Previsão de gastos de um consumidor

Relatório de formalização

Clévio Orlando de Oliveira Junior
Departamento de Ciência e Tecnologia
Universidade Federal de São Paulo
São José dos Campos, Brasil
cleviojr20@gmail.com

Luiz Otávio Medeiros de Portella Passos
Departamento de Ciência e Tecnologia
Universidade Federal de São Paulo
São José dos Campos, Brasil
luizopassos33@gmail.com

I. Introdução e motivação

Dada uma base de dados referente a compras feitas em uma loja numa black friday , esse trabalho visa prever quanto um consumidor irá gastar e o que irá comprar baseado em seu perfil socioeconômico. Com essa análise cabe as lojas decidirem como precificar seus produtos atendendo assim os seus clientes da melhor forma e adotando estratégias de marketing e venda para de forma que maximize seus lucros. Vale ressaltar que esse trabalho usa apenas uma data do ano onde ocorrem baixa de preços, porém espera-se ser usado para o dia a dia do varejo.

II. Proposta

A. Objetivos

Entender as compras feitas e os hábitos dos consumidores utilizando das seguintes variáveis: idade, profissão, cidade etc.
Os dados acima podem ser utilizados para prever por exemplo, a quantia que um consumidor irá gastar no dia da black friday.
Identificar a loja que mais se adéqua ao perfil do consumidor Com os dados, poderia ser previsto também a idade do consumidor e o tipo de produto que costuma comprar.

B. O que será entregue

De acordo com [3], o problema é especificamente uma regressão onde pretende-se treinar um modelo que melhor prevê o atributo *Purchase* (valor da compra). Pretende-se comparar diferentes algoritmos por meio das medidas de avaliação. O modelo entregue será baseado um MLP (Multilayer Perceptron) que é uma rede neural com várias camadas, a eficácia desse algoritmo será comparada ao Random Forest Regressor e a Regressão Linear Múltipla entre outras.

III. Fundamentação Teórica

A. O que é uma Rede Neural

Em ciência da computação e campos relacionados, redes neurais artificiais (RNAs) são modelos computacionais inspirados pelo sistema nervoso central de um animal (em particular o cérebro) que são capazes de realizar o aprendizado de máquina bem como o reconhecimento de padrões. Redes neurais artificias geralmente são apresentadas como sistemas de ”neurônios interconectados, que podem computar valores de entradas”, simulando o comportamento de redes neurais biológicas.

B. O Porquê de uma Rede Neural

São esperados melhores resultados ao usar MLP já que: Segundo [6] o modelo de rede neural produziu predições mais acuradas de valores de energia metabolizável verdadeira de amostras de farinha de carne e ossos, quando comparado com modelos de mínimos quadrados parciais e regressão linear múltipla. Em [7] mostrou-se que que o método de rede neural é mais acurado do que os modelos tradicionais de regressão para a predição de produção de ovos, e os autores de [8] verificaram que o modelo de rede neural de base radial apresentou predições mais acuradas do crescimento de frangos de corte do que às obtidas por modelos de regressão múltipla.

C. o que é um Multilayer Perceptron

Um perceptron e um tipo de rede neural artificial com alimentação pra frente mais simples que existe. A perceptron multicamadas (MLP) é uma rede neural semelhante à perceptron, mas com mais de uma camada de neurônios em alimentação direta [1]. Tal tipo de rede é composta por camadas de neurônios ligadas entre si por sinapses com pesos.

D. Medidas de Avaliação

Medidas que dizem o quão bom um modelo é a partir do seu treino e teste, a medida de avaliação mais comum é a acurácia que é a proporção de predições corretas, sem levar em consideração o que é positivo e o que é negativo. Esta medida é altamente suscetível a desbalanceamentos do conjunto de dados e pode facilmente induzir a uma conclusão errada sobre o desempenho do sistema, logo serão usadas outras duas métricas mais consistentes.

