Relátorio: Trabalho 1 - Introdução a aprendizado de máquina 2020-2

1. Objetivo
   O objetivo deste trabalho é o desenvolvimento de um classificador que auxilie na decisão de aprovação de crédito

2. Os dados

   O conjunto de dados utilizado apresenta informações sobre pessoas que passaram pelo sistema de análise de crédito de uma instituição financeira.
   Foram fornecidos dois conjuntos de dados, um para treino e outro para testes.

3. Tratamento de dados

Colunas:

'id_solicitante', 'produto_solicitado', 'dia_vencimento',
'forma_envio_solicitacao', 'tipo_endereco', 'sexo', 'idade',
'estado_civil', 'qtde_dependentes', 'grau_instrucao', 'nacionalidade',
'estado_onde_nasceu', 'estado_onde_reside',
'possui_telefone_residencial', 'codigo_area_telefone_residencial',
'tipo_residencia', 'meses_na_residencia', 'possui_telefone_celular',
'possui_email', 'renda_mensal_regular', 'renda_extra',
'possui_cartao_visa', 'possui_cartao_mastercard',
'possui_cartao_diners', 'possui_cartao_amex', 'possui_outros_cartoes',
'qtde_contas_bancarias', 'qtde_contas_bancarias_especiais',
'valor_patrimonio_pessoal', 'possui_carro',
'vinculo_formal_com_empresa', 'estado_onde_trabalha',
'possui_telefone_trabalho', 'codigo_area_telefone_trabalho',
'meses_no_trabalho', 'profissao', 'ocupacao', 'profissao_companheiro',
'grau_instrucao_companheiro', 'local_onde_reside',
'local_onde_trabalha', 'inadimplente'.

Inicialmente, o conjunto de dados foi analisado usando a ferramenta “profile_report()” da biblioteca pandas, com isso foi possível descobrir várias informações importantes a respeito do conjunto de dados.

- Por apresentar valores nulos, alta cardinalidade ou por ser de baixa utilidade as seguintes variáveis foram retiradas do “dataframe” inicial, utilizando o método “drop”:

"id_solicitante", "grau_instrucao", "codigo_area_telefone_trabalho", "estado_onde_trabalha", "codigo_area_telefone_residencial", "estado_onde_reside", "estado_onde_nasceu", "local_onde_trabalha", "possui_telefone_celular", ‘'grau_instrucao_companheiro', 'profissao_companheiro'.

- As seguintes variáveis foram codificadas como binárias com o método “binarizer”:

"possui_telefone_residencial", "vinculo_formal_com_empresa". "possui_telefone_trabalho", "tipo_endereco".

- Nas seguintes variáveis categóricas foi aplicado o “one hot encoding” com o método “get_dummies” do pandas:

"forma_envio_solicitacao", "sexo".

- Ao utilizar o método "simple imputer" para realizar o preenchimento de colunas com células vazias com o valor da moda da variável, observou-se, posteriormente, uma queda na acuŕacia do classificar, por isso, utilizou-se o método “dropna” para retirar as amostras com valores vazios.

- Foi aplicada uma escala nas variáveis do conjunto utilizando o métod “MinMasScaler

4. Modelos

- Com o uso da ferramenta “GridSearchCV” do scikit learn foram testados no conjunto de treinamento os modelos Random Forest, Classificador KNN e Regressão logística, assim como várias combinações de parâmetros. Desta forma, os melhores resultados foram o seguintes:

```
   Model              Best score
--------------------------------------
        random_forest 0.58941
kneighbors_classifier 0.56918
  logistic_regression 0.57766
```

Com isso, o modelo random forest foi escolhido como definitivo para o projeto com os seguintes parâmetros:

```
classificador = RandomForestClassifier(
        n_estimators=300,
        random_state=12345,
        max_depth=6,
        min_samples_split=2,
        min_samples_leaf=1,
        min_weight_fraction_leaf=0.0
)
```

Obtendo a acurácia final de 0,57320.