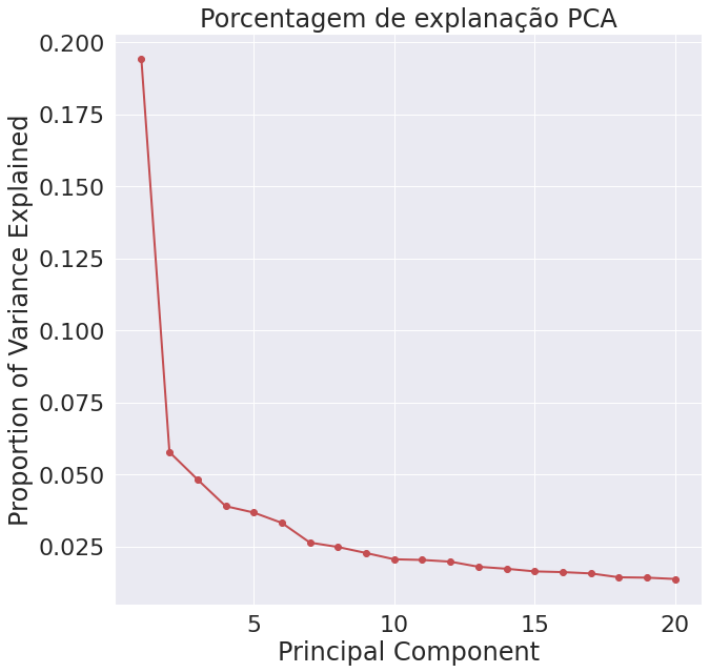

Figura 6: Explicabilidade por cada componente.

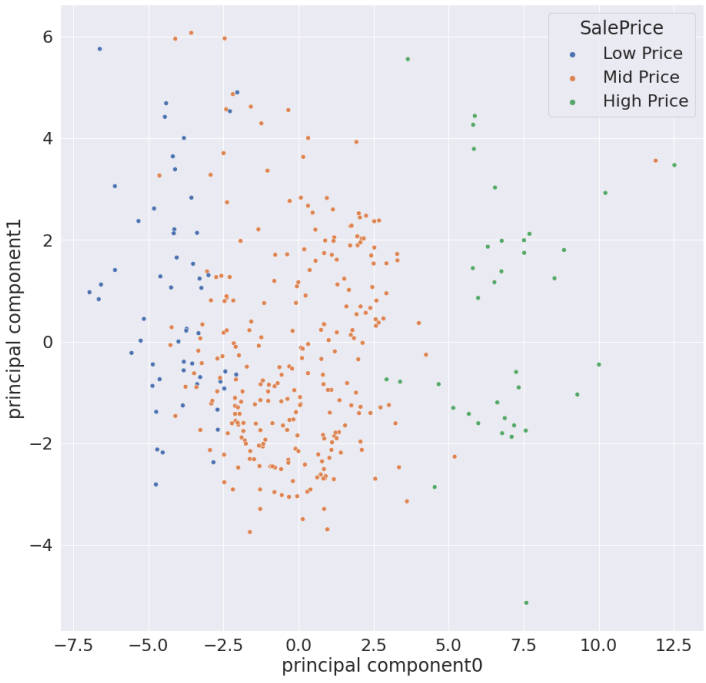

Figura 7: Scatter das principal component0 e principal component1.

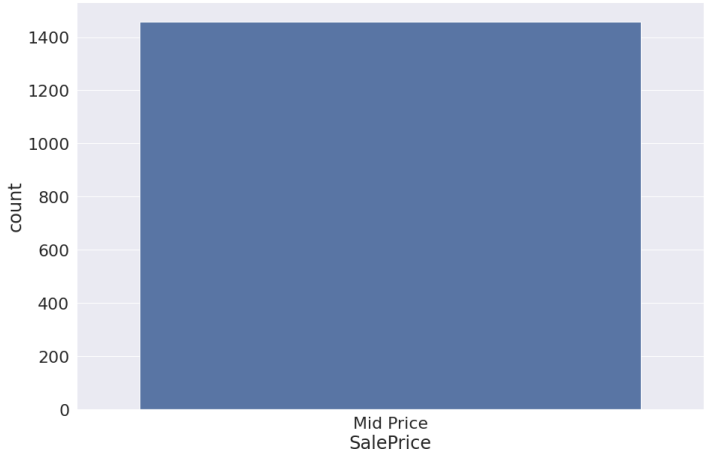

Figura 8: Gráfico de distribuição dos dados novos preditos.

Referências

[1] Mehdi Naseriparsa and Mohammad Mansour Riahi Kashani (2013) *Combination of PCA with SMOTE Resampling to Boost the Prediction Rate in Lung Cancer Dataset*, Islamic Azad University.

[2] Ian T. Jolliffe and Jorge Cadima (2016) *Principal component analysis: a review and recent developments*, Royal Society.

[3] Andy Martin del Campo (2019) *Using PCA in a Machine Learning Pipeline*, Medium.