1) R2 Score: O coeficiente de determinação, também chamado de R², é uma medida de ajustamento de um modelo estatístico linear generalizado, como a regressão linear, em relação aos valores observados. O R² varia entre 0 e 1, indicando, em percentagem, o quanto o modelo consegue explicar os valores observados. Quanto maior o R², mais explicativo é o modelo, melhor ele se ajusta à amostra. È dada pela fórmula:

$$R^2 = \frac{SQ_{\text{exp}}}{SQ_{\text{tot}}} = 1 - \frac{SQ_{\text{res}}}{SQ_{\text{tot}}}.$$

2) Raiz Quadrada do Erro Médio (RMSE): A Raiz Quadrada do Erro Médio (RMSE) é o desvio padrão do erro de predição. O erro de predição é a medida da distância de cada ponto da linha da regressão. Essa é a medida de quão espalhados os pontos estão, em outras palavras, diz o quão concentrados os dados estão da linha do melhor ajuste.
A fórmula é:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(f_i - o_i)^2}$$

IV. Trabalhos Relacionados

Na base de dados desse paper que pode ser encontrada em [1] foram realizadas análises estatísticas como vistas em [2], [3] e [4]. Foram aplicadas tarefas de regressão para predizer o atributo Purchase em [9], [10] e [11], sendo esse Regressão linear simples, regressão polinomial, elastic net, lasso ridge e Random forest regressor. Em [5], [6] e [7] têm-se artigos que usam a tarefa de regressão em outras bases para comparar diferentes predições a partir de algoritmos de regressão básica como linear e polinomial com o MLP.

V. Base de Dados

O dataset tem 537577 linhas, e 12 colunas. As linhas se referem as compras feitas e as colunas aos atributos que estão descritos abaixo:

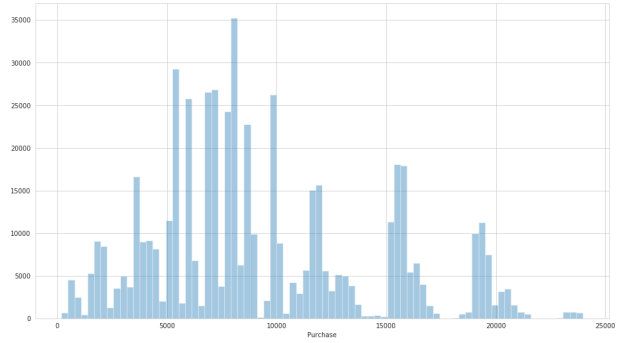
- *User_id*: Id único de cada consumidor, existem 5891.
- *Product_id*: Id único do produto, existem 3623.
- *Gender*: Indica o gênero de quem fez a compra.
- *Age*: Indica o grupo de idade de quem fez a compra.
- *Ocuppation*: Indica a profissão, já definido de 0 a 20.
- *City_category*: Cidade onde mora o consumidor, estão divididas em 'A', 'B' e 'C'.
- *Stay*_In_*Current_City_Years*: Tempo que o consumidor mora na mesma cidade.
- *Marital_Status*: 1 caso o consumidor seja casado e 0 caso contrário
- *Product_Category*_1 até _3: Categoria do produto, todos os 3 estão rotulados com números.
- *Purchase*: Valor gasto.

É válido dizer que a maioria dos atributos com exceção de *Purchase*, *User*_id e *Product*_id são categóricos.

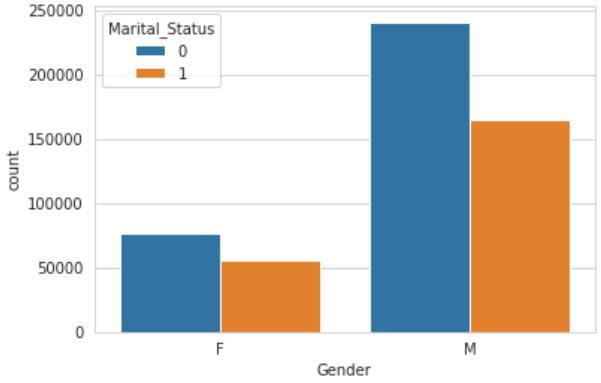
VI. Análise Estatística da Base

Segundo [2] e [5] obtêm-se as seguintes informações:

