UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO
CENTRO TECNOLÓGICO
DEPARTAMENTO DE ENGENHARIA ELÉTRICA
PROJETO DE GRADUAÇÃO

Alan Zanão da Penha

Predição de vendas no setor de moda baseado em Aprendizado de Máquina no contexto da competição Data Mining Cup 2018

Vitória-ES

Julho/2018

Alan Zanão da Penha

Predição de vendas no setor de moda baseado em Aprendizado de Máquina no contexto da competição Data Mining Cup 2018

Parte manuscrita do Projeto de Graduação do aluno Alan Zanão da Penha, apresentado ao Departamento de Engenharia Elétrica do Centro Tecnológico da Universidade Federal do Espírito Santo, como requisito parcial para obtenção do grau de Engenheiro Eletricista.

Vitória-ES

Julho/2018

Alan Zanão da Penha

Predição de vendas no setor de moda baseado em Aprendizado de Máquina no contexto da competição Data Mining Cup 2018

Parte manuscrita do Projeto de Graduação do aluno Alan Zanão da Penha, apresentado ao Departamento de Engenharia Elétrica do Centro Tecnológico da Universidade Federal do Espírito Santo, como requisito parcial para obtenção do grau de Engenheiro Eletricista.

Aprovado em 04 de julho de 2018.

COMISSÃO EXAMINADORA:

Prof. Dr. Jorge Leonid Aching Samatelo
Universidade Federal do Espírito Santo
Orientador

Prof. Dr. Patrick Marques Ciarelli
Universidade Federal do Espírito Santo
Examinador

Msc. Clebeson Canuto dos Santos
Universidade Federal do Espírito Santo
Examinador

Vitória-ES

Julho/2018

A minha mãe, guerreira e exemplo de vida.

AGRADECIMENTOS

A minha mãe, pelo suporte ao longo de todos esses anos, que mesmo com todas as dificuldades da vida nunca desistiu de nenhuma batalha, superando o impossível para ver os três filhos formados, e pelos valiosos princípios de vida que foram essenciais norteadores do meu caminho acadêmico.

As minhas irmãs Amanda e Keila, pelo incentivo durante o meu percurso e por torcerem pelo meu sucesso.

Ao meu orientador Prof. Dr. Jorge Leonid Aching Samatelo, pela disposição em todas as reuniões realizadas, pela compreensão nos momentos em que tive que me dedicar a problemas pessoais e pelos ensinamentos de uma área tão vasta como a de Aprendizado de Máquina.

Aos meus amigos do Aerodesign. Juntos formamos verdadeiramente uma equipe que com muita dedicação, após muitos finais de semana e noites viradas na UFES, conseguimos voar alto em 2013, obtendo o que considero a maior conquista que já obtive. Projeto que guardo com carinho pois é o lugar onde me apaixonei por competições, evolui muito como pessoa e tive os melhores momentos de toda a minha graduação.

Aos meus amigos do curso de Engenharia Elétrica por caminharmos e crescermos juntos.

Um agradecimento especial a dois grandes amigos de intercâmbio: Feliz Lintz e Alex Brazadio, que em 2015 me apresentaram a competição *Data Mining Cup* e foram meus companheiros de equipe. Essa que veio a se tornar a motivação pessoal para esse trabalho.

À banca examinadora pela aceitação do convite e pelo tempo investido na leitura e avaliação desse trabalho.

A todos vocês, muito obrigado!

RESUMO

A predição de vendas no setor de moda possui peculiaridades em relação às demais áreas, sendo que uma boa acurácia representa um diferencial competitivo importante das empresas. As técnicas envolvendo Aprendizado de Máquina na área começaram a ser desenvolvidas nas últimas duas décadas em substituição aos métodos tradicionais, que envolviam formulações matemáticas simplificadas para previsão de vendas de grupos de produtos em intervalos de tempo normalmente mensais. O setor de moda, por possuir uma grande variedade de produtos com curto ciclo de vida, tem o interesse de desenvolver modelos com previsão diária de vendas para cada produto em específico. O presente Projeto de Graduação propõe a análise de modelos de previsão de vendas diária, baseados em Aprendizado de Máquina, de uma loja online alemã de artigos esportivos, cujo banco de dados é disponibilizado no contexto da competição *Data Mining Cup 2018*. Para auxiliar e facilitar a tarefa foi realizada uma análise exploratória dos dados, uso de técnicas de pré-processamento e *Feature Engineering*. Diversos classificadores foram treinados utilizando validação cruzada e validação por *holdout*, além da seleção dos melhores hiperparâmetros através da técnica *Grid Search*. Especificamente, é obtida uma média da área sob a curva de característica de operação do receptor de $0,823$ e $0,820$, utilizando os modelos de *MultiLayer Perceptron* e *Gradient Boosting Machine*, quando testado no banco de dados da competição.

Palavras-chave: Predição de vendas, Varejo de Moda, Aprendizado de Máquina.

ABSTRACT

The sales forecast in the fashion sector has specific features in relation to the other areas, and a good accuracy represents a competitive advantage of the companies. Machine Learning techniques in the field began to be developed over the past two decades in place of traditional methods, which involved simplified mathematical formulations for predicting sales of product groups at normally monthly intervals. The fashion industry is interested in developing models with daily sales forecast for each specific product because the wide variety of products with a short life cycle. The present Graduation Project proposes to analyse daily sales forecast models, based on Machine Learning, of a German online store of sporting goods, whose database is made available in the context of the competion *Data Mining Cup 2018*. An Exploratory Data Analysis, pre-processing techinques and Feature Engineering are performed before development the models. Then, several classifiers are trained using cross-validation and Grid Search to obtain the hyperparameters that provide the best performance. Specifically, an area average under the receiver operating characteristic curve of 0.823 e 0.820 is obtained using MultiLayer Perceptron and Gradient Boosting Machine, when tested in the competition database.

Keywords: Sales forecasting, Fashion retail, Machine Learning.

LISTA DE FIGURAS

Figura 1 – Regressão Logística 22
Figura 2 – Estrutura da árvore do problema *Iris Flower* 23
Figura 3 – Arquitetura de uma rede MLP 23
Figura 4 – Arquitetura de um *ensemble bagging* 24
Figura 5 – Arquitetura de um *ensemble boosting* 25
Figura 6 – Arquitetura de um *ensemble stacking* 25
Figura 7 – Arquitetura de *Random Forest* 26
Figura 8 – Exemplo de modelos com *underfitting* e *overfitting* 27
Figura 9 – Validação *holdout* 27
Figura 10 – Validação cruzada 28
Figura 11 – Quantidade de variáveis *size* para um mesmo PID 33
Figura 12 – Quantidade de produtos por marcas. 34
Figura 13 – Quantidade de produtos em estoque para produtos com menos de 50 unidades. 34
Figura 14 – Mapa de calor das variações de preços durante o período de treino e teste. 35
Figura 15 – Série temporal de vendas por dia da loja. 36
Figura 16 – Mapa de calor de vendas de acordo com as categorias do atributo *size*. 37
Figura 17 – Mapa de calor de vendas de acordo com as categorias do atributo *color*. 37
Figura 18 – Mapa de calor de vendas de acordo com as categorias do atributo *brand*. 38
Figura 19 – Exemplo da aplicação de codificação *one-hot-encoding*. 40
Figura 20 – Fluxograma de trabalho. 43
Figura 21 – Curva ROC com validação cruzada 48
Figura 22 – Curva ROC com validação *holdout* 49
Figura 23 – Atributos do conjunto de dados do desafio DMC 2016 56
Figura 24 – Modelagem da equipe vencedora da competição DMC 2016 57
Figura 25 – Modelagem vencedora da competição AirBnB 2016 58

LISTA DE TABELAS

Tabela 1 – *Dataset* estendido 16
Tabela 2 – Atributos disponíveis no *dataset train* 30
Tabela 3 – Cinco primeiros registros do conjunto de dados *train* 30
Tabela 4 – Atributos disponíveis no *dataset items* 31
Tabela 5 – Cinco primeiros registros do conjunto de dados *items* 31
Tabela 6 – Cinco primeiros registros do conjunto de dados *prices* 31
Tabela 7 – *Ranking* dos cinco valores de *size* mais comum 33
Tabela 8 – Feriados alemães considerados 40
Tabela 9 – *Dataset* estendido 41
Tabela 10 – Melhores hiperparâmetros obtidos pelo *Grid Search*. 47
Tabela 11 – Resultado do treinamento com validação cruzada. 47
Tabela 12 – Resultado do treinamento com validação cruzada usando como treinamento os meses de Outubro de 2017 a Janeiro de 2018. 48
Tabela 13 – Resultado do treinamento com validação *holdout* usando como treinamento os meses de Outubro de 2017 a Dezembro de 2017 e como teste o mês Janeiro de 2018. 50

LISTA DE ABREVIATURAS E SIGLAS

AUC	*Area Under the Curve*
AM	Aprendizado de Máquina
DMC	*Data Mining Cup*
DT	*Decision Tree*
EDA	*Exploratory Data Analysis*
FE	*Feature Engineering*
GBM	*Gradient Boosting Machine*
GPU	*Graphics Processing Unit*
MD	Mineração de Dados
ML	*Machine Learning*
MLP	*Multilayer Perceptron*
PID	*Product Identification*
PG	Projeto de Graduação
RF	*Random Forest*
ROC	*Receiver Operating Characteristic*
SGD	*Stochastic Gradient Descent*
SKU	*Stock Kepping Units*
UFES	Universidade Federal do Espírito Santo

SUMÁRIO

1 INTRODUÇÃO . . . 12
1.1 Apresentação . . . 12
1.2 Objetivos . . . 13
1.3 Metodologia Adotada . . . 14
1.4 Trabalhos relacionados . . . 15
1.5 Estrutura do Texto . . . 16

2 TEORIA . . . 18
2.1 Introdução . . . 18
2.2 Predição de vendas . . . 18
2.3 Aprendizado de Máquina (AM) . . . 20
2.3.1 Definição . . . 20
2.3.2 Terminologia . . . 20
2.4 Modelos . . . 21
2.4.1 Regressão logística . . . 21
2.4.2 *Decision Tree* (DT) . . . 22
2.4.3 *Multi Layer Perceptron* (MLP) . . . 22
2.4.4 *Ensemble Learning* . . . 24
2.4.5 *Random Forest* (RF) . . . 25
2.5 Validação . . . 26
2.6 Resumo . . . 28

3 SOLUÇÃO PROPOSTA . . . 29
3.1 Introdução . . . 29
3.2 Desafio . . . 29
3.2.1 Cenário . . . 29
3.2.2 Tarefa . . . 30
3.2.3 Dados . . . 30
3.3 Análise exploratória dos dados . . . 32
3.3.1 Atributo: PID e *size* . . . 32
3.3.2 Atributo: *Brand* . . . 33
3.3.3 Atributo: *Stock* . . . 33
3.3.4 Atributo: *Price* . . . 35
3.3.5 Atributo: *Units* . . . 35
3.4 Pré-Processamento . . . 38
3.4.1 Valores faltantes . . . 38

3.4.2 Novos atributos 39
3.4.3 Codificações 39
3.4.4 Normalização 40
3.5 Modelagem do problema 41
3.6 Fluxograma 42
3.7 Resumo 43

4 RESULTADOS 44
4.1 Introdução 44
4.2 Detalhes de implementação 44
4.3 Métricas 44
4.4 Treinamento 46
4.5 Resultados 47
4.6 Resumo 49

5 CONCLUSÕES E PROJETOS FUTUROS 51
5.1 Conclusões 51
5.2 Temas a serem pesquisados 51

REFERÊNCIAS 52

APÊNDICES 55

APÊNDICE A – COMPETIÇÕES 56
A.1 Data Mining Cup 2016 56
A.2 AirBnB: Reservas de novos usuários 58

APÊNDICE B – AGREGAÇÃO DO ATRIBUTO *SIZE* 60

1 INTRODUÇÃO

1.1 Apresentação

A predição de vendas assume um papel importante nas empresas, sendo uma vantagem competitiva, por buscar equilíbrio entre estoque mínimo e disponibilidade quando um produto é requisitado pelos consumidores. Historicamente, métodos qualitativos e quantitativos são utilizados, usando opinião de especialistas e métodos matemáticos simplificados, respectivamente (BESSA, 2018). Aksoy, Ozturk e Sucky (2012) ressaltam que os atores envolvidos não são apenas o fabricante e o varejista, mas também todos os que compõem a cadeia de suprimentos (*Supply Chain*) do produto.

Segundo Garcia e Reis (2006), uma adequada previsão é justificada pois uma produção maior do que a demanda implica em: (*i*) maiores custos de operacionalização envolvendo produção, logística, inventário e armazenamento; (*ii*) oportunidade do capital empatado/imobilizado; (*iii*) obsolência de produtos; (*iv*) retorno e devoluções.

Em contrapartida, uma produção menor do que a demanda implica em: (*i*) falta de produtos e consequente perda da confiança dos clientes; (*ii*) perda de economias de escala[1].

Logo, os modelos devem buscar prover uma previsão que estabeleça um equilíbrio entre a sub e sobre demanda, uma vez que um desequilíbrio para qualquer um dos lados incorre em perdas financeiras, tanto diretas como indiretas.

Especificamente na indústria da moda, a predição de demanda é complexa por possuir peculiaridades. Segundo Yu, Hui e Choi (2012), Mostard, Teunter e Koster (2011): as companhias trabalham com uma extensa variedade de produtos com curto ciclo de vida, fortemente influenciado por sazonalidades, eventos promocionais, condições climáticas, campanhas de marketing, datas festivas, fatores sociais e econômicos.

