CPQD AutoML Algorithm

11 de Março de 2022

Conteúdo

Lista de Figuras

Lista de Tabelas

1 Análise de Outliers

Os outliers são dados que se diferenciam drasticamente de todos os outros. Em outras palavras, um outlier é um valor que foge da normalidade e que pode (e provavelmente irá) causar anomalias nos resultados obtidos por meio de algoritmos e sistemas de análise. Entender os outliers é fundamental em uma análise de dados por pelo menos dois aspectos: os outliers podem viesar negativamente todo o resultado de uma análise; o comportamento dos outliers pode ser justamente o que está sendo procurado.

1.1 DBscan

Isolation Forests é um modelo de detecção de anomalias que faz uso de um conjunto de dados onde o alvo, neste caso a anomalia da fraude, do modelo contém poucas amostras entre tantos dados normais. A ideia do modelo é construir Árvores para isolar essas anomalias. Em outras palavras, a floresta de Isolamento é um conjunto de Árvores de Isolamento. Método parecido com a da nossa querida Random Forest.

1.2 Isolation Forest

Isolation Forests é um modelo de detecção de anomalias que faz uso de um conjunto de dados onde o alvo, neste caso a anomalia da fraude, do modelo contém poucas amostras entre tantos dados normais. A ideia do modelo é construir Árvores para isolar essas anomalias. Em outras palavras, a floresta de Isolamento é um conjunto de Árvores de Isolamento. Método parecido com a da nossa querida Random Forest.
Utilizando o método Isolation Forest foram detectados 8 outliers neste dataset, correspondendo a uma proporção de 5.33% do conjunto de amostras.
A orbservação do outliers pode ser feita nas tabelas 1-2, onde serão mostradas as tabelas de descrição dos outliers, tando de todas as features categóricas quanto de todas as features numéricas. Também é mostrado, na figura 1, a distribuição dos outliers em relação ao restante da população do dataset.

Tabela 1: Descrição Features Numéricas dos Outliers: Isolation Forest

	sepal.length	sepal.width	petal.length	petal.width
count	8.00	8.00	8.00	8.00
mean	6.35	3.44	3.90	1.25
std	1.47	0.73	2.82	1.08
min	4.30	2.30	1.10	0.10
25%	5.40	2.90	1.27	0.28
50%	6.50	3.70	3.80	1.20
75%	7.70	3.85	6.48	2.23
max	7.90	4.40	6.90	2.50

Tabela 2: Descrição Features Categóricas dos Outliers: Isolation Forest

	variety
count	8
unique	2
top	Setosa
freq	4

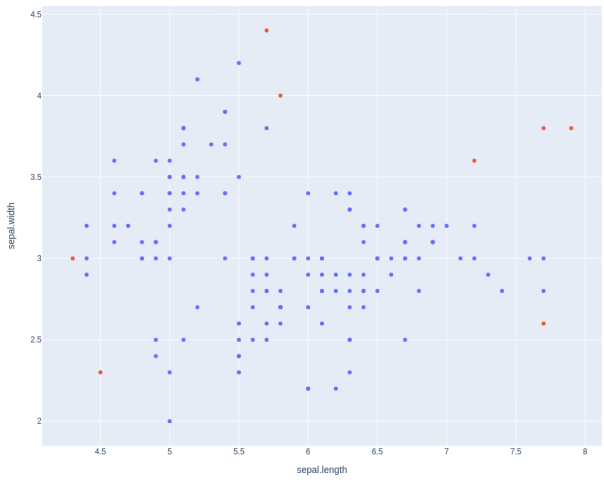

Figura 1: Visualização dos outliers: Isolation Forest

2 Análise de Cluster

O clustering ou análise de agrupamento de dados é o conjunto de técnicas de prospecção de dados (data mining) que visa fazer agrupamentos automáticos de dados segundo o seu grau de semelhança. O critério de semelhança faz parte da definição do problema e, dependendo, do algoritmo. A cada conjunto de dados resultante do processo dá-se o nome de grupo, aglomerado ou agrupamento (cluster). O procedimento de agrupamento (clustering) também pode ser aplicado a bases de texto utilizando algoritmos de prospeção de texto (text mining), onde o algoritmo procura agrupar textos que falem sobre o mesmo assunto e separar textos de conteúdo diferentes.

2.1 Feature Permutation 2 dim.

A melhor separação de grupos ocorreu nas features: ['petal.length', 'petal.width'] com uma quantidade de 3 grupos. A análise multidimensional em 2 dimensões fornece insights de quais features são mais importantes e que distinguem grupos entre si. Esta análise é realizada realizando a permutação das features do dataset, procurando os agrupamentos que melhor distinguem as amostras. Este tipo de operação traz diversos insights sobre quais features são mais importantes num ponto de vista de agrupamento de amostras.
A matriz de separação Tab. 4 mostra a distribuição populacional normalizada dentro de cada grupo de cluster. Note que valores maiores significam que há maior presença daquela população dentro daquele grupo. Este tipo de tabela também pode ser chamado de matriz de confusão, do inglês Confusion Matrix.

Tabela 3: Matriz de separação do melhor agrupamento

	0	1	2
Versicolor	0.94	0.00	0.06
Virginica	0.02	0.00	0.98
Setosa	0.00	1.00	0.00

2.2 Multidimensional

A melhor separação de grupos ocorreu nas features: ['sepal.length', 'sepal.width', 'petal.length', 'petal.width'] com uma quantidade de 3 grupos. A análise multidimensional de clusters é semelhante a análise anterior, mas utilizando todas as features disponíveis no dataset. Esta análise tem o potencial da melhor separação entre grupos, se esta existir. Existem casos em que o score de separação é menor em relação às análises bi-dimensionais ou tri-dimensionais, significando que provavelmente aquela combinação é a mais determinante para a separaçao de grupos.
A matriz de separação Tab. 4 mostra a distribuição populacional normalizada dentro de cada grupo de cluster. Note que valores maiores significam que há maior presença daquela população dentro daquele grupo. Este tipo de tabela também pode ser chamado de matriz de confusão, do inglês Confusion Matrix.