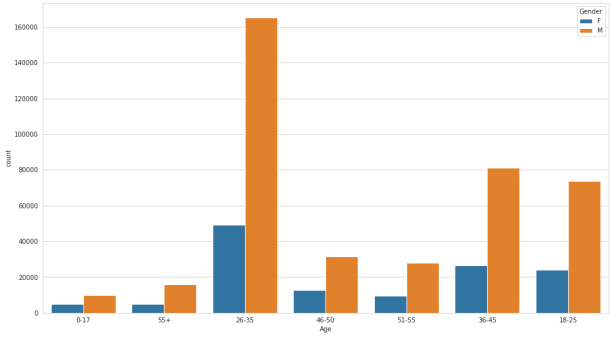
- A faixa etária de 26 a 45 anos gastou mais de 3 bilhões.



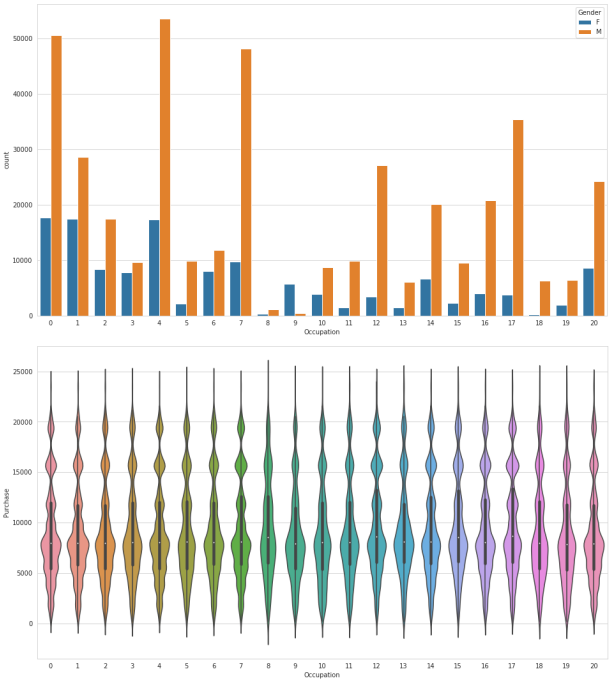
- O valor de compra está mais concentrado entre 5000 e 10000.



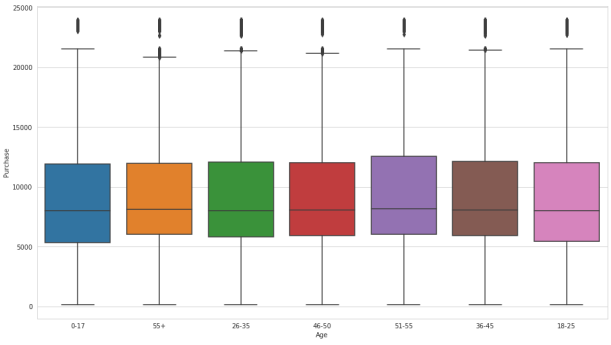
- A maioria dos clientes que vão as compras são homens solteiros.