Thomassey e Fiordaliso (2006) também ressaltam sobre essa complexidade: a maioria dos itens é substituída na próxima coleção e suas vendas, que muitas vezes têm vida curta (6-12 semanas), sofrem interferência de inúmeros fatores, sendo que esses não são estritamente controlados nem identificados, podem depender do próprio item (cor, preço), distribuidor (número de lojas, marketing), clientes ou fatores externos (meteorologia, férias). Além

[1] Uma economia de escala é caracterizada pelo fato de que quanto maior a quantidade ordenada, menor é o custo do pedido por unidade de produto. Alguns fatores que levam a estas economias de escala são custos fixos de transporte, emissão de ordens, recebimento de pedidos, descontos por quantidade, dentre outros.

disso, esses fatores nem sempre estão disponíveis e apresentam influências diferentes nas vendas.

Devido ao mercado globalizado, crescimento dos *e-commerces* e desenvolvimento de *hardwares* e *softwares*, técnicas mais avançadas de predição utilizando Aprendizado de Máquina (AM) começaram a serem desenvolvidas nas últimas duas décadas, em detrimento dos métodos tradicionais (NENNI; GIUSTINIANO; PIROLO, 2013).

A melhora na previsão gera uma série de benefícios econômicos, sendo almejada por vários varejistas. A empresa "Zara" obteve um aumento real do faturamento no período de liquidação de 6% (correspondente a 90 milhões de dólares em 2008) e a empresa "Rue La La" estimou um aumento de 10%, segundo os estudos em (CARO; GALLIEN, 2012) e (FERREIRA; LEE; SIMCHI-LEVI, 2016).

Batista (2017) realizou um mapeamento dos estudos que utilizam técnicas de previsão na indústria de vestuário, tendo sido selecionados 25 artigos entre 2005 e 2015. 48% dos artigos analisados adotam métodos híbridos de previsão de vendas, enquanto 32% utilizam métodos individuais já desenvolvidos por autores em livros de referências. Os demais são referentes a revisão bibliográfica e proposta de métodos. Nos métodos híbridos, as técnicas mais relevantes identificadas foram: Redes Neurais, Aprendizado de Máquina e Lógica Fuzzy. No geral, os estudos apresentam as características: previsão de vendas de grupo de produtos, com intervalo semanal ou mensal, onde estão disponíveis dados históricos de coleções passadas.

1.2 Objetivos

Objetivo Geral

- Desenvolver um modelo de previsão de vendas diárias de produtos de uma loja de artigos esportivos utilizando AM, no contexto da competição *Data Mining Cup* (DMC) 2018.

Objetivos Específicos

- Realizar uma revisão bibliográfica de trabalhos que estudam o problema de predição de vendas;
- Desenvolver modelos de classificação binária da data de venda de um conjunto de produtos;
- Avaliar diferentes configurações de modelos e *Feature Engineering* (FE) buscando a que melhor se adequa à predição de venda;
- Submeter o resultado na competição DMC 2018;
- Submeter um artigo sobre a pesquisa efetuada ao Congresso Científico Têxtil e de Moda.

1.3 Metodologia Adotada

O trabalho foi desenvolvido conforme as etapas apresentadas a seguir.

- **Etapa 1.** Estudo teórico
  1. Estudo sobre os modelos de AM, baseado principalmente nos livros (GÉRON, 2017) e (HENRIK; MARK, 2017).
  2. Revisão sobre competições passadas, tanto da DMC como da plataforma *Kaggle*.
- **Etapa 2.** Análise dos dados.
  3. Utilização dos módulos *pandas, matplotlib, seaborn* e *bokeh* para visualização dos dados.
  4. Busca por *outlier*, valores faltantes ou corrompidos e dados com falta de padronização.
- **Etapa 3.** Pré processamento e FE.
  5. Derivação e criação de atributos.
  6. Codificação dos dados categóricos utilizando *one-hot-encoding* e normalização dos dados numéricos.
- **Etapa 4.** Treino e testes dos modelos de classificação.
  6. Teste para selecionar os modelos que forneçam bons resultados e que são treinados em um intervalo de tempo adequado ao disponível para a competição.
  7. Seleção dos melhores hiperparâmetros através da técnica de *Grid Search*.

8. Treinamento e teste dos modelos utilizando avaliação de validação cruzada e *holdout*.
9. Teste utilizando diferentes configurações de *ensemble stacking*.

- **Etapa 5.** Documentação
  9. Documentar as etapas de desenvolvimento.
  10. Desenvolver a dissertação do PG.

1.4 Trabalhos relacionados

Clearance Pricing Optimization for a Fast-Fashion Retailer: Caro e Gallien (2012) implementaram um sistema de remarcação de preço na empresa "Zara", que até 2007 era feito de forma manual e informal. O desafio enfrentado consistia em precificar e predizer a demanda de produtos que nunca foram vendidos anteriormente e que representam a maior parte das vendas da empresa estudada. Após a implementação de um modelo paramétrico houve um aumento de aproximadamente 6% na receita.

Analytics for an Online Retailer: Demand Forecasting and Price Optimization: Ferreira, Lee e Simchi-Levi (2016) apresentam um estudo de caso da loja online de moda "Rue La La", que enfrentava a mesma dificuldade da empresa "Zara", possuindo muitos artigos com poucos históricos de vendas. Foram utilizados técnicas de *clustering* e AM. Os resultados apresentaram um aumento de 10% na receita no grupo de teste.

Fourier Analysis for Demand Forecasting in a Fashion Company: (FUMI et al., 2013) é um trabalho no qual os autores fazem uso da transformada de Fourier para previsão da demanda de artigos de moda de uma companhia italiana, comparando a acurácia com heurísticas tradicionalmente utilizadas (predileção, projeção usando técnicas matemáticas e estatísticas simplificadas). O estudo buscou propor o uso de uma técnica simples para empresas de pequeno e médio porte, que não podem arcar com o custo de um sistema complexo, e que ofereça resultados melhores do que os métodos tradicionais.

Ferramentas para a indústria de vestuário: previsão da demanda: Ekami, Ming e Brito (2014) realizaram um trabalho que avalia métodos de previsão de demanda como ferramenta de otimização do planejamento estratégico da empresa. A ótica adotada é em relação a indústria de vestuário e não do varejista. No estudo de caso realizado, utilizou-se método de média móvel e índice de sazonalidade, demonstrando que o uso de técnicas simplificadas ainda é o padrão no cenário nacional, principalmente nas empresas de médio porte.

A seguir são resumidas as principais informações e características de estudos correlatos com a predição a ser desenvolvida neste trabalho, e que demonstra a dificuldade a ser enfrentada no que tange a comparação de estudos, pois existem divergência dos estudos.

Tabela 1 – *Dataset* estendido

Estudo	Histórico	Demanda	Intervalo de previsão	Técnicas
Fildes, Goodwin e Lawrence (2006)	3 anos	agregada	semanal	Lógica *fuzzy*, redes neurais e algoritmos evolutivos
Thomassey e Fiordaliso (2006)	3 anos	agregada	semanal	*Clustering* e DT
Au, Choi e Yu (2008)	2 anos	agregada	diário	*Evolutionary Neural Network*
Aksoy, Ozturk e Sucky (2012)	Obtido através de especialistas	agregada	mensal	Redes Neurais e inferência *fuzzy*
Xia et al. (2012)	5 anos	agregada	mensal	*Extreme Learning Machine with adaptive metrics of inputs*

Fonte – Fonte: Autoria própria

Os trabalhos citados, no geral, são estudos sobre a previsão da demanda de forma agregada em períodos semanais ou mensais, tendo disponível dados históricos de mais de uma estação, com a utilização de técnicas diversificadas. Ressalta-se, ainda, que não foi encontrada nenhuma publicação que tratasse a predição diária de vendas, item a item, de uma loja varejista, pelo autor. Com isso, não é possível obter conclusões específicas de trabalhos já realizados, de forma a balizar as diretrizes do PG e comparação de métricas.

1.5 Estrutura do Texto

O presente trabalho está estruturado da seguinte maneira:

- **Capítulo 1**: Esse capítulo tem como objetivo contextualizar o problema aqui estudado, que envolve predição de vendas, e um levantamento dos estudos correlatos de modo a situar as pesquisas da área. Além disso, são apresentados os objetivos da realização deste PG.
- **Capítulo 2**: Esse capítulo apresenta os fundamentos teóricos envolvendo a predição de vendas e os conceitos de AM, os principais modelos a serem utilizados, as técnicas de validação e avaliação dos modelos desenvolvidos.
- **Capítulo 3**: Esse capítulo apresenta os detalhes do desafio estipulado pela competição DMC 2018 e a proposta de solução, que foi concebida através de classificação

binária.

- **Capítulo 4**: Esse capítulo apresenta as métricas de desempenho usadas nos experimentos, os resultados obtidos e uma breve análise comparativa entre os modelos.
- **Capítulo 5**: No capítulo final deste trabalho são apresentadas as conclusões e as recomendações de trabalhos futuros.

2 REFERENCIAL TEÓRICO

2.1 Introdução

Este capítulo apresenta as características da predição de vendas no setor de moda e a base teórica de AM. Na primeira seção é abordado a predição de vendas, iniciando-se pela importância de uma predição acurada. Em seguida, são apresentadas as principais características envolvendo a predição de vendas no segmento de vestuário, com foco naquelas que são diferentes em relação à predição de demais setores. Na segunda seção é apresentada a definição da área de AM e as terminologias, de forma a facilitar a leitura. Na seção seguinte são abordados os principais modelos matemáticos que serão utilizados na solução proposta: Regressão Logística, *Decision Tree* (DT), *Random Forest* (RF), *Multi Layer Perceptron* (MLP) e *Ensemble*. Na última seção são apresentados os métodos mais comuns utilizados para validar um modelo: validação *holdout* e validação cruzada.

2.2 Predição de vendas

Predição acurada pode levar ao aumento da satisfação do consumidor, redução do estoque, eficiência na produção, redução de produtos obsoletos, melhor gerenciamento logístico, melhor negociação com fornecedores, melhores informações para determinar promoções e preço (STOCK, 2001).

Nenni, Giustiniano e Pirolo (2013) explicam que a indústria de moda sofreu uma fase de transição nos últimos 20 anos, elencando os principais fatores para essa mudança:

- Necessidade de reduzir custos em busca de aumentar a vantagem competitiva fez com que houvesse mudanças dos locais de produção devido ao menor custo da mão-de-obra.
- Globalização contribuiu para o aumento no tempo de aprovisionamento (*lead-time*) das matérias primas e produtos.
- O aumento da exigência dos consumidores exigiu das empresas atender às especificações tanto do produto quanto do serviço, intensificada ainda mais em tempo de mudanças dinâmicas no *design* e expectativas variadas de produtos;
- Tecnologia que fez aumentar drasticamente o conhecimento instantâneo de novas tendências e marcas.

Através das mudanças ocorridas, a predição de vendas é complexa por possuir peculiarida-

des:

- As companhias trabalham com produtos de curto ciclo de vida, desenhados para captar as tendências do momento, tipicamente de poucos meses ou até mesmo de semanas (NENNI; GIUSTINIANO; PIROLO, 2013);
- Poucos dados históricos disponíveis (um ciclo de vida curto, entre 1 e 2 meses) (FILDES; GOODWIN; LAWRENCE, 2006);
- Influência de sazonalidades ocasionadas por condições climáticas, eventos promocionais, fatores sociais e econômicos (MOSTARD; TEUNTER; KOSTER, 2011);
- Extensa variedade de produtos e alto número de SKU (*Stock Kepping Units*[1]) resulta em baixo volume de vendas por SKU e a demanda dentro de uma mesma linha de produtos pode variar significativamente. Com isso, mesmo que a demanda agregada possa ser predita com bom grau de acurácia, é difícil predizer como que a demanda será distribuída dentre os muitos produtos oferecidos (FILDES; GOODWIN; LAWRENCE, 2006; NENNI; GIUSTINIANO; PIROLO, 2013);
- Previsões de médio prazo (um horizonte de tempo de uma estação ou coleção) (FILDES; GOODWIN; LAWRENCE, 2006);
- Muitos itens novos a cada coleção lançada;
- A baixa previsibilidade por causa da alta volatilidade da demanda dificulta a previsão da demanda total dentro de um período, muito menos semana a semana ou item a item (CHRISTOPHER; PECK, 1997; NENNI; GIUSTINIANO; PIROLO, 2013);
- Alta impulsividade no momento da compra, o que leva muitos consumidores a decidir pela compra sem avaliar a necessidade (NENNI; GIUSTINIANO; PIROLO, 2013);

Nenni, Giustiniano e Pirolo (2013) concluem dizendo que, devido a essas características, as predições são comumente imprecisas e que gradualmente é aceito por quem trabalha na indústria e nas pesquisas de previsões que a demanda no setor de moda não pode ser prevista.

Segundo Ferreira, Lee e Simchi-Levi (2016) um aspecto que vem sendo trabalhado conjuntamente com a predição de vendas na última década, principalmente na fase de liquidação da coleção de moda, é o sistema de preços dinâmicos, que visa minimizar ou até zerar o estoque residual. A maioria desses sistemas assumem que o ciclo de vida do produto é longo e que o estoque é infinito. Todavia, na área de moda, ambos devem ser levados em consideração: idade do produto e estoque.

Os métodos tradicionais, ainda bastante utilizados no cenário nacional, como observado

[1] O SKU é um código ou referência dada a cada item de acordo com sua apresentação - forma, cor, tamanho.

por Bessa (2018), são classificados em dois grupos: qualitativos e quantitativos.

Os métodos qualitativos são baseados nas opiniões de especialistas do setor, e devido ao aspecto subjetivo sofrem de viés, sendo de difícil padronização. Todavia, são métodos simples que não requerem grandes investimentos em *softwares* especializados.

Os métodos quantitativos utilizam técnicas matemáticas e estatísticas, onde se busca extrapolar tendências futuras através do comportamento no passado, sendo adequadas para demandas de alto volume sem mudanças abruptas. As técnicas mais utilizadas são: média móvel, suavização exponencial, decomposição clássica, ARIMA, Box e Jenkins model, modelos de regressão, etc.