2.3 Comparação dos Métodos

Na tabela 5 abaixo podemos observar a comparação de todos os métodos de clustering testados. O método com o maior score será considerado o melhor método, sendo este utilizado nas próximas análises.

Tabela 4: Matriz de separação do melhor agrupamento

	0	1	2
Versicolor	0.10	0.00	0.90
Virginica	0.80	0.00	0.20
Setosa	0.00	1.00	0.00

Tabela 5: Análise de score do agrupamento.

	Feature Permutation 2 dim.	Multidimensional
petal.length - petal.width	0.66	NaN
sepal.width - petal.length	0.59	NaN
sepal.length - petal.length	0.58	NaN
sepal.length - sepal.width - petal.length - pet...	NaN	0.53

2.4 Insights - Variáveis Numéricas

Insights obtidos das variáveis numéricas estão disponíveis nas tabelas 6-7, onde são apresentadas as diferenças das médias e variâncias entre a população geral e cada um dos grupos. A idéia é facilitar a observação de tendências distintas em cada um dos grupos, em relação a população geral. A tabela de variância é importante de um ponto de vista de análise da variação das features dentro de cada um dos grupos. A ideia é que a variância dentro de um grupo específico seja menor em relação a população em geral.
Figuras também são aliadas importantes na visualização de dados. Nas figuras 2-3 estão presentes duas figuras que representam a variação de média e variância de cada grupo em relação a população geral. Dados variando para a cor azul significam que a variação é negativa, enquanto dados variando para cores vermelhas significam que a variação é positiva.

Tabela 6: Diferença de Média entre População - Grupos

cluster_group	0	1	2
sepal.length	0.87	-0.81	0.01
sepal.width	-0.04	0.39	-0.30
petal.length	1.76	-2.27	0.56
petal.width	0.79	-0.95	0.20

A maior diferença populacional positiva foi detectada na feature petal.length e no grupo 0, com valor de 1.76. A maior variação negativa foi na feature petal.length e no grupo 1, com o valor registrado de -2.27

2.5 Insights - Variáveis Categóricas

É possível observar nas tabelas 8-9 alguns dados notáveis extraídos dos grupos da base de dados. Estas tabelas de diferenças máximas e mínimas visam demonstrar as diferenças de distribuição entre a população e os grupos.

Tabela 7: Diferença de Variância entre População - Grupos

cluster_group	0	1	2
sepal.length	-0.33	-0.47	-0.34
sepal.width	-0.18	-0.10	-0.10
petal.length	-1.22	-1.53	-1.20
petal.width	-0.42	-0.64	-0.47

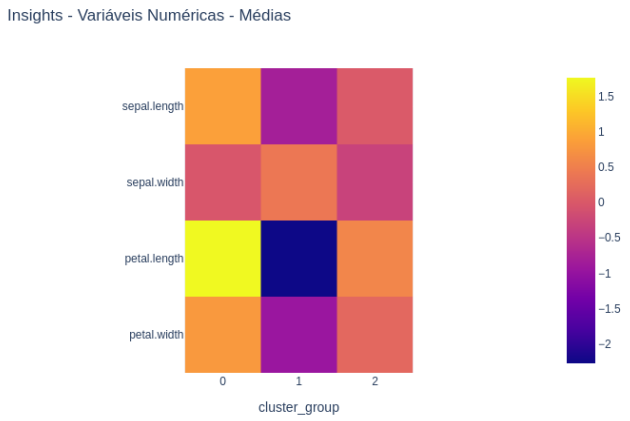

Figura 2: Diferença de Média entre População - Grupos

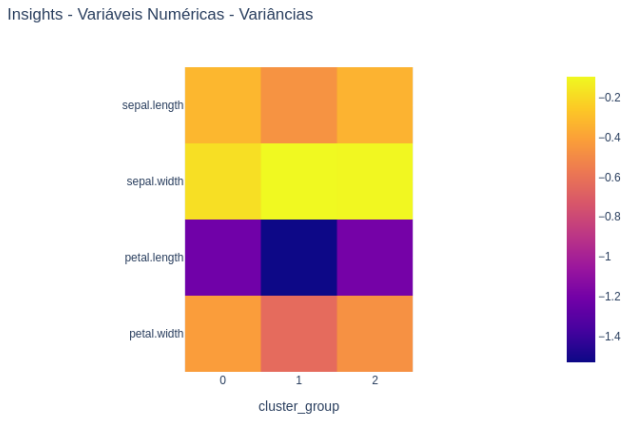

Figura 3: Diferença de Variância entre População - Grupos

A linha *Contagem* mostra a contagem daquela classe dentro do grupo, A linha *Proporção* mostra a proporção da classe em relação a população do grupo, A linha *Diferença da População* mostra a diferença de proporção daquela população no grupo em relação a população geral.

Tabela 8: Diferença Máxima de População entre Grupos e Dataset

	variety
maior ocorrencia	Setosa
contagem	46
proporção	1.0
diferença da população	0.676056
grupo	1

Tabela 9: Diferença Mínima de População entre Grupos e Dataset

	variety
maior ocorrencia	Versicolor
contagem	45
proporção	0.833333
diferença da população	0.481221
grupo	2