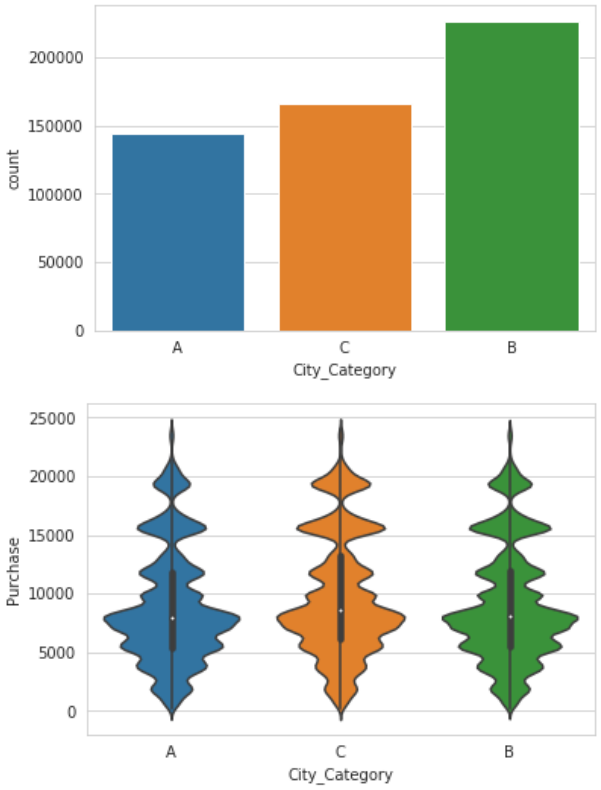
- A faixa etária predominante é de 26 a 35 anos, isso explica o fato da maioria dos clientes serem solteiros.



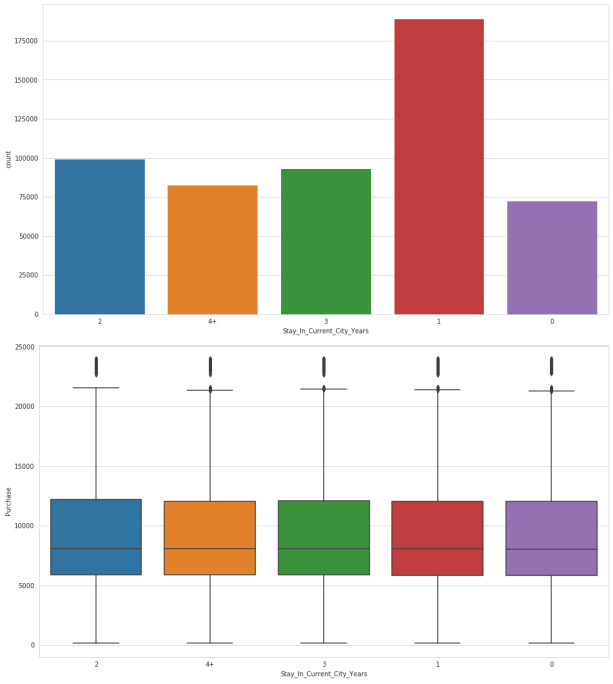
- Apesar de não ser possível saber o nome de cada profissão, as profissões 4 e 0 são as que mais compraram, porém, estão em maior número. Isso significa que a profissão 13, que estava em menor número mas teve uma valor similar nas compras, é a mais bem paga. Também é possível ver que trabalhadores estão presentes em maior quantidade que trabalhadoras.



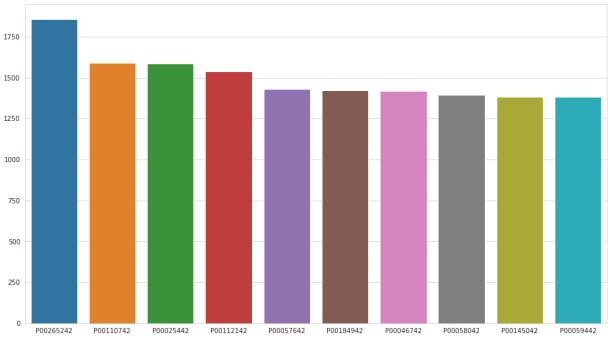
- A quantidade de compras em cada faixa etária é semelhante, isso significa que os produtos em promoção podiam ser utilizados por qualquer público. Sabe-se que as vendas foram mais aproveitadas pela faixa de 26 a 35 anos e que os que menos aproveitaram tinham entre 0 e 17 anos, porém, nesses 2 casos, a quantidade de compras é similar.