2.3 Aprendizado de Máquina (AM)

2.3.1 Definição

É uma área de estudo especializada no desenvolvimento de sistemas que aprendem de forma automatizada a partir de dados (HENRIK; MARK, 2017).

A definição mais difundida e formal é dado por (MITCHELL, 1997): "Um programa aprende a partir da experiência (E) com relação a alguma classe de tarefas (T) e medida de desempenho (P), se seu desempenho na tarefa em T, medido por P, melhora com a experiência". Em essência, o sistema aprende se o mesmo consegue melhorar sua habilidade, em determinada tarefa, a medida que é alimentado por dados.

2.3.2 Terminologia

No desenvolvimento cronológico da área, diferentes terminologias foram criadas para os conceitos básicos. A seguir são apresentados os principais termos, suas definições e características.

- **Atributo ou característica**: Dado elementar que caracteriza um registro. Pode ser do tipo binário, numérico, categórico, temporal, textual, espacial. É chamado na literatura inglesa de *feature*.
- **Registro ou instância**: Conjunto de atributos referente a um determinado objeto ou evento, descritos por um vetor.

- **Classe, rótulo ou atributo-alvo**: Atributo que se deseja aprender, também conhecido como valor-alvo, *label* ou *target*.
- **Conjunto de dados**: É um conjunto de registros, que contém valores dos atributos que podem estar ou não associados com uma classe.
  - **Conjunto de treinamento**: Conjunto de dados que possui atributos de predição associados a um atributo-alvo (classe ou rótulo), utilizados para treinar o modelo.
  - **Conjunto de validação**: Utilizado para ajustes no modelo obtido pelo algoritmo de aprendizagem, sendo diferente dos dados utilizado na fase de treinamento.
  - **Conjunto de teste**: Conjunto de dados cujo o atributo-alvo não é fornecido ao classificador, utilizado para avaliar o modelo obtido, através de métricas previamente estabelecidas.

2.4 Modelos

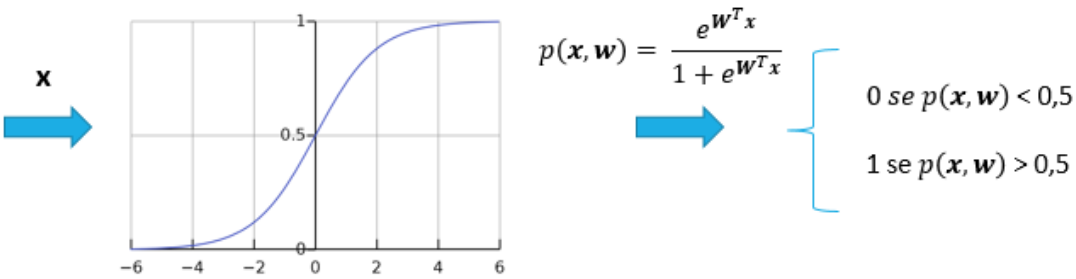
De forma geral os algoritmos de AM podem ser classificados em dois grupos: (*i*) supervisionado (um conjunto de registros com a saída conhecida é fornecida para treinamento); (*ii*) não supervisionado (a classe a que determinado registro pertence é desconhecida e o algoritmo busca determinar padrões através do agrupamento dos registros de acordo com as similaridades ou diferenças desses). Os modelos a serem utilizados para a previsão de vendas neste trabalho são do tipo supervisionado, especificamente: Regressão Logística, *Decision Tree* (DT), *Random Forest* (RF), *Multi Layer Perceptron* (MLP) e *Ensemble*. A seguir esses modelos são apresentados de forma resumida.

2.4.1 Regressão logística

Na regressão logística, o atributo alvo corresponde a um dado binário. O objetivo do modelo portanto é estimar a classe, dentre duas possíveis, a que um determinado registro pertence, de acordo com os atributos de predições disponíveis.

Esse modelo pode ser dividido em duas etapas, conforme Figura 1: a primeira determinada por uma função discriminante, que estima a probabilidade de um atributo-alvo pertencer a determinada classe (através do vetor de parâmetro W aprendido durante o treinamento) em função dos atributos de um registro (X); e a segunda determinada por uma função de ativação, que busca decidir de fato a qual classe um registro pertence (BISHOP, 2006).

Figura 1 – Regressão Logística



Fonte: Autoria própria

2.4.2 *Decision Tree* (DT)

As técnicas baseadas em árvores, que podem ser aplicados a problemas de regressão e classificação, consistem na subdivisão recursiva de um conjunto de valores com base no *dataset* de treino utilizado, resultando nas regras que é representado por um fluxograma em árvore.

A estrutura é formada por nós, onde cada registro é encaminhado para um determinado ramo (subconjunto) de acordo com os seus atributos e a condição gerada pelo treinamento do modelo, ou seja, os registros que satisfazem uma condição seguem para um nó enquanto os que não satisfazem seguem para um outro. O processo de divisão em ramos é repetido até que um critério de parada seja atingido.

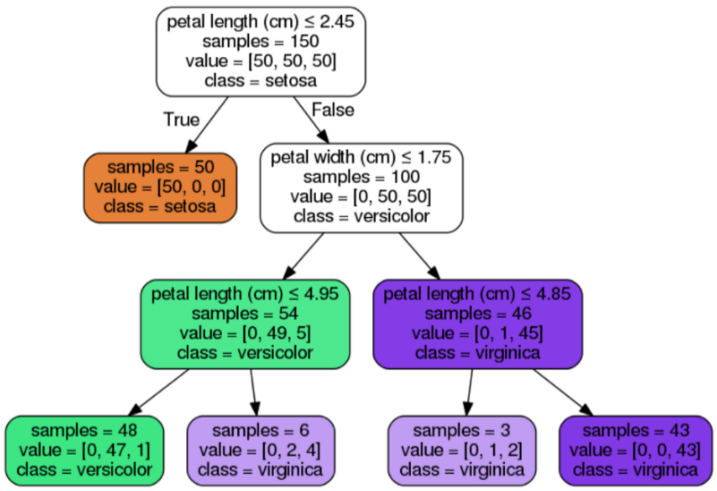
Na Figura 2 é possível visualizar a estrutura da árvore de decisão do problema clássico de classificação *Iris Flower*, utilizado em livros da área como introdução ao modelo DT. Três tipos de flores são classificados de acordo com o comprimento (*length*) e largura (*width*) das pétalas (*petal*). Após um registro percorrer as condições aprendidas durante o treinamento, a flor é classificada em setosa (cor laranja), versicolor (cor verde) e virginica (cor roxa).

2.4.3 *Multi Layer Perceptron* (MLP)

As redes neurais artificiais (ANN) são técnicas computacionais que desenvolvem um modelo matemático inspirado na estrutura neurológica dos seres humanos e que aprendem através das experiências (ROJAS, 1996).

A arquitetura tradicional de uma rede neural artificial é o MLP. Esta arquitetura pode ser vista na Figura 3 que consiste em uma rede *feedforward* com uma ou mais camadas

Figura 2 – Estrutura da árvore do problema *Iris Flower*

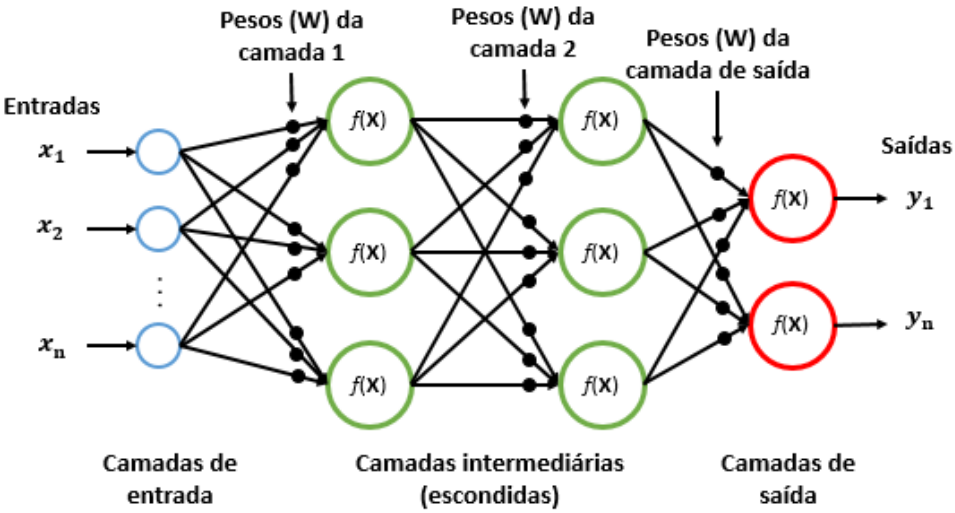


Fonte: Autoria própria

ocultas entre a camada de entrada e a de saída (BISHOP, 2006).

O algoritmo de treinamento usado comumente para determinar os pesos de uma rede MLP é o algoritmo *backpropagation*, que é trabalhado em duas etapas. Na primeira etapa, o sinal de entrada é propagado da camada de entrada até a camada de saída, calculando-se a previsão. Na segunda etapa, o erro da previsão realizada em relação ao valor real é propagado da camada de saída para a camada de entrada, ajustando os pesos de modo a minimizar o erro. Esse processo é repetido iterativamente sobre todo o conjunto de treinamento até que se atinja o ponto que minimize a função custo estipulada para o treinamento (BISHOP, 2006).

Figura 3 – Arquitetura de uma rede MLP



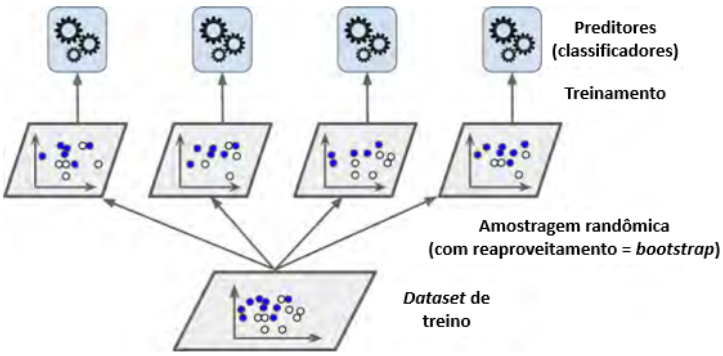
Fonte: Autoria própria

2.4.4 *Ensemble Learning*

Chamada na literatura como *ensemble*, é utilizado normalmente na parte final de um projeto, quando bons preditores já estão disponíveis. O método combina a previsão de diferentes modelos de AM em um preditor com melhor desempenho. Como relatado em Domingos (2012) e Géron (2017), esse método é amplamente utilizado em competições de análise de dados, sendo o mais famoso o da competição *Netflix Prize*. A competição lançada em 2006 pela empresa *Netflix*, desafiava as equipes a conseguir uma melhoria de 10% no sistema de recomendação de vídeos da empresa, sendo o prêmio de um milhão de dólares para a primeira equipe a conseguir. A competição é um marco em competições de *Data Science*, pois a meta só foi atingida após a junção de três equipes (*BellKor*, *BigChaos* e *Pragmatic Chaos*), que combinaram seus modelos através da técnica *ensemble*, vecendo a competição em junho de 2009. As principais técnicas são:

- *Bagging*: o método criado por Breiman (1996), busca reduzir a variância das previsões, combinando os resultados de vários classificadores fracos[2] em diferentes subamostras de um mesmo conjunto de dados. Para que se tenha êxito, é necessário que pequenas mudanças na base de treinamento provoque grandes mudanças na saída, como em árvores de decisão e redes neurais (GÉRON, 2017).

Figura 4 – Arquitetura de um *ensemble bagging*

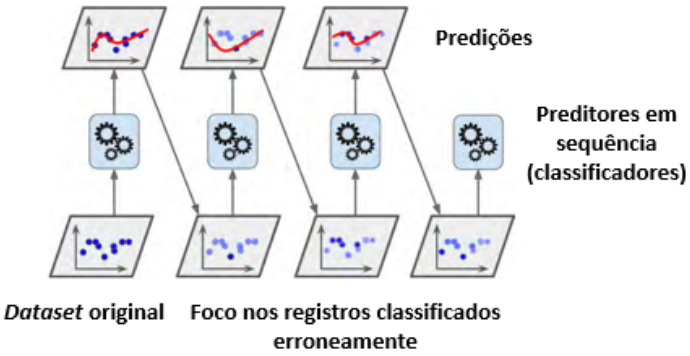


Fonte: Adaptado de Géron (2017)

- *Boosting*[3]: a ideia do método é construir preditores sequenciais, de modo que cada um melhore a predição feita pelo precursor, focando nas instâncias mal classificadas, ou seja, os modelos recorrem ao conjunto de dados que o modelo anterior classificou erroneamente. Os algoritmos mais populares baseados neste método são o *AdaBoosting* e o *Gradient Boosting* (GÉRON, 2017, p. 191).

[2] Classificador cujo desempenho é ligeiramente melhor do que adivinhação aleatória.
[3] Signifca aceleração, impulsão.

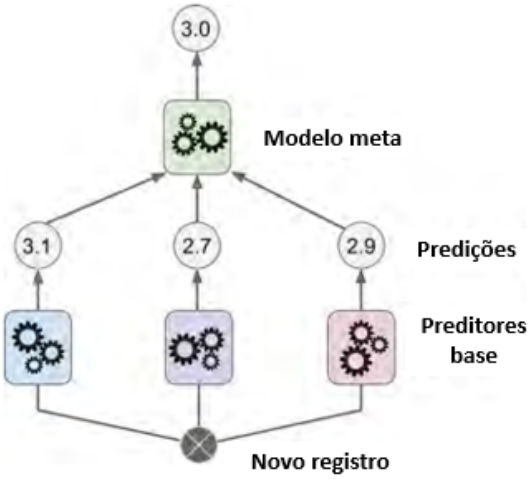
Figura 5 – Arquitetura de um *ensemble boosting*



Fonte: Adaptado de Géron (2017)

- *Stacking*: esse método é utilizado para combinar modelos de diferentes tipos em uma estrutura de camadas, conforme ilustrado na Figura 6. Na primeira camada são treinados diferentes modelos, chamados de modelos bases. Na segunda, um outro modelo, chamado de modelo meta, irá aprender através das previsões realizadas pelos modelos bases (VEEN; DAT; SEGNINI, 2015; FLENNERHAG, 2018).

