MACHINE LEARNING: APLICAÇÃO DE REDES NEURAIS ARTIFICIAIS EM UM PROBLEMA DE RISCO DE CRÉDITO[1]

Gustavo Martignago Barbosa [2], Analucia Vieira Fantin[3]

[1] Vinculado ao projeto "Machine Learning: estudo de aplicações no setor financeiro"
[2]Acadêmico do Curso de Ciências Econômicas – ESAG – Bolsista PIBIC/CNPq
[3] Orientador, Departamento de Ciências Econômicas – ESAG – analucia.fantin@edu.udesc.br

O objetivo do projeto de pesquisa foi o estudo de modelos de *Machine Learning* através de uma aplicação prática na área de classificação de riscos de crédito. Para isso, primeiramente, foi necessário realizar uma capacitação em linguagem Python, que foi a linguagem de programação utilizada durante a pesquisa.

A primeira plataforma utilizada para o aprendizado da lógica e sintaxe básica foi a plataforma *Snakify (*https://snakify.org/pt/), a qual contém diversos exercícios que exploram conceitos, como: listas, dicionários, operadores lógicos e arrays. Durante esse processo, a orientadora Analucia Vieira Fantin também disponibilizou alguns exercícios fora da plataforma, com o intuito de aprofundar o conhecimento.

Após o desenvolvimento inicial da lógica e a familiarização com a linguagem escolhida, foi utilizada a plataforma *DataCamp* (https://www.datacamp.com/), que contém diversos cursos de diversas linguagens de programação. A plataforma possibilitou aprofundar o conhecimento na linguagem Python e estudar o uso de bibliotecas importantes para *Machine Learning*, como: Numpy, Pandas e *Sci-kit Learn*. O primeiro curso realizado dentro da plataforma foi o *Introduction to Statistics in Python*, com o foco em aprofundar questões matemáticas, usando a biblioteca Numpy, e criar uma base para os próximos cursos. O curso seguinte foi o *Supervised Learning with sci-kit learn*, cujo objetivo foi introduzir a biblioteca *Sci-kit Learn* e criar os primeiros algoritmos de *Machine Learning*. Este último curso mostra como tratar e separar o conjunto de dados, utilizando o método *train test split*, como criar modelos, como treiná-los e como avaliar e otimizar os mesmos. O terceiro e último curso realizado na plataforma foi o *Machine Learning for Finance in Python*, o qual é focado exatamente no que o projeto propõe: introduzir modelos lineares, aprofundar o conhecimento na biblioteca *Sci-kit Learn*, apresentar as Redes Neurais Artificiais e analisar os resultados.

Finalizada a fase de capacitação, foi iniciado o estudo de caso, o qual envolvia um problema de risco de crédito. Para o desenvolvimento do projeto, foi utilizado um conjunto de dados fictícios contendo 32000 linhas de dados. Cada linha contém 12 informações (variáveis, também denominadas *features*) sobre o sobre o histórico de pessoas que já solicitaram empréstimo em uma data passada. As variáveis são: idade, renda, tipo de moradia, motivo do empréstimo, nota do empréstimo, valor, status do empréstimo, taxa de juros, uma variável categórica que diz se a pessoa já tem histórico de atrasar o pagamento, tempo de trabalho em anos, histórico de crédito e porcentagem da renda. Com o intuito de reduzir o tempo de processamento dos modelos, foram utilizadas apenas 1000 linhas de dados, ou seja, informações de apenas 1000 mutuários. Com base nessas variáveis, o objetivo do modelo foi prever se o mutuário aponta ser um bom ou mau pagador. Essa última informação fica armazenada em uma variável binária denominada "status do empréstimo", sendo 0 (bom pagador) ou 1 (mau pagador).

As variáveis tipo de moradia, motivo do empréstimo, nota do empréstimo e histórico de atraso de pagamento, por serem variáveis categóricas, foram transformadas em valores binários através da função *OneHotEncoder*, da biblioteca *Sci-kit learn*.

Para desenvolvimento do modelo, os dados foram divididos em 2 grupos: grupo de treino contendo 70% dos dados, e grupo de teste contendo os restantes 30%. A divisão foi realizada através da função *train test split*, da biblioteca *Sci-kit learn*, onde os dados são distribuídos de forma aleatória e devem manter o tipo original de distribuição estatística.

O modelo da rede neural foi inicialmente construído a partir de alguns valores de hiperparâmetros mais comumente utilizados na literatura. Hiperparâmetros são parâmetros externos do modelo, que precisam ser definidos antes do treinamento do modelo. Exemplos comuns de hiperparâmetros incluem a taxa de aprendizado, o número de camadas em uma rede neural, o tamanho do lote (batch size) durante o treinamento, o número máximo de iterações (épocas), entre outros. O modelo de rede neural proposto neste trabalho foi construído contendo três camadas, a primeira sendo com 64 neurônios, de dimensão 25, correspondente ao número de colunas da base de dados, utilizando o método de ativação *'relu'*. A camada seguinte contendo 32 neurônios, também utilizando *'relu*' como sua função de ativação. E a última camada contendo apenas um neurônio, com a função de ativação *'sigmoid'*.

Um segundo modelo otimizado foi construído para comparação de resultados. Neste segundo modelo os hiperparâmetros foram ajustados e otimizados, com o intuito de se obter um melhor desempenho do modelo. Para a otimização, foi utilizada a rotina *RandomSearch* da biblioteca *keras-turner*. Esta rotina possui como entrada um conjunto de intervalos definidos para cada hiperparâmetro. O modelo de otimização busca, através de todas as possíveis combinações de hiperparâmetros, o melhor valor para a acurácia do modelo. Para os dados utilizados, a melhor acurácia encontrada foi de 0,9132.

Os resultados obtidos com os dois modelos de redes neurais, com e sem otimização dos hiperparâmetros, estão contidos na Tabela 1.

Tabela 1. *Resultados Rede Neural e Rede Neural Otimizada*

Métrica	Rede Neural	Rede Neural Otimizada
Precisão	0.8755	0.9132
Especificidade	0.9431	0.9763
Sensibilidade	0.6111	0.6667
F1 Score	0.6667	0.7198

O modelo de rede neural com otimização provou ser útil, pois observou-se um aumento em todas as métricas, indicando uma melhoria no desempenho do modelo, produzindo resultados mais precisos.

Palavras-chave: Machine Learning. Redes Neurais. Risco de Crédito