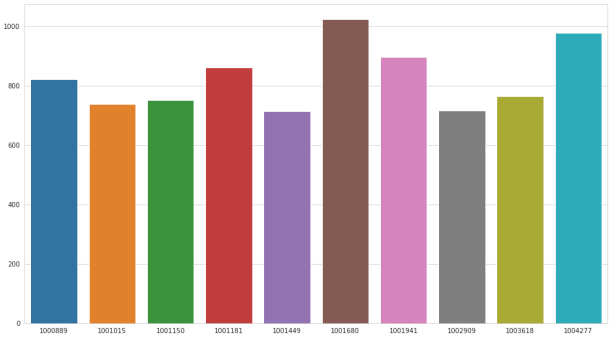
- A maioria das pessoas que compraram vieram da cidade B, porém, a quantidade de compras estava bem equilibrada e a menor parte das pessoas vieram da cidade A, e através disso infere-se que as pessoas da cidade A possuem uma melhor condição financeira que as outras.



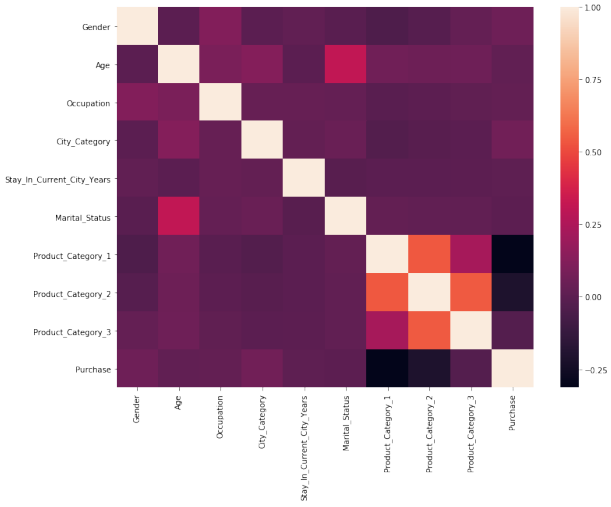
- A maioria das pessoas que compraram haviam se mudado há pouco tempo, o que confirma o fato de que as pessoas gastam mais quando mudam de cidade.



- Produtos mais vendidos.



- Consumidores que mais gastaram.



- Do heatmap de [10] vê-se que a correlação entre *Product_Category_1* e *Product_Category_2* é alta comparada com a correlação da *Product_Category_1* e *Product_Category_3*.
- Também é possível observar que existe uma correlação negativa entre *Purchase* e *Product_Category_1*, e também entre *Purchase* e *Product_Category_2*.
- A idade e o estado civil possuem uma alta correlação.
- Nenhuma variável possui alta correlação com *Purchase* e isso mostra que a mesma depende de todas as outras variáveis.

VII. Protocolo Experimental

A. Pré processamento

No estudo realizado por [5], nas colunas *Product_Category_2* e *Product_Category_3* têm 31% e 69% dos dados com *NaN* nas linhas respectivamente (valores faltantes), em *Product_Category_1* o tipo do dado é inteiro, logo os missing values serão substituídos por 0.
Também precisa-se mudar o tipo de dados de float para int.
Substitui-se string '4+' em *Age* pelo inteiro 4 e mudar o tipo de dados para int.
Regularizar os atributos categóricos para deixar os dados acessíveis pelo algoritmo.
Aplica-se o OneHotEncoding em algumas colunas. Perceba que não foi incluso *Gender* já que ele é binário, M ou F.

B. Divisão da Base

Para uma melhor comparação com outros estudos feitos, adota-se o mesmo padrão divisão feito por esses, sendo 70% da base para treino e os 30% restantes para teste. Para esse problema a linguagem usada será python, e a biblioteca usada para o aprendizado será a sklearn e keras.

C. Algoritmos e Parâmetros

Para a obtenção de melhores resultados vamos usar o MLP de 2 bibliotecas diferentes, o primeiro MLPRegressor() da biblioteca em Python Scikit-learn, sendo esse mais simples pois não permite o ajuste da rede neural por camadas, ou seja, as camadas de neurônios de entrada, intermediárias e a de saída possuem as mesmas características. A segunda biblioteca que será usada sera o Keras, também em Python, onde o MLP pode ser encontrado importando Sequential(), que permite o ajuste específico das características de cada camada de neurônios.