Figura 6 – Arquitetura de um *ensemble stacking*



Fonte: Adaptado de Géron (2017)

2.4.5 *Random Forest* (RF)

Proposto por (HO, 1995) e melhorado por (BREIMAN, 2001), esse modelo ganhou popularidade durante as últimas décadas devido ao seu bom desempenho computacional, facilidade de uso e melhoria do desempenho em relação ao DT (RASCHKA, 2015). O modelo consiste em um *ensemble* que realiza *Bagging* sobre modelos de DT. A diferença

para o método *Bagging* é que nem todos os atributos são considerados na condição de cada árvore, aumentando a aleatoriedade e o desempenho do *ensemble* resultante.

Cada árvore é construída a partir do treinamento de um conjunto de amostras aleatória, tanto de registros como de atributos. A divisão de um nó busca o melhor resultado para as amostras aleatórias, e não sobre o conjunto de dados originais.

Figura 7 – Arquitetura de *Random Forest*

Fonte: Autoria própria

2.5 Validação

Após a adoção e treinamento do modelo de AM, é preciso verificar a eficiência da predição. Utilizar o próprio conjunto de treinamento não é ideal, uma vez que pode fornecer índices de desempenho que não refletem a capacidade de generalização[4] do sistema para novas instâncias. Problemas de *overfitting*[5] e *underfitting*[6], exemplificados na Figura 8, podem ocorrer, sendo necessário utilizar técnicas como validação cruzada e regularização para evitá-los.

Existem dois tipos de validação, segundo Kohavi (1995): *holdout* e validação cruzada.

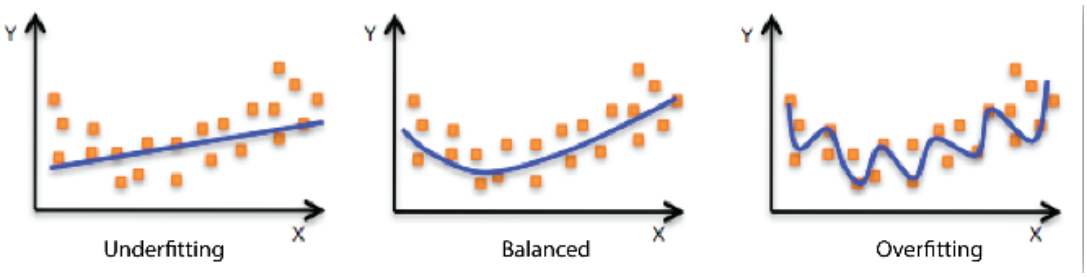
No método de *holdout*, o conjunto de dados é dividido em subconjunto de treinamento e subconjunto de teste, conforme a Figura 9. A escolha da proporção a ser dividido o conjunto de dados reflete numa decisão de *trade-off*: quanto maior o subconjunto de

[4] A habilidade de generalização se refere ao desempenho em classificar padrões que não foram utilizados durante o treinamento.

[5] Efeito de sobre-ajustamento, onde o modelo se ajusta bem ao conjunto de dados observados, mas é ineficaz para prever novos resultados.

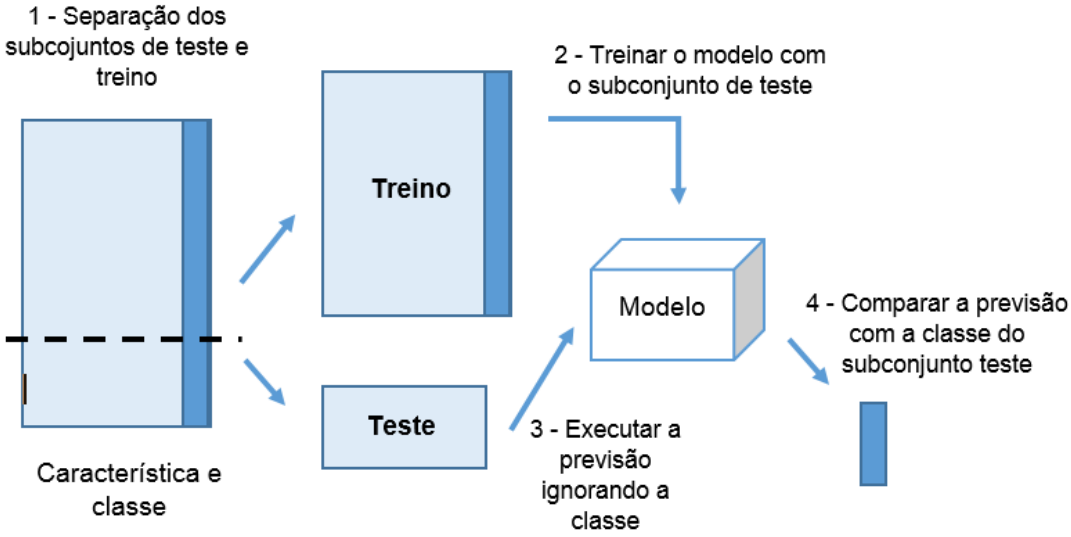
[6] Efeito de sub-ajustamento, onde o modelo não consegue capturar adequadamente o comportamento dos dados.

Figura 8 – Exemplo de modelos com *underfitting* e *overfitting*



Fonte: *Developer Guide - Amazon Machine Learning*

Figura 9 – Validação *holdout*

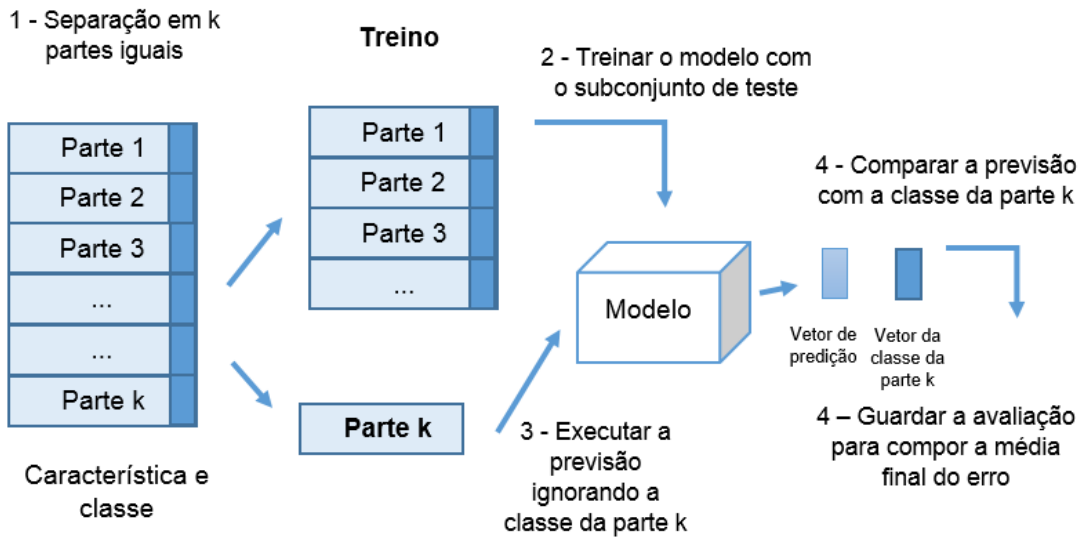


Fonte: Adaptado de Henrik e Mark (2017)

treinamento, maior é a capacidade de generalização. Por outro lado, quanto menor o subconjunto de testes, menor é a confiança na capacidade preditiva do modelo.

No método de validação cruzada, em uma rodada é realizada o particionamento do conjunto de dados, realizando o treinamento em $k-1$ subconjuntos e testando o modelo no subconjunto restante, conforme a Figura 10. Para reduzir a variabilidade, um determinado número de rodadas são executadas em diferentes partições e a validação resultante é combinada (por exemplo, pela média) para estimar a predição final do modelo. Uma melhor estimativa do erro é fornecida, porém demanda maior tempo de processamento computacional.

Figura 10 – Validação cruzada



Fonte: Adaptado de Henrik e Mark (2017)

2.6 Resumo

Através desse capítulo é possível concluir que a predição de vendas no setor *fashion* possui características específicas, o que torna a tarefa de predição desafiante. Nos últimos anos foram efetuados vários estudos sobre este problema, com a utilização de uma ampla variedade de técnicas de análise de dados, normalmente focadas em período mensais ou semanais, diferente da predição diária estipulada pela competição. No próximo capítulo é apresentado o problema a ser estudado, que é o desafio proposto no contexto da competição DMC 2018, a análise exploratória dos dados (EDA), o pré-processamento e a técnica de FE.

3 PROPOSTA

3.1 Introdução

Neste capítulo é apresentado o problema específico de predição de vendas no contexto da competição DMC 2018. Sendo assim, o capítulo divide-se em três seções, além desta introdução. Na primeira seção é descrito o problema-desafio estipulado pela competição, apresentando as principais características dos dados fornecidos através da Análise Exploratória dos Dados (EDA). Além de esclarecer o caso real que será estudado, visa-se agregar informações que sejam úteis para as etapas subsequentes. Na segunda seção é realizada a etapa de pré-processamento que visa formatar os dados para que seja possível aplicar os modelos, derivação e criação de novos atributos que contribuem para uma melhor predição. Na terceira seção são apresentadas as abordagens a serem testadas para a solução do problema, juntamente com a justificativa do porquê da adoção de cada uma.

3.2 Desafio

DMC é uma competição internacional organizada pela empresa alemã *Prudsys AG*, envolvendo problemas de *e-commerce*. No ano de 2018, um total de 193 equipes participantes de 148 instituições em 47 países, tiveram um intervalo de 6 semanas para submissão dos resultados, que consistiu na predição de vendas de produtos esportivos.

3.2.1 Cenário

Um varejista de produtos esportivos usa preços dinâmicos para controlar quando os itens serão vendidos em sua loja *online* e para isso uma boa predição de quando a venda se realizará é necessária para ajustar os preços mais conveniente. Todavia, o tempo para que um produto seja vendido depende não somente do preço, mas também de outros atributos como marca, tamanho e grupo do produto (PRUDSYS, 2018).

3.2.2 Tarefa

A tarefa estipulada para as equipes participantes consiste em utilizar os dados de vendas de quatro meses (de Outubro de 2017 até Janeiro de 2018) para desenvolver um modelo que preveja os dias em que os produtos se esgotarão no mês subsequente (Fevereiro de 2018).

3.2.3 Dados

Os dados fornecidos são reais, de uma loja *online* alemã, porém parte das informações foram codificadas de forma a tornar os dados anônimos. Por exemplo, não é especificado o nome do produto, mas sim o PID (*Product Identification*) do mesmo.

Foram disponibilizados três conjuntos de dados, em arquivos com extensão CSV, especificamente:

- *train.csv*: possui os eventos de venda durante o período de treinamento, tendo sido registrado um total de 135.117 vendas durante os quatro meses. Produtos não vendidos para determinado dia não são representados no conjunto de dados. Os cinco primeiros registros podem ser visualizados na Tabela 3.

Tabela 2 – Atributos disponíveis no *dataset train*

Atributo	**Descrição**	**Tipo de dado**	**Dados não nulos**
date	Data do registro de vendas	Número natural	135117
PID	Identificação do produto	*String*	135117
size	Tamanho do produto	*String*	134941
units	Unidades vendidas	Número natural	135117

Fonte – Fonte: Autoria própria

Tabela 3 – Cinco primeiros registros do conjunto de dados *train*

date	**PID**	***size***	*units*
2017-10-01	14393	2 (37-39)	1
2017-10-01	10069	36	2
2017-10-01	10069	35	1
2017-10-01	16221	L	1
2017-10-01	11317	L	1

Fonte – Fonte: Autoria própria

- *items.csv*: apresenta os atributos de todos os produtos que ocorrem durante o período de treinamento e de predição. Um produto é unicamente identificado pelos atributos *pid* e *size*. Os cinco primeiros registros podem ser visualizados na Tabela 5.

Tabela 4 – Atributos disponíveis no *dataset items*

Atributo	**Descrição**	**Tipo de dado**	**Dados não nulos**
PID	Identificação do produto	*String*	135117
size	Tamanho do produto	*String*	134941
color	Cor do produto	*String*	12824
brand	Marca do produto	*String*	12824
rpp	Preço de referência	Número real	12824
mainCategory	Categoria Principal	Número natural	12824
category	Categoria	Número natural	12824
subCategory	Sub Categoria	Número natural	11690
stock	Unidades do produto em estoque	Número natural	12824
releaseDate	Data em que foi disponibilizado	Formato AAAA-MM-DD	12824

Fonte – Fonte: Autoria própria

Tabela 5 – Cinco primeiros registros do conjunto de dados *items*

PID	***size***	*color*	*brand*	*rrp*	*mainCategory*	*category*	*subCategory*	*stock*	*releaseDate*
10000	XL (158-170)	gruen	Nike	25.33	1	7	25.0	1	2017-10-01
10001	L	schwarz	Jako	38.03	1	7	16.0	1	2017-10-01
10003	3 (35-38)	weiss	Jako	12.63	1	7	13.0	1	2017-10-01
10003	4 (39-42)	weiss	Jako	12.63	1	7	13.0	1	2017-10-01
10003	5 (43-46)	weiss	Jako	12.63	1	7	13.0	1	2017-10-01

Fonte – Fonte: Autoria própria

- *prices.csv*: apresenta os preços dos produtos durante o período de treinamento e de predição, ou seja, os cinco meses (Outubro de 2017 - Fevereiro de 2018). Os cinco primeiros registros podem ser visualizados na Tabela 6.

