INTELIGÊNCIA ARTIFICIAL

Credit Risk Analysis
Supervised Learning

Grupo 39

Mariana Ramos – up201806869
Pedro Ferreira – up201806506
Pedro Ponte – up201809694

Especificação do Projeto

O objetivo deste trabalho é utilizar *Supervised Learning* para prever o grau de risco de um empréstimo (*loan grade*). Para isso, contamos com um dataset de 855969 amostras de empréstimo. Estas amostras contêm os 73 atributos, sendo alguns dos mais importantes:

- **loan_amnt** – Quantidade de dinheiro solicitada pelo cliente;
- **int_rate** – Taxa de juros do empréstimo;
- **grade** – Grau do empréstimo, com valores A, B, C, D, E, F, G. Quanto mais próximo de A for o grau do empréstimo, menor será o risco deste não ser pago;
- **annual_inc** – Rendimento anual do cliente;
- **purpose** – Motivo principal para o pedido de empréstimo;
- **installment** – Valor pago por mês pelo empréstimo;
- **term** – tempo até o empréstimo estar pago.

Tirando partido dos diferentes valores destes atributos, utilizaremos vários classificadores para avaliar o possível grau de cada empréstimo, com uma taxa de acerto aceitável.

Referências e Trabalho Relacionado

- https://www.kaggle.com/rameshmehta/credit-risk-analysis
- https://rstudio-pubs-static.s3.amazonaws.com/190551_15f6124632824534b7e397ce7ad2f2b8.html
- https://rstudio-pubs-static.s3.amazonaws.com/263968_5057ec1f5a2e48a89aab7f568fc37ade.html
- Slides das aulas teóricas e fichas realizadas nas aulas teórico-práticas
- https://pandas.pydata.org/, https://scikit-learn.org/stable/, https://numpy.org/, https://matplotlib.org/, https://keras.io/

Detalhes da implementação

Ferramentas

Para o desenvolvimento deste projeto iremos usar **Python** e as suas bibliotecas que tornam mais simples o desenvolvimento de projetos de aprendizagem supervisionada, cujos algoritmos associados são normalmente bastante complexos.

- **Scikit-learn** – extensa biblioteca com um enorme leque de algoritmos úteis à aprendizagem computacional;
- **Keras -** biblioteca dedicada a redes neuronais;
- **Pandas** – biblioteca usada para análise de dados;
- **Numpy** – biblioteca usada para análise de dados;
- **Matplolib** – biblioteca para visualização de dados.

Ambiente de desenvolvimento

Jupyter Notebook – ambiente de desenvolvimento ideal para o desenvolvimento de projetos, permitindo um *workflow* rápido graças às suas funcionalidades que permitem intercalar código com visualização de dados e correr rapidamente *snippets* de código.

Detalhes da implementação

Algoritmos

Neste projeto temos como objetivo implementar as seguintes técnicas de classificação:

- **Nearest Neighbor (K-NN)** - Tenta classificar um objeto com base nos K elementos mais próximos/semelhantes;
- **Naive Bayes (NB)** – Assume que todos os valores são independentes, tirando partido disso para a previsão de um elemento;
- **Support Vector Machines (SVM)** – Tenta traçar uma fronteira entre as várias classes e classificar o elemento consoante a região a que pertence;
- **Neural Networks (ANN)** – Rede constituída por vários "neurônios" que comunicam entre si de forma a prever um dado *outcome*;
- **Decision Trees –** Técnica que prevê os resultados através de uma abordagem algorítmica que identifica maneiras de dividir um conjunto de dados com base em diferentes condições.

Trabalho realizado

Até ao momento, estivemos a trabalhar no pré-processamento dos dados fornecidos, pois o *dataset* inicial continha demasiadas colunas, o que tornava muito complicada a análise dos dados de modo a construir um modelo de aprendizagem supervisionada aceitável.

Deste modo, removemos bastante colunas que continham informações que não acrescentavam grande valor para a construção do modelo (por ex. *id*, *member_id*), que continham informações que só são possíveis de conhecer depois do empréstimo ser feito (por ex. *funded_amnt*, *out_prncp*), que continham apenas um valor único (por ex. *policy_code*) ou que continham uma elevada percentagem de valores em falta (por ex. *verification_status_joint*, *annual_inc_joint*).

Para além da remoção das colunas, começamos também a resolver os casos em que faltavam valores em algumas colunas e a converter os valores de colunas categóricas em valores numéricos de forma a poderem ser analaisados.