1) Usando Scikit-learn: Características gerais da rede neural no primeiro experimento:

- Neurônios da camada oculta: 200
- Função de Ativação: ReLU
- Nº máximo de épocas : 1000
- Tamanho do Lote: 256
- Aceitação de 10 épocas com melhora menor que 0.0001

Com os melhores resultados usando essa biblioteca, os mesmos parâmetros serão selecionados para treinar o MLP do keras.

2) Usando Keras: Com essa biblioteca há uma maior liberdade quanto a criação da rede neural, nesse caso serão colocadas 5 camadas de neurônios, uma pra entrada, outra pra saída e três camadas ocultas. A camada de entrada terá 128 saídas para as camadas ocultas, essa por sua vez 256, até que a ultima terá apenas uma saída. O tamanho do lote será reduzido para 64, visando maior precisão sem afetar a performance já que as camadas de entrada e ocultas possuem maior números de ligações. Características gerais da rede neural no segundo experimento:

- Camadas Ocultas: 3
- Função de Ativação: ReLU(Camadas de entrada e ocultas) e Linear(Camada de saída)
- Nº máximo de épocas: 200
- Tamanho do Lote: 64
- Taxa de Aprendizagem:Constante

Um novo experimento foi feito, dessa vez usando o tamanho do lote em 256 e as épocas em 513, assim os parâmetros serão os mesmos do experimento do scikit.

- Camadas Ocultas: 3
- Função de Ativação: ReLU(Camadas de entrada e ocultas) e Linear(Camada de saída)
- Nº máximo de épocas: 513
- Tamanho do Lote: 256
- Taxa de Aprendizagem:Constante

VIII. Discussão dos Resultados

1) Usando Scikit-learn: No primeiro experimento, após 513 épocas o algoritmo não obteve melhora na validação sendo paralisado com uma taxa de acerto de 0,526043 (aproximadamente 53%). As medidas de avaliação obtidas pelo teste foram as seguintes:

- Taxa de acerto do teste (*R2 score*) : 0,533096355036375
- Raiz quadrada do erro médio (*RMSE*): 3403,032482510383

Com uma taxa de acerto da R2 em 53% e o RMSE em 3403 já se obteve melhora em relação aos algoritmos de regressão sem o uso de redes neurais, apesar de ainda não obtermos resultados próximos a 90%

A. Usando Keras

Após a rodagem do segundo experimento obteve-se os seguintes resultados:

- Taxa de acerto do teste (*R2 score*) : 0,6276900911393912
- Raiz quadrada do erro médio (*RMSE*): 3038,8184870067175

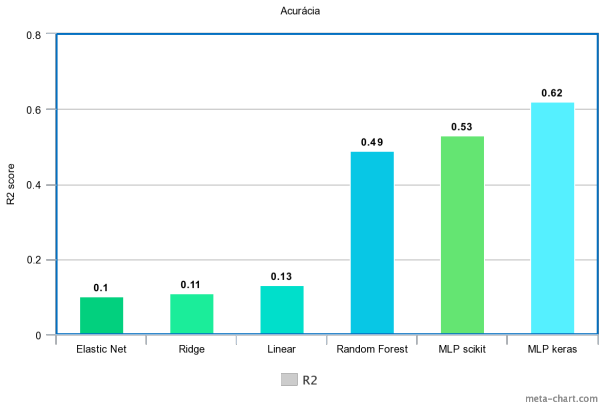
Percebe-se a melhora em ambas medidas de avaliação. Antes de qualquer conclusão serão analisados os dados do terceiro experimento também com uso da biblioteca keras.

Após a rodagem das 513 épocas obteve-se os seguintes resultados:

- Taxa de acerto do teste (*R2 score*) : 0,6277264499620959
- Raiz quadrada do erro médio (*RMSE*): 3038,6701018071926

Pouca diferença foi notada. A melhora pode ser ligada ao fato da personalização da rede por camadas.