Tabela 6 – Cinco primeiros registros do conjunto de dados *prices*

PID	***size***	2017-10-01	2017-10-02	2017-10-03	2017-10-04	2017-10-05	2017-10-06
19671	39 1/3	133.31	133.31	133.31	133.31	133.31	133.31
19671	40	133.31	133.31	133.31	133.31	133.31	133.31
19671	41 1/3	133.31	133.31	133.31	133.31	133.31	133.31
19671	42	133.31	133.31	133.31	133.31	133.31	133.31
19671	42 2/3	133.31	133.31	133.31	133.31	133.31	133.31

Fonte – Fonte: Autoria própria

3.3 Análise exploratória dos dados

A análise visa identificar características particulares do problema em questão, buscando identificar informações importantes como valores faltantes, detecção de *outliers*, dispersão dos dados, dados sem padronização, dentre outros. Para isso são utilizados métodos visuais (gráficos de linhas, histogramas, mapas de calor) e estatísticos (percentual, média, dispersão, principalmente usando *boxplot*) a cerca dos dados. Nesse contexto, na continuação é descrita as análises dos atributos PID, *size*, *brand*, *stock*, *price* e *units*.

3.3.1 Atributo: PID e *size*

O atributo PID foi disponibilizado como forma de tornar anônima a identificação do artigo, que é feita juntamente com o atributo *size*, conforme especificado na competição. Por exemplo, na Tabela 5 os registros com o PID 10003 e tamanhos 3 (35-38), 4 (39-42), 5 (43-46), devem ser tratados como produtos diferentes.

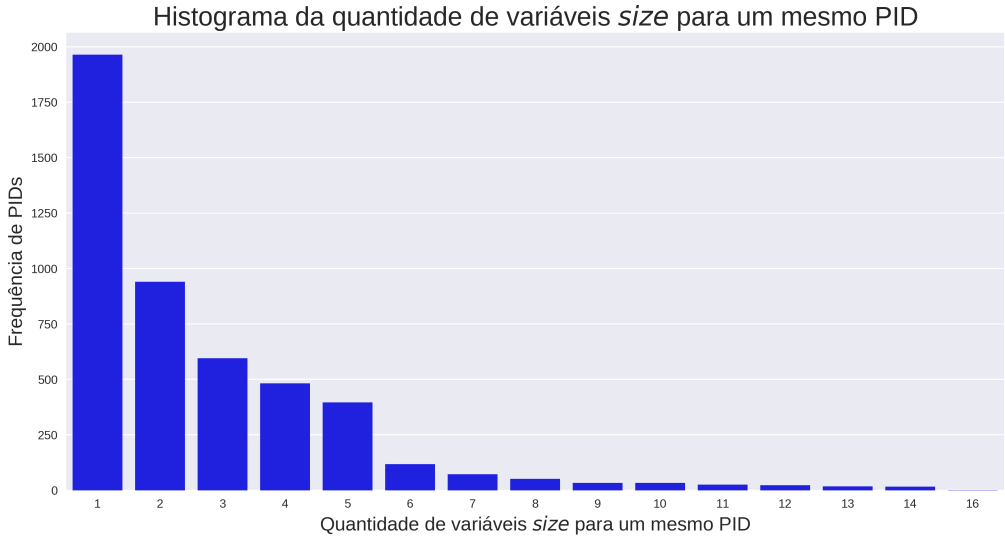
Observando os valores do atributo *size*, foi possível perceber que existem registros com informação similar, mas representados com códigos diferentes e que podem ser agrupados de forma a se ter melhores dados de entrada para o treinamento do modelo. Por exemplo: M (38/40), M (38/42), M (38-42), M (38 - 42), 38/40 (M / L), M (38-42), possuem a intenção de demonstrar uma unidade de medida similar, que podem ser agrupadas como M. Os dados que foram agregados são descritos no anexo B.

O histograma da Figura 11 mostra que quase dois mil PIDs possuem apenas um valor de tamanho e que a maioria dos produtos possuem menos do que cinco opções de tamanho.

Quase metade dos produtos possuem os tamanhos M, L, XL e S, conforme Tabela 7, existindo assim uma quantidade pequena de categorias de tamanho para vestuário superior (camisas, vestidos). Por outro lado, calçados e vestuário inferior (calças, bermudas) possuem uma maior quantidade de categorias de tamanho.

As informações mais relevantes desses atributos são sintetizadas: (*i*) a loja possui 12.824 produtos diferentes; (*ii*) existem 4.770 PIDs; (*iii*) os tamanhos M, L, XL, S concentram cerca de 40% dos produtos, conforme Tabela 7; (*iv*) 2.431 produtos (19% do total) foram vendidos apenas uma única vez durante o período de quatro meses para treinamento.

Figura 11 – Quantidade de variáveis *size* para um mesmo PID



Fonte: Autoria própria

Tabela 7 – *Ranking* dos cinco valores de *size* mais comum

Valor	Percentual de produtos
M	12.26
L	12.22
XL	9.04
S	8.36
44	7.78

Fonte – Fonte: Autoria própria

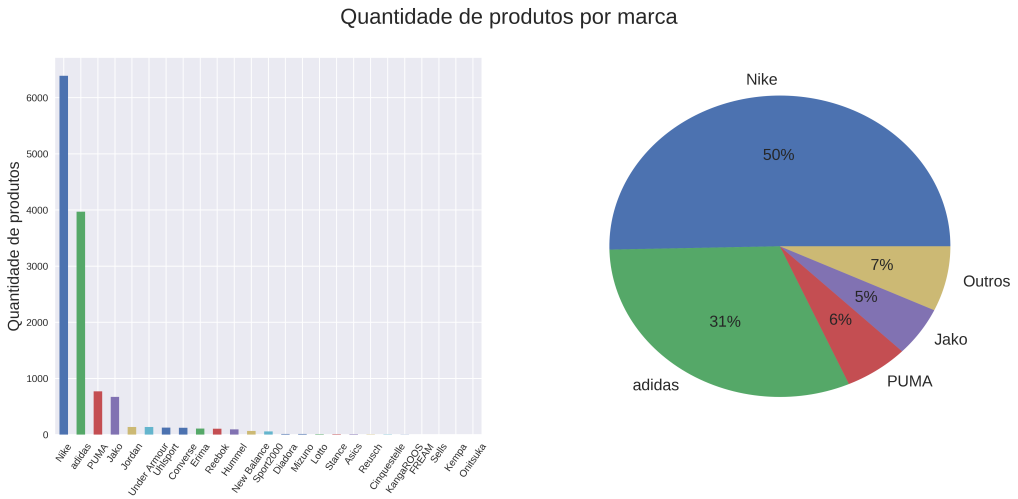
3.3.2 Atributo: *Brand*

As marcas dos produtos disponíveis se concentram em quatro grandes marcas, conforme a Figura 12, representando um total de 93% dos itens disponíveis. Essas marcas são: *Nike* (50%), *Adidas* (31%), *Puma* (6%) e *Jako* (5%). Outros 7% são produtos com marcas menos famosas. Ao total existem 25 marcas diferentes.

3.3.3 Atributo: *Stock*

A maior parte dos produtos possui poucas unidades em estoque, como é possível observar no histograma da Figura 13. Dos 12.824 produtos disponíveis 7.616 possui uma única unidade em estoque (59,39%), 2.110 possui duas unidades em estoque (16,45%) e 880 possui

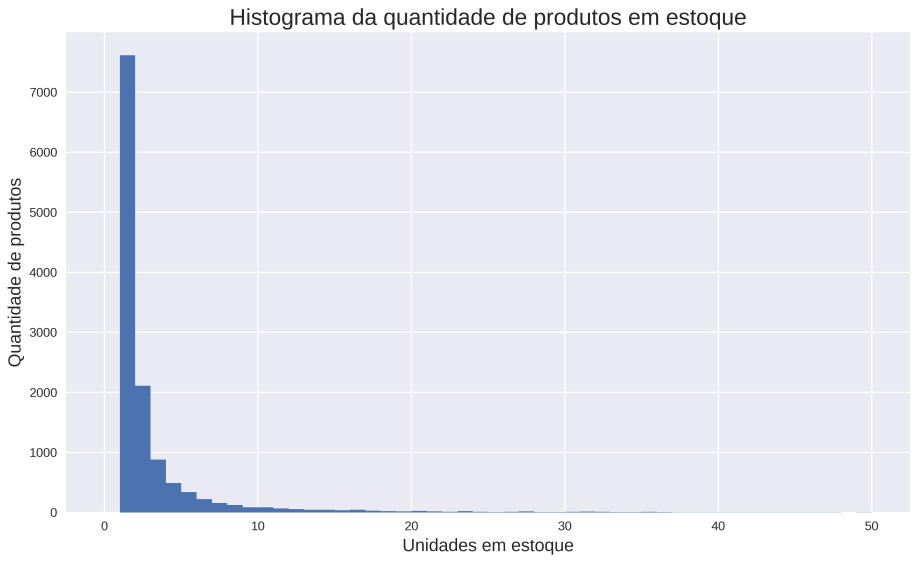
Figura 12 – Quantidade de produtos por marcas.



Fonte: Autoria própria

três unidades em estoque (6,86%). Essa informação fornece indicativo da complexidade da predição, pois cerca de 75% dos produtos possuem menos de duas unidades em estoque.

Figura 13 – Quantidade de produtos em estoque para produtos com menos de 50 unidades.



Fonte: Autoria própria

As informações relevantes encontradas para o atributo são: (*i*) a quantidade de produtos em estoque é de 45.302 unidades; (*ii*) cerca de 75% dos produtos, possui estoque com menos de 2 unidades; (*iii*) o produto com maior estoque possui 459 unidades.

3.3.4 Atributo: *Price*

Dos 12.824 produtos disponíveis, 1.866 produtos (14,55%) sofrem variação de preço. É possível observar as variações no mapa de calor da Figura 14, onde o eixo horizontal representa os dias, o eixo vertical os produtos e a escala de cor a variação percentual de preços. Logo que o produto é lançado ele geralmente possui um preço maior (cores avermelhadas) e a maioria sofre redução (cores azuladas) justamente no período de teste (Fevereiro 2018), devido a intenção de liquidar o estoque da estação.A região do gráfico na cor cinza claro mostra que nesse período o produto não estava disponível para venda.

Figura 14 – Mapa de calor das variações de preços durante o período de treino e teste.

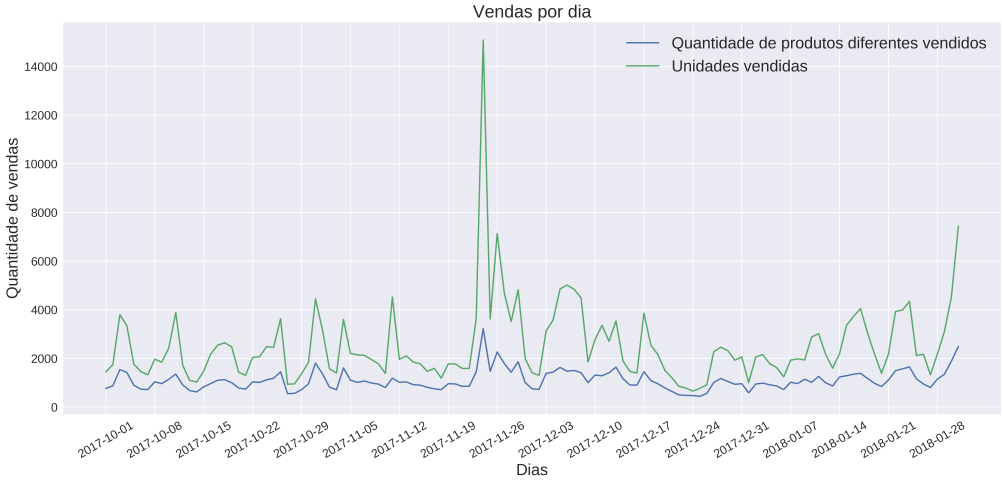
Fonte: Autoria própria

3.3.5 Atributo: *Units*

O atributo *units* fornece a quantidade de unidades vendidas em cada registro de venda. Em busca de compreender o comportamento de demanda da loja, a soma das vendas para

cada dia foram agrupadas e o gráfico da Figura 15 foi obtido. É possível observar um comportamento periódico, com dois períodos atípicos: 24 de Novembro (*BlackFriday*) e 31 de Janeiro (início da temporada de liquidação).

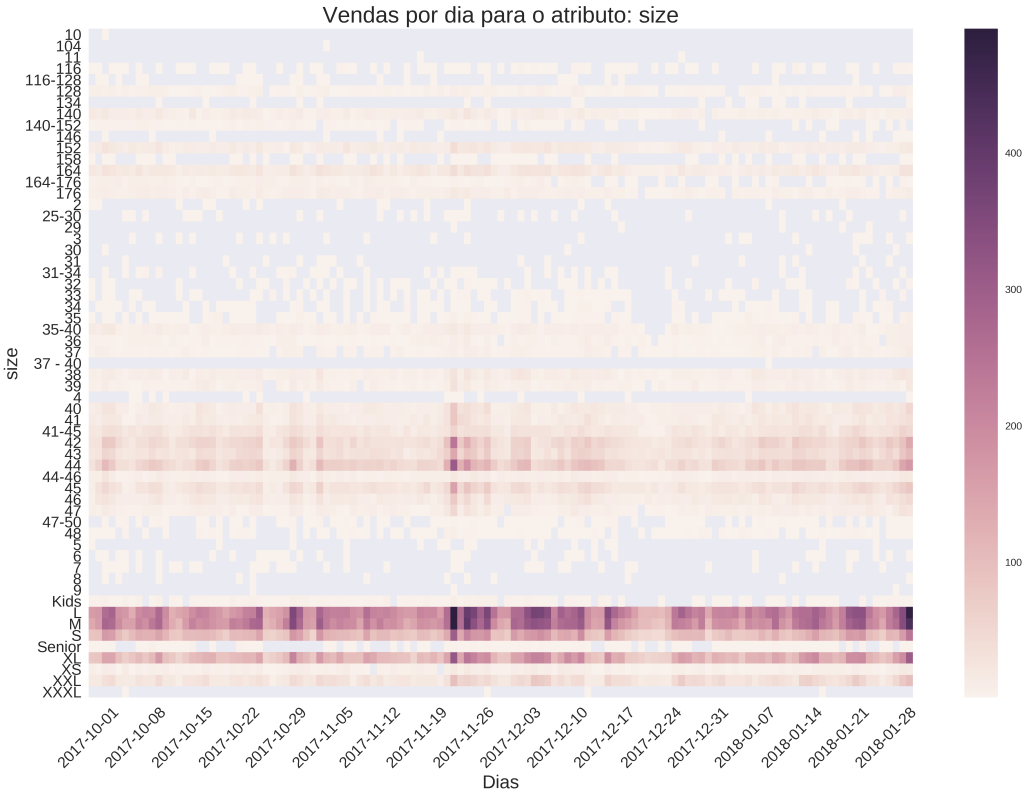
Figura 15 – Série temporal de vendas por dia da loja.



