A mineração de dados tem aplicações diversas, geralmente se visando extrair conhecimento de conjuntos de dados que a princípio não tem uma regra bem definida que os descreva, buscando se resolver uma tarefa de classificação ou regressão.

Neste contexto, foram utilizadas técnicas de mineração de dados para resolver um problema de regressão, que se tratava de inferir o valor de diamantes por meio de um conjunto de nove atributos: carat, cut, color, clarity, depth, table, x, y, e z, onde se utilizando destes atributos físicos da pedra era desejado inferir um preço.

Tal técnica pode ser utilizada para precificar de forma automática pedras preciosas e por meio disso buscar erros ou possíveis fraudes.

Foi utilizado durante todo o projeto os conceitos e técnicas discutidas em aula, desde os métodos de se tratar os dados, os modelos utilizados, as práticas abordadas.

Trabalhos Relacionados

Foram utilizados três trabalhos relacionados que tratavam deste problema, o Diamond Prices Prediction with 99% accuracy, feito pelo usuário PREETI MADAN, Random Forest Diamond Price Prediction, feito pelo usuário ABDU0CH, e DiramondPrices\_RegressionModels, feito pelo usuário MAHYAR ARANI, as principais contribuições destes trabalhos vem de auxílios de sintaxe da linguagem, onde existe operações que são facilitadas por funções e comandos.

A contribuição mais marcante veio do trabalho Diamond Prices Prediction with 99% accuracy, com uma estratégia de aplicar uma função invertível no atributo alvo visando diminuir a variação dos dados, desta forma auxiliando o modelo a funcionar melhor para todos os conjuntos de valores.

Desenvolvimento

Falando primeiramente dos dados, temos 53,940 instancias no banco de dados, onde todas ele tem os seguintes atributos, o quilate, uma medida continua do peso da pedra, o corte, um valor categórico nominal da qualidade da lapidação da pedra, a cor, outra variável categórico nominal que representa qual a cor da pedra, a profundidade, um valor continuo que representa a altura da pedra, a “mesa”, uma medida continua do diâmetro da pedra, o preço, representado como um valor continuo e as dimensões x, y e z da pedra, todas elas continuas.

Na etapa de tratamento dos dados, não foi encontrado nenhuma instancia que tenha um resultado nulo para algum atributo, foram encontrados 149 valores repetidos, porem como a base da dados é bem extensa isso não é um problema, foi realizada uma busca por Outliers, e foi contatado a existência de alguns, desta forma foi utilizado o cálculo do Z Score para remover esses dados que não se enquadravam no esperado, como a base de dados é extensa não há problemas de remover os dados.

Os atributos categóricos foram tratados se utilizando do one hot encoder, foi realizada uma discretização dos dados contínuos, ocorreram testes sem se utilizar tal técnica, mas seu uso se fez valido.

Durante a analise da relação entre o preço e os demais atributos os mais significativos foram o quilate, e as dimensões x, y e z. Foi realizada uma busca por multicolinearidade, onde caso fosse encontrado um resultado significativo entre duas variáveis utilizar uma para inferir a outra seria de grande valor.

Para se analisar os resultados obtidos foram utilizados o erro quadrado médio, a raiz do erro quadrado médio e o R Square, onde as duas primeiras medidas se baseiam em um somatório de erros e a segunda em uma diferença entre duas curvas.

Como técnica para melhora de resultados foi realizada a operação de Log na base 2 no valor do preço, em todos os dados entrada, de forma geral manipular os dados de teste pode gerar resultados errados, porém, como a função de Log é invertível podemos a qualquer momento obter os dados originais.

Foram utilizados 4 modelos, um modelo de regressão linear que atribui para cada atributo um coeficiente e a partir de um conjunto de atributos infere um resultado, o modelo Ridge que realiza uma manobra semelhante a regressão linear, porem somando alguns pesos ao resultado para tentar encontrar uma melhor inferência, o modelo Elasticnet que é um método hibrido entre o método Ridge e o método Lassa, de forma geral um modelo Elasticnet possui um conjunto maior de pesos a serem regulados do que um modelo Ridge e por último temos uma floresta aleatória, que utiliza uma serie de alvores de decisão para inferir um resultado.

Resultados

Por meio de uma sequencia de testes que estão apresentados anteriormente, foram encontradas as melhores operações a serem realizadas para ganhar eficiência dos modelos, sendo tal sequência, a remoção dos Outliers e utilização de todos os atributos na inferência, a discretização, a normalização piorou os resultados, a aplicação de Log no preço dos dados de treino e aplicação da exponencial nos dados previstos, se obtendo o melhor resultado sem se manipular os dados de teste, e aplicando Log nos dados de teste também chegamos ao melhor resultado obtido.

Foi possível por meio das diversas técnicas aplicadas se encontrar alguns modelos com mais de 95% de acurácia, sendo assim viável se utilizar um sistema automatizado para prever o valor de um diamante de se tendo como entradas as características físicas da pedra, por meio de modelos de baixo custo computacional.