Apesar dos resultados não terem atingido acurácias de R2 superiores a 90%, obtiveram melhores avaliações que outros algoritmos de machine learning estudados em [9] e [11], isso esta diretamente relacionado ao fato de que redes neurais podem lidar com comportamentos não lineares quanto que alguns outros algoritmos não. Os gráficos abaixo são comparativos em relação a R2 score entre os estudos de [9], [10] e os usando o MLP.



Apesar de melhores resultados nas medidas de avaliação, não se alcançaram acurácias da R2 próximas a 90%, isso pode estar relacionado há diferentes fatores, o principal deles é a mostrado na tabela de correlação vista na seção Análise dos Dados, já que não há fatores específicos que possam prever unicamente o alvo que é Purchase, necessitando assim da junção de todos outros atributos.
Segundo [11], o atributo que mais afeta em Purchase é o tipo de produto, o qual neste estudo é eliminado da previsão do estimador.
Além disso trabalha-se com uma base de mais de meio milhão de atributos, o que dificulta ainda mais a performance dos algoritmos.

IX. Conclusão

Como visto na seção acima, o uso de rede neural para a previsão do preço se mostrou melhor do que algoritmos mais simples de regressão linear. Apesar de ainda apresentar um erro alto a rede neural se mostrou no melhor caminho para a predição, desta forma estudos mais avançados têm um ponto de partida no que diz respeito á escolha do algoritmo quanto a essa base de dados e objetivos propostos por esse paper. Os experimentos se mostraram difíceis de estimar parâmetros empiricamente, o que é ainda é uma grande dificuldade da inteligência artificial.

Referências

[1] Mehdi Dagdoug, "Black Friday", Disponível em: https://www.kaggle.com/mehdidag/black-friday [Acessado 10 de Abril, 2019]

[2] Loai Abdalsam "Sales Data Analysis Report", Disponível em: https://www.kaggle.com/loaiabdalslam/sales-data-analysis-report [Acessado 10 de Abril, 2019]

[3] Victor Hugo Pereira "Black Friday Datasets", Disponível em: https://www.kaggle.com/panamby/black-friday-dataset [Acessado 10 de Abril, 2019]

[4] Sp "Black Friday data exploration", Disponível em: https://www.kaggle.com/shamalip/black-friday-data-exploration [Acessado 5 de Maio, 2019]

[5] Sean Choi"BlackFriday_EDA_RandomForestPrediction", Disponível em: https://www.kaggle.com/sungsujaing/blackfriday-eda-randomforestprediction [Acessado 5 de maio, 2019]

[6] Perai, A. H. et al. A comparison of artificial neural networks with other statistical approaches for the prediction of true metabolizable energy of meat and bone meal. Poultry Science, Champaign, v. 89, p. 1562-1568, July 2010.

[7] Wang, B. Y.; Chen, S. A.; Roan, S. W. Comparison of regression and artificial neural network on egg production. Journal of Animal and Veterinary Advances, Kuala Lumpur, v. 11, n. 14, p. 2503-2508, 2012.
[8] Ahmadi, H.; Golian, A. Growth analysis of chickens fed diets varying in the percentage of metabolizable energy provided by protein, fat, and carbohydrate through artificial neural network. Poultry Science, Champaign, v. 89, n. 1, p. 173-179, Jan. 2010.
[9] Roshan Sharma "Black Friday Regression Analysis" Disponível em: https://www.kaggle.com/roshansharma/black-friday-regression-analysis [Accessado 6 de Junho, 2019].
[10] Sriharsha Atyam "EDA and Predictions using various ML models" Disponível em: https://www.kaggle.com/sriharshaatyam/eda-and-predictions-using-various-ml-models [Accessado 6 de Junho, 2019].
[11] Alireza "Black Friday - Analysis, Regression and Clustering" Disponível em: https://www.kaggle.com/arkhoshghalb/black-friday-analysis-regression-and-clustering [Accessado 6 de Junho, 2019].