Fonte: Autoria própria

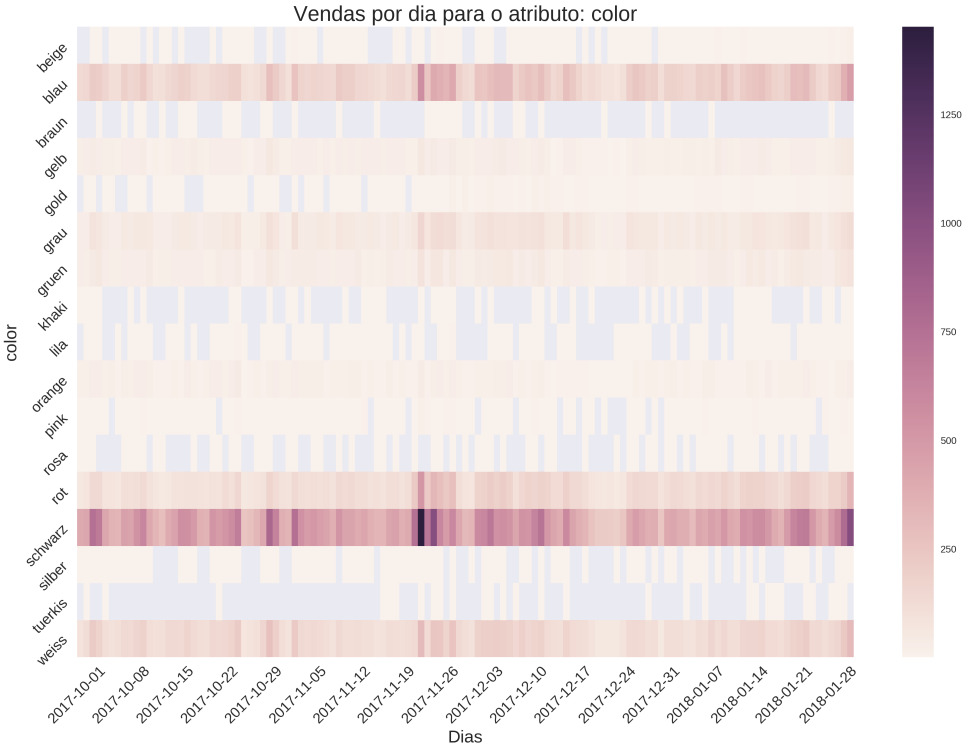
As Figuras 16, 17 e 18 mostram o mapa de calor de vendas para as categorias dos atributos *size*, *color* e *brand*. Na Figura 16 é possível observar que os tamanhos L, M, S, XL, 42, 43, 44 e 45, possuem maior volume de vendas, na Figura 17 que as cores com mais vendas são *schwarz* (preto), *rot* (vermelho), *blau* (azul) e *weiss* (branco) e na Figura 18 que as marcas mais vendidas são Nike e Adidas.

Figura 16 – Mapa de calor de vendas de acordo com as categorias do atributo *size*.



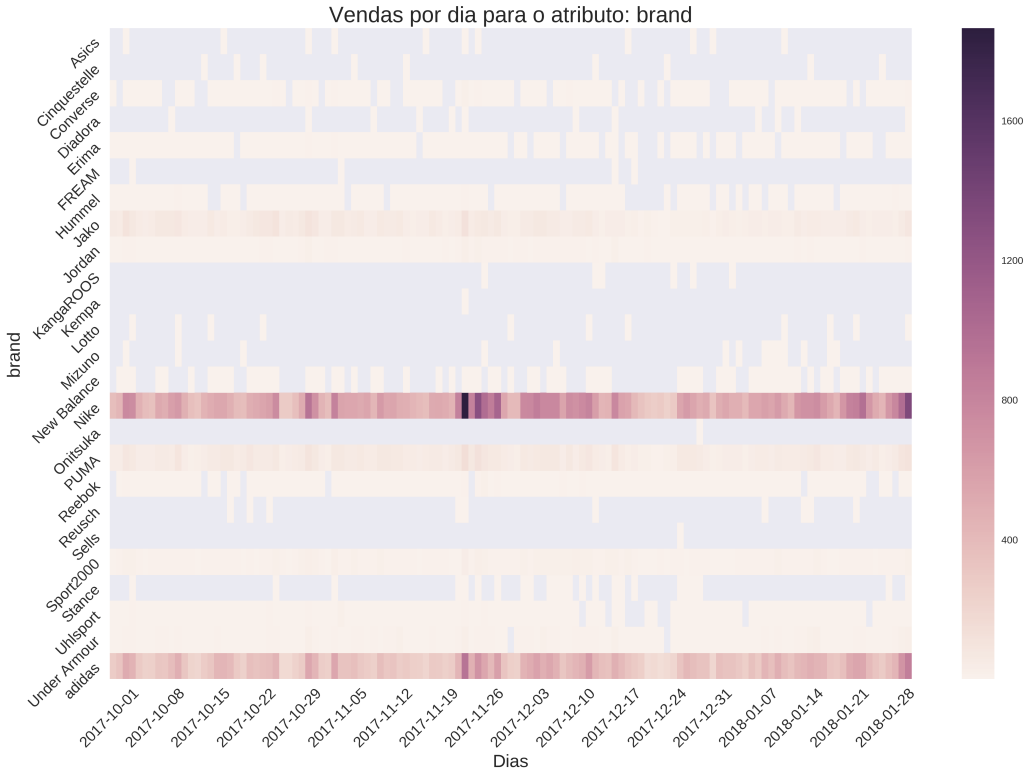
Fonte: Autoria própria

Figura 17 – Mapa de calor de vendas de acordo com as categorias do atributo *color*.



Fonte: Autoria própria

Figura 18 – Mapa de calor de vendas de acordo com as categorias do atributo *brand*.



Fonte: Autoria própria

3.4 Pré-Processamento

3.4.1 Valores faltantes

Os atributos com valores faltantes identificados não apresentam dificuldades de tratamento, pois existem em baixo volume e em situações específicas. Esses são detalhados a seguir:

- 1134 produtos possuem valores faltantes para o atributo *subcategory*, sendo que esses produtos pertencem a *mainCategory* “15” e a *category* “24”, “16” ou “33”. Por ser um atributo categórico, foi criada a categoria “0” para os valores faltantes do atributo *subcategory*.
- 9 produtos possuem valores faltantes para o atributo *size*, que são todos os produtos pertencentes a *subcategory* “27”. Como o tamanho é tratado como um atributo categórico, foi criado também a categoria “0” para os valores faltantes do atributo *size*.

3.4.2 Novos atributos

Novos atributos podem ser obtidos através da derivação dos atributos fornecidos, de forma a simplificar os modelos de predição, ou através da obtenção de dados externos que não foram fornecidos. A seguir são apresentados os atributos derivados ou criados.

- **Divisão do atributo data em ano, mês e dia**: Tal operação visa facilitar a identificação de padrões venda pelo modelo de acordo com o dia do mês, como o dia em que há um maior número de vendas devido ao recebimento de salário dos clientes. As informações de ano e mês não possuem relevância para a predição, uma vez que só foi disponibilizado o histórico de vendas de apenas uma coleção de moda (coleção de outono/inverno do hemisfério norte). Caso fosse disponibilizado dados dos anos anteriores, a informação do mês se tornaria importante por provavelmente guardar padrões de consumo similares ao ano corrente.
- **Criação do atributo "dia da semana"**: Através das análises exploratórias realizadas, como na Figura 15 é possível perceber que existem padrões de venda que se repetem a cada sete dias, sendo útil portanto um atributo que identifique os dias da semana. Com o auxílio do módulo *datetime* em *Python*, foi criado um atributo do dia da semana, sendo a segunda-feira o primeiro dia e o domingo o último dia.
- **Criação do atributo "feriado"**: As datas festivas podem alterar o comportamento de demanda dos clientes. Por isso foi criado um atributo que realiza a contagem de dias até determinado feriado (obtido por fonte externa e explicitados na Figura 8), de forma progressiva e começando 10 dias antes, conforme a Equação 3.1. Por exemplo, os registros do dia 20 de dezembro receberá no atributo Feriado o valor 5, pois esse dia antecede o Natal em 5 dias.
- **Derivação do atributo "preço"**: Foi criado um atributo que guarda a informação da diferença percentual de um preço em determinado dia em relação ao preço médio do produto durante o tempo em que esteve disponível. Por exemplo: o produto (16484, 41) possui, no dia 19 de Fevereiro, um preço 18,44% menor do que o preço médio.

$$Feriado(dia_{\ atual}) = 10 - (data_{\ feriado} - dia_{\ atual}) \tag{3.1}$$

3.4.3 Codificações

A maioria dos algoritmos de AM tem maior eficiência sobre valores numéricos, havendo assim necessidade de transformar características categóricas em características binárias. Na

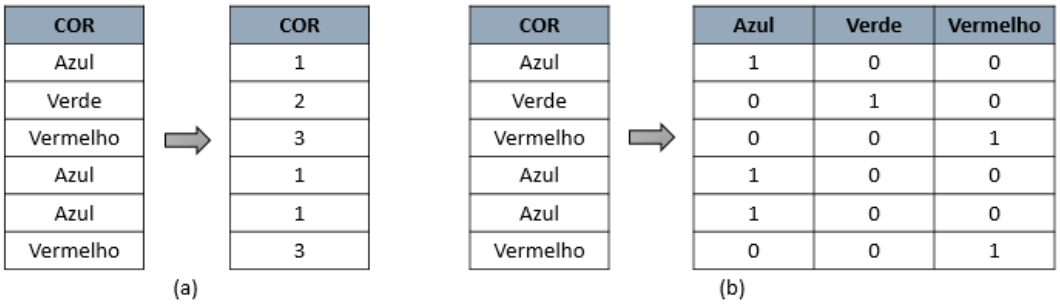
Tabela 8 – Feriados alemães considerados

date	*vacation*	*Day of week*
2017-10-03	German Unity Day	2
2017-10-31	Day of Reformation	2
2017-11-01	All Saint	3
2017-11-22	Repentance	3
2017-12-06	Nicholas Day	3
2017-12-25	Christmas	2
2018-01-01	New Year's Day	1
2018-01-06	Three Kings Day	6
2018-02-12	Carnival	1
2018-02-14	Valentine's Day	3

Fonte – Fonte: Autoria própria

Figura 19 (*a*) é realizada uma codificação, na qual cada categoria é transformada em um valor numérico. Esse tipo de codificação não é a mais adequada, pois os algoritmos assumem que dois valores próximos são mais similares do que dois valores distantes (GÉRON, 2017). Uma alternativa amplamente utilizada é a codificação *one-hot-encoding*, apresentado na Figura 19 (*b*), que cria um atributo binário para cada categoria, onde apenas um atributo codificado por instância pode ser igual a um (*hot*) enquanto os demais serão zero (*cold*).

Figura 19 – Exemplo da aplicação de codificação *one-hot-encoding*.



Fonte: Autoria própria

A técnica de *one-hot-encoding* é utilizada nos atributos: *size*, *color*, *brand*, *mainCategory*, *category*, *subCategory*, *weekday* e *day*.

3.4.4 Normalização

Os algoritmos de AM não tem boa eficiência sobre atributos numéricos com escalas muito diferentes, sendo necessário aplicar transformação que normalize os dados, ou seja, que passe a ter média zero e desvio padrão unitário. É particularmente útil em algoritmos de redes neurais ou baseados em distância, como o K-nn, acelerando o treinamento (HAN; PEI, 2012).

A normalização, realizada através da função *StandardScaler* do módulo *Scikit-Learn*, é aplicada aos atributos numéricos: *rrp*, *price* e *feriado*.

3.5 Modelagem do problema

Os dados históricos de vendas da loja revelaram que 41% dos produtos tiveram menos do que 5 registros de vendas entre os meses de Outubro de 2017 a Janeiro de 2018, e que 75% dos produtos possui menos de duas unidades em estoque para venda no mês Fevereiro de 2018. Por esse motivo, decidiu-se por modelar o problema como uma classificação binária (venda × não-venda), ao invés de uma classificação multiclasse que trás maior dificuldade de análise.

Para isso, o atributo *units* foi binarizado, onde o valor 1 (um) representa "venda". Porém, o *dataset* de treino disponibilizado possui apenas os casos de sucesso de venda, não sendo representado os produtos que não foram vendidos para determinado dia. Com isso, foi necessário acrescentar os dados de insucesso para que fosse possível treinar adequadamente o classificador. Com a extensão, o *dataset* passou de 135.117 para 1.644.692 registros, representando um desbalanceamento do atributo-alvo de 8,2%. Na Tabela 9 é apresentado um intervalo do *dataset* estendido, onde é possível observar o atributo *sold*, o qual será o atributo-alvo da classificação binária.

