Relatório Trabalho 1: Classificação

Gabriela dos Santos e Santos[1]

415 Palavras

DRE:118044310

March 5, 2021

1 Introdução

O presente trabalho consiste em um modelo de classificação que visa dar apoio a aprovação de crédito. O modelo utilizado foi o de Random Forest Classifier, e as bibliotecas utilizadas para pré-processamento, processamento e visualização dos dados foram Pandas e Seaborn. O script do código foi escrito na linguagem Python utilizando a IDE Jupyter Notebook.

2 Métodos

2.1 Pré -Processamento

O pré-processamento dos dados se deu na importação dos mesmos de arquivo CSV para DataFrame com o auxilio do pandas. Em seguida, foi retirado do DataFrame colunas com dados inconsistentes como se o usuário possuia telefone no trabalho, o código de área desse telefone e algumas colunas com dados repetidos.

2.2 Processamento

Algumas colunas ditas como relevantes no pré-processamento possuiam valores não definidos. Logo, foi necessário a substituição desses dados não definidos por valores genéricos de fácil leitura para o modelo. Para evitar propagações altas de erros foi utilizado substituição desses valores por valores de mediana da coluna. Todo esse processamento foi realizado nos dados de teste e nos dados de treinamento utilizando funções do Pandas.

2.3 Visualização

Feito o processamento dos dados foi necessário a plotagem do mesmo para saber como funciona a sua dispersão. Para isso foi utilizada a função pairplot do Seaborn. A função pairplot plota gráficos de dispersão utilizando de duas em duas dimensões, o que forma uma matriz de gráficos. Foi utilizada para essa matriz as colunas "tipo residencia", "qtde dependentes", "renda mensal regular", "ocupacao" e "inadimplente", sendo "inadimplente" o resultado da dispersão.

2.4 Random Forest Classifier

O modelo Random Forest Classifier é uma mistura de alguns modelos de predição, ele consiste basicamente na criação de diversas árvores de decisão que vão se ajustando de acordo

com a entrada de dados. Na implementação do algoritmo, o método foi importado como uma funnção do sklearn. Logo em seguida foi passado como parametros os dados de treinamento e teste e foi gerada a predição do modelo.

3 Resultados

O modelo teve uma predição de aproximadamente 50,4 por cento o que pode ser considerado razoável.

4 Discussões

Antes da utilização do Random Forest Clasifier foi utilizado o modelo de Support Vector Machine (SVM), que consiste basicamente em dividir em partes o gradiente de dados e de acordo com a margem da divisão classificar os dados. O modelo apresentou uma predição de aproximadamente 48 por cento, que não foi considerada tão satisfatória em relação ao Random Forest. O Random Forest apresentou aproximadamente 55 por cento de predição correta na competição do Kaggle e com isso conlui-se que o seu desempenho é melhor que o do SVM nesse banco de dados.