Tabela 9 – *Dataset* estendido

date	pid	size	units	sold
2017-10-01	10312	42	0	0
2017-10-01	10312	42 2/3	1	1
2017-10-01	10312	43 1/3	1	1
2017-10-01	10312	44	1	1
2017-10-01	10312	44 2/3	2	1
2017-10-01	10312	45 1/3	0	0
2017-10-01	10312	46	0	0
2017-10-01	10312	46 2/3	0	0
2017-10-01	10319	38	0	0
2017-10-01	10319	39	0	0
2017-10-01	10319	40	0	0
2017-10-01	10320	39/42	0	0
2017-10-01	10320	43/46	0	0
2017-10-01	10322	47	0	0
2017-10-01	10330	164	0	0
2017-10-01	10331	L (152-158)	1	1

Fonte – Fonte: Autoria própria

Portanto, para o treinamento do modelo de predição utiliza-se o *dataset* de treino estendido, onde estão disponíveis os dados históricos de vendas dos meses de Outubro de 2017 a Janeiro

de 2018 (154 dias), totalizando 1.644.692 registros "rotulados", ou seja, o atributo-alvo *sold* é conhecido ("venda" do produto em determinado dia, valor binário 1, "não-venda" do produto em determinado dia, valor binário 0, podendo ser melhor compreendido observando a Tabela 9).

Já a predição a ser realizada é feita para o mês de Fevereiro de 2018, sendo que o *dataset* de teste possui 359.072 registros (28 dias $\times$ 12.824 produtos), onde o atributo-alvo *sold* não é "rotulado", ou seja, o atributo-alvo é desconhecido.

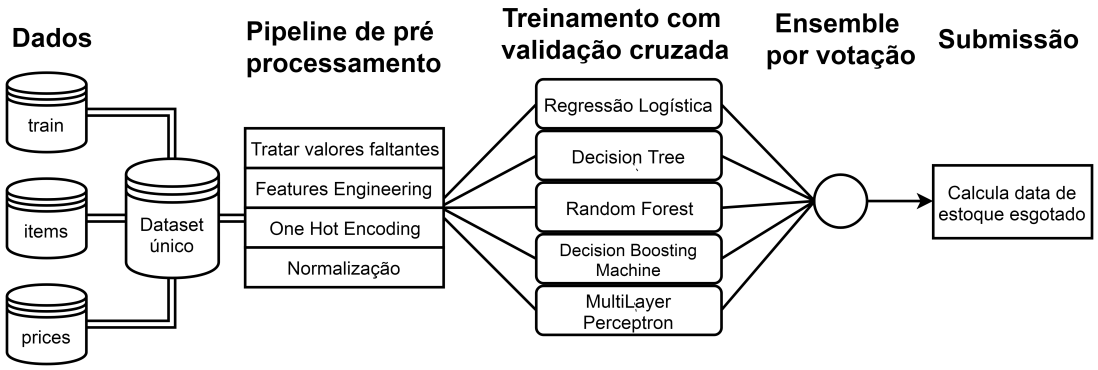
3.6 Fluxograma

Foi realizado o estudo de competições de *Data Science*, visando identificar as principais técnicas e abordagens utilizadas pelos vencedores, de modo a nortear o fluxograma deste trabalho do presente PG. As competições *DMC2016* e *AirBnB2016* são analisadas no apêndice A. Além disso, etapas de treinamento foram realizadas em competições da plataforma *Kaggle*[1] como forma de familiarizar-se com as ferramentas utilizadas em competições.

O fluxograma de trabalho foi determinado de acordo com as principais abordagens vencedoras e encontra-se na Figura 20. Inicialmente as bases de dados disponíveis são agrupadas em um conjunto de dados único e estendido, conforme explicado na seção 3.5. Em seguida uma *pipeline* de pré-processamento e FE é realizada, com a intenção de tratar os dados e adequá-los para os algoritmos de aprendizagem. Na terceira etapa é realizada a seleção dos hiperparâmetros e treinamento dos modelos. Por fim é realizada a predição e submissão do resultado.

[1] É uma plataforma de competições em que os competidores produzem modelos, sem restrição de técnicas, para problemas de predição ou classificação de empresas, onde as melhores soluções podem ser premiadas financeiramente. Em 2017 foi comprada pela *Google*, mesmo ano em que atingiu a marca de um milhão de membros.

Figura 20 – Fluxograma de trabalho.



Fonte: Autoria própria

3.7 Resumo

A competição DMC estipulou um desafio com grande complexidade por se tratar de uma previsão que não possui um consenso na literatura e cuja abordagem é inédita (predição diária de item a item). As dificuldades são ampliadas principalmente por dois fatores: (*i*) ausência de dados das coleções de moda das estações anteriores; (*ii*) grande quantidade de produtos com poucos registros de venda no conjunto de treinamento.

4 RESULTADOS

4.1 Introdução

Neste capítulo são apresentados os resultados obtidos pelos modelos de AM. De tal modo, inicia-se descrevendo o treinamento dos modelos de acordo com as abordagem propostas, discriminando os resultados de cada etapa através das métricas de validação. Por fim, é feita uma análise dos resultados de modo a compor o modelo mais promissor.

4.2 Detalhes de implementação

A linguagem *Python* foi escolhida pois é amplamente utilizada na área de computação científica, sendo a mais popular nas comunidades envolvendo competição de *Data Science*, como a plataforma *Kaggle*. Sua popularidade se dá devido ao alto nível da linguagem, natureza interativa e ao ecossistema de bibliotecas científicas.

As etapas de treinamento e testes foram realizadas utilizando *Scikit-Learn*, que é um módulo de AM na linguagem *Python*. Nele são disponibilizados algoritmos para treinamento de regressão, classificação, métricas de avaliação, além de ser projetado para interoperar com as bibliotecas de computação numérica *NumPy* e *SciPy* (BUITINCK et al., 2013).

O ambiente de programação utilizado foi o *Google Colaboratory*, que consiste em uma máquina virtual que executa um *Jupyter Notebook*, disponibilizado pela *Google* para ensino e pesquisas na área de AM. A escolha se deu por ser gratuito, possuir as principais bibliotecas de computação científica integrada sem necessidade de instalação e ter boas configurações de *hardware* (destaque para GPU Tesla K80 e 16 GB de RAM).

4.3 Métricas

Considerando que a previsão de vendas foi modelada como um problema de classificação binária, a avaliação utilizará quatro categorias básicas para a construção das métricas de desempenho.

- Verdadeiros Positivos (VP): número de registros classificados como verdadeiros, quando de fato são verdadeiros;
- Verdadeiros Negativos (VN): número de registros classificados como falso, quando de fato são falsos;
- Falsos Positivos (FP): número de registros classificados como verdadeiros, quando deveriam ser classificados como falsos;
- Falsos Negativos (FN): números de registros classificados como falsos, quando deveriam ser classificados como verdadeiros;

A seguir são apresentadas as definições das métricas de desempenho e as formulações matemáticas para o cálculo das mesmas.

- Acurácia (Ac): avalia a capacidade de classificar corretamente os dias que há venda de determinado produto, em essência, a porcentagem de acerto. Essa métrica sozinha não é adequada, pois ao lidar com dados desbalanceados poderá indicar uma alta taxa de acerto;
- Precisão (Pr): avalia o quanto das predições realizadas são de fato verdadeiras, representada matematicamente pela razão entre o número de acertos (VP) e a soma das predições classificadas como verdadeiras ($VP + FP$).
- *Recall* ou Sensibilidade (Se): avalia o quanto dos itens corretos foram encontrados, ou seja, a razão entre o número de acertos (VP) e a soma das predições que deveriam ser classificadas como verdadeiras ($VP + FN$).
- $F1$: é a média harmônica entre a precisão e a sensibilidade.

As relações que definem essas métricas são apresentadas nas Equações (4.1) – (4.4):

$$Ac = \frac{VP + VN}{VP + VN + FP + FN} \tag{4.1}$$

$$Pr = \frac{VP}{VP + FP} \tag{4.2}$$

$$Se = \frac{VP}{VP + FN} \tag{4.3}$$

$$F_1 = 2\frac{Se.Pr}{Se + Pr}. \tag{4.4}$$

A análise ROC[1] é um método visual para avaliar o desempenho de diferentes modelos de classificação. A curva representa a relação bidimensional entre:

- Taxa de verdadeiros positivos (eixo das ordenadas), que representa a proporção de

[1] Abreviação de *Receiver Operating Characteristics*, refletindo a utilização original da técnica em aplicações de processamento de sinal.

registros positivos que são corretamente classificados como positivos.

$$TVP = \frac{VP}{VP + FN} \tag{4.5}$$

- Taxa de falsos positivos (eixo abscissas), que representa a proporção de registros negativos que são classificados erroneamente como positivos.

$$TFP = \frac{FP}{FP + VN} \tag{4.6}$$

Cada ponto na curva ROC representa um par de sensibilidade/especificidade correspondente a um limiar de decisão. Quanto mais próximo do canto superior esquerdo, melhor é o desempenho do classificador, pois significa que o modelo possui uma alta proporção de acertos na classificação de verdadeiros positivos e uma baixa proporção de falsos classificados erroneamente. A linha diagonal representa um modelo de predição aleatória.

Existe ainda uma métrica que reduz a curva ROC a um valor numérico, útil para comparar classificadores, que é a área sob a curva (AUC), sendo um dos índices mais usados para sintetizar a qualidade da curva.

4.4 Treinamento

Em essência, um modelo de AM é uma função matemática com uma quantidade de parâmetros que serão aprendidos a partir dos dados, através do processo de treinamento. Existem também os hiperparâmetros, que são fixados antes do treinamento e são responsáveis por adicionar ao modelo propriedades como complexidade ou capacidade de aprendizado. Como forma de selecionar a melhor configuração para cada modelo, segundo Becker (2013), é possível utilizar técnicas como *Grid Search*, *Random Search* e abordagens baseadas em otimização por gradiente.

A técnica adotada neste trabalho é o *Grid Search*, que é tido como o mais simples e o mais utilizado quando o número de hiperparâmetros é pequeno (BECKER, 2013). Esse recurso busca os melhores hiperparâmetros, dentro dos especificados pelo usuário, através do treinamento e avaliação de todas as combinações possíveis. A Tabela 10 apresenta os hiperparâmetros testados, sendo que os em negrito foram os melhores obtidos.

Tabela 10 – Melhores hiperparâmetros obtidos pelo *Grid Search*.

Modelo	**Hiperparâmetros**
DT	*max_features*: 30, 50, **70**; *max_depth*: 3, 5, **10**; *min_samples_split*: **2**, 4; *min_samples_leaf*: 1, **3**;
RF	*n_estimators*: 10, **30**, 60; *max_features*: 10, 25, **50**; *max_depth*: **5**, 10; *min_samples_split*: 2, **5**; *min_samples_leaf*: **1**, 3;
GBM	*n_estimators*: 100, **150**; *max_features*: 10, **40**; *max_depth*: **3**, 7; *min_samples_split*: 3, **5**; *min_samples_leaf*: **1**, 2, 4;
MLP	*hidden_layer_sizes*: 10, 20, (5, 5), **(10, 10)** *alpha*: 0.01, 0.1, 0.5, **2**, 5

Fonte – Fonte: Autoria própria

4.5 Resultados

Na Tabela 11 é possível visualizar os resultados obtidos na classificação binária entre venda e não-venda, para cada modelo treinado com os hiperparâmetros obtidos com o *Grid Search*. A acurácia possui valores próximos de $0,9$ devido aos dados de treinamento serem desbalanceados (cerca de 10% dos registros são classificados como venda). Os modelos de Regressão Logística e *Gradient Boosting Machine* (GBM) apresentaram melhor precisão, porém com pior *recall*. Em contrapartida, o modelo MLP apresenta uma menor precisão, mas com o melhor resultado de *recall*.

Observa-se contudo que essas métricas possuem valores baixos em relação às predições usuais, provavelmente devido à presença de muitos produtos com nenhum ou com poucos registros de vendas.

Tabela 11 – Resultado do treinamento com validação cruzada.

Modelo	**Acurácia**	**Precisão**	**Recall**	**F1**
Regressão Logística	$0,91516$	$0,59401$	$0,21313$	$0,31370$
DT	$0,88483$	$0,29290$	$0,10576$	$0,18359$
RF	$0,89025$	$0,54352$	$0,16547$	$0,23456$
GBM	$0,91443$	$0,60373$	$0,18354$	$0,28150$
MLP	$0,91132$	$0,52797$	$0,23851$	$0,32858$

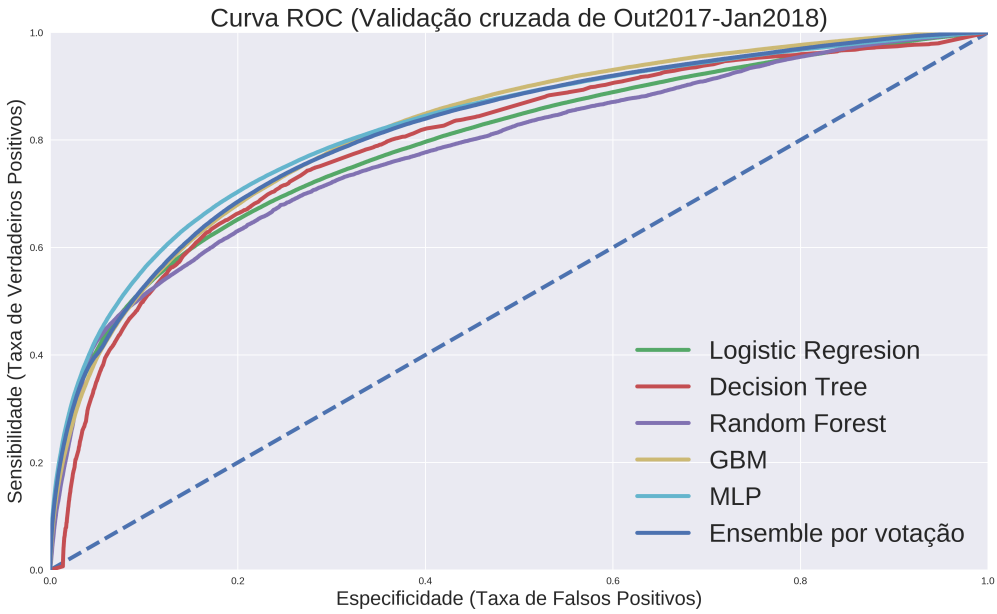
Fonte – Fonte: Autoria própria

Os modelos são comparados dentre duas situações de avaliação: (*i*) Validação cruzada, usando os 4 meses de treinamento; (*ii*) *holdout*, usando 3 meses para treinamento (Outubro de 2017 a Dezembro 2017) e um mês para teste (Janeiro de 2018).

Inicialmente os modelos são avaliados através da validação cruzada, que tem como vantagem a utilização integral de todos os dados de treinamento, mas como desvantagem a avaliação do desempenho sobre os dados utilizados no treinamento do modelo.

A Tabela 12, apresenta os valores AUC para os modelos testados com validação cruzada, com 4 partições, dos 4 meses de treinamento. Os modelos MLP e GBM apresentam o melhor desempenho, $0,823$ e $0,820$, respectivamente.

Figura 21 – Curva ROC com validação cruzada



Fonte: Autoria própria

Tabela 12 – Resultado do treinamento com validação cruzada usando como treinamento os meses de Outubro de 2017 a Janeiro de 2018.

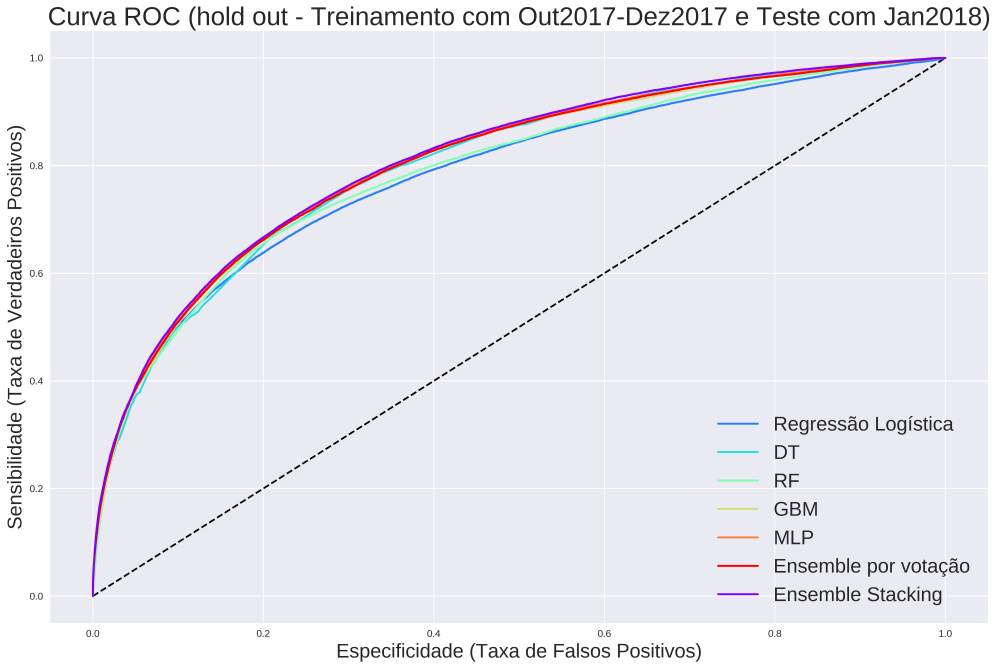
Modelo	**AUC**
Regressão Logística	$0,793$
DT	$0,795$
RF	$0,781$
GBM	$0,820$
MLP	$0,823$
ensemble por votação	$0,817$

Fonte – Fonte: Autoria própria

A segunda avaliação é realizada através do método *holdout*, que possui como vantagem a avaliação do desempenho sobre dados não utilizados no processo de aprendizado, mas como desvantagem a menor quantidade de dados disponíveis para treinamento.

A Tabela 13 apresenta os valores AUC para os modelos testados com validação *holdout*, usando 3 meses de treinamento (Outubro de 2017 - Dezembro de 2017) e um mês para teste (Janeiro de 2018). Os modelos MLP e GBM apresentam o melhor desempenho individual, ambos com $0,805$. O *ensemble* por votação apresenta uma pequena melhora, para $0,806$, sendo que o *ensemble stacking*, usando um classificador *Boosting* como modelo meta, apresenta o melhor desempenho alcançando $0,812$.

Figura 22 – Curva ROC com validação *holdout*



Fonte: Autoria própria

Os modelos treinados usando menos dados de treinamento (avaliação por *holdout*) apresentam um rendimento inferior aos modelos treinados usando todos os dados (avaliação por validação cruzada). Além disso, é possível perceber na curva ROC da avaliação por *holdout*, Figura 22, que as previsões possuem mais similaridade em todos os pontos de *limiar* (cada ponto na curva ROC representa um par de sensibilidade/especificidade correspondente a um limiar de decisão), do que na curva ROC da avaliação por validação cruzada, Figura 21.

4.6 Resumo

Nesse capítulo foram detalhados os aspectos da implementação, como os principais *softwares* utilizados. Em seguida foram treinados os modelos de Regressão Logística, DT, RF, GBM

Tabela 13 – Resultado do treinamento com validação *holdout* usando como treinamento os meses de Outubro de 2017 a Dezembro de 2017 e como teste o mês Janeiro de 2018.

Modelo	**AUC**
Regressão Logística	$0,786$
DT	$0,802$
RF	$0,791$
GBM	$0,805$
MLP	$0,805$
ensemble por votação	$0,806$
ensemble* por *stacking	$0,812$

Fonte – Fonte: Autoria própria

e MLP, e avaliados segundo as métricas de acurácia, precisão, *recall*, *F1* através da técnica de *Grid Search*. Conhecendo as melhores configurações de hiperparâmetros, os modelos foram treinados e avaliados utilizando validação cruzada e validação *holdout*. É possível observar que a utilização de uma quantidade maior de dados, possibilitada na validação cruzada, fornece métricas de desempenho AUC superiores às métricas obtidas quando se utiliza *holdout* com o uso de um mês a menos para treinamento. A técnica de *ensemble* mais promissora é a que usa *stacking*, porém é preciso utilizar uma quantidade de dados de treinamento reduzido para treinamento dos modelos base, reservando uma parte para treinamento do modelo meta.

5 CONCLUSÕES E PROJETOS FUTUROS

5.1 Conclusões

O objetivo principal deste trabalho foi desenvolver um modelo de previsão de vendas diária de produtos de uma loja de artigos esportivos usando modelos de AM. A abordagem proposta é baseada nas principais técnicas utilizadas em competições de *Data Science*, sendo o fluxo de trabalho balizado de acordo com soluções vencedores de competições similares, como visto no Apêndice A. O problema foi modelado como uma classificação binária, entre "venda" e "não-venda", motivado pelas características do conjunto de dados disponíveis, já que 75% dos produtos possui menos de duas unidades em estoque e é necessário prever a venda diária durante um mês completo. Os modelos de Regressão Logística, DT, RF, GBM, MLP e *ensemble*, foram testados utilizando a técnica de seleção de hiperparâmetros *Grid Search*, e posteriormente treinados e avaliados utilizando os métodos de validação cruzada e *holdout*. Como resultado, os modelos GBM e MLP apresentaram o melhor desempenho, obtendo métricas AUC de $0,820$ e $0,823$, respectivamente, ao serem treinados utilizando os registros de vendas de Outubro de 2017 a Janeiro de 2018 e avaliação por validação cruzada. Por não ter sido encontrado pelo autor nenhuma publicação que trata sobre predição de vendas diária, item a item, espera-se contribuir fornecendo as primeiras métricas de desempenho para comparação de trabalhos futuros. Finalmente, obteve-se a 38ª colocação com a solução apresentada nesse trabalho dentre 193 equipes participantes.

5.2 Temas a serem pesquisados

Uma abordagem observada em anos anteriores e que poderia trazer melhorias para as predições é a utilização de dados externos, o que requer uma maior quantidade de tempo coletando e tratando os dados. Na competição desse ano, essa tarefa foi dificultada devido ao não fornecimento do nome da loja online e dos produtos. Recomenda-se para trabalhos futuros: inclusão de dados de pesquisa em mecanismos de buscas online, informações de redes sociais e fatores socioeconômicos.

Uma segunda abordagem sugerida é a inclusão de dados históricos de outras lojas de material esportivo, principalmente no mês de fevereiro (intervalo de teste da predição), que por ser um mês de liquidação tende a ter um padrão de consumo diferente dos quatro meses anteriores.

REFERÊNCIAS

AKSOY, Asli; OZTURK, Nursel; SUCKY, Eric. **A decision support system for demand forecasting in the clothing industry**. [s.n.], 2012. v. 24. 221-236 p. Disponível em: <https://doi.org/10.1108/09556221211232829>. Citado na página 12.

AU, Kin-Fan; CHOI, Tsan-Ming; YU, Yong. **Fashion retail forecasting by evolutionary neural networks**. [S.l.: s.n.], 2008. v. 114. 615-630 p. Citado na página 16.

BATISTA, Valquiria Constancio. Dissertação de mestrado. **Previsão de vendas : um estudo no varejo de vestuário**. Universidade Federal do Rio Grande do Sul, 2017. Disponível em: <http://hdl.handle.net/10183/173191>. Citado na página 13.

BECKER, Cassiano. Dissertação de mestrado. **Hyper-parameter optimization for manifold regularization learning**. Universidade Estadual de Campinas, 2013. Disponível em: <http://repositorio.unicamp.br/jspui/handle/REPOSIP/259699>. Citado na página 46.

BESSA, Adriano Bezerra. Dissertação de mestrado. **Previsão de vendas no varejo de moda com modelos de redes neurais**. Fundação Getúlio Vargas, 2018. Disponível em: <http://hdl.handle.net/10438/23890>. Citado 2 vezes nas páginas 12 e 20.

BISHOP, Christopher M. **Pattern Recognition and Machine Learning (Information Science and Statistics)**. Berlin, Heidelberg: Springer-Verlag, 2006. ISBN 0387310738. Citado 2 vezes nas páginas 21 e 23.

BREIMAN, Leo. **Bagging Predictors**. Hingham, MA, USA: Kluwer Academic Publishers, 1996. v. 24. 123–140 p. ISSN 0885-6125. Disponível em: <http://dx.doi.org/10.1023/A:1018054314350>. Citado na página 24.

BREIMAN, Leo. **Random Forests**. Hingham, MA, USA: Kluwer Academic Publishers, 2001. v. 45. 5–32 p. ISSN 0885-6125. Disponível em: <https://doi.org/10.1023/A:1010933404324>. Citado na página 25.

BUITINCK, Lars et al. **API design for machine learning software: experiences from the scikit-learn project**. [S.l.: s.n.], 2013. 108–122 p. Citado na página 44.

CARO, Felipe; GALLIEN, Jérémie. **Clearance Pricing Optimization for a Fast-Fashion Retailer**. Institute for Operations Research and the Management Sciences (INFORMS), Linthicum, Maryland, USA: INFORMS, 2012. v. 60. 1404–1422 p. ISSN 0030-364X. Disponível em: <http://dx.doi.org/10.1287/opre.1120.1102>. Citado 2 vezes nas páginas 13 e 15.

CHRISTOPHER, M.; PECK, H. **Managing logistics in fashion markets**. [S.l.: s.n.], 1997. v. 8. 63-74 p. Citado na página 19.

DOMINGOS, Pedro. **A Few Useful Things to Know About Machine Learning**. New York, NY, USA: ACM, 2012. v. 55. 78–87 p. ISSN 0001-0782. Disponível em: <http://doi.acm.org/10.1145/2347736.2347755>. Citado na página 24.

EKAMI, K.H.O; MING, W.C.; BRITO, J.F.S. **Ferramentas para a indústria de vestuário: previsão de demanda**. [S.l.: s.n.], 2014. Citado na página 15.

FERREIRA, Kris Johnson; LEE, Bin Hong Alex; SIMCHI-LEVI, David. **Analytics for an Online Retailer: Demand Forecasting and Price Optimization**. [S.l.: s.n.], 2016. v. 18. 69-88 p. Citado 3 vezes nas páginas 13, 15 e 19.

FILDES, Robert; GOODWIN, Paul; LAWRENCE, Michael. **The Design Features of Forecasting Support Systems and Their Effectiveness**. Amsterdam, The Netherlands, The Netherlands: Elsevier Science Publishers B. V., 2006. v. 42. 351–361 p. ISSN 0167-9236. Disponível em: <http://dx.doi.org/10.1016/j.dss.2005.01.003>. Citado na página 19.

FLENNERHAG, Sebastian. **Introduction to Python Ensembles**. 2018. Disponível em: <https://www.dataquest.io/blog/introduction-to-ensembles/>. Citado na página 25.

FUMI, Andrea L et al. **Fourier Analysis for Demand Forecasting in a Fashion Company**. [S.l.: s.n.], 2013. Citado na página 15.

GARCIA, Eduardo; REIS, Leonardo Machado e Virgílio José Ferreira Filho Leticia. **Gestão De Estoques - Otimizando a Logística e a cadeia de Suprimentos**. [S.l.]: E-Papers Serviços Editoriais, 2006. Citado na página 12.

GÉRON, A. **Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems**. O'Reilly Media, 2017. ISBN 9781491962244. Disponível em: <https://books.google.com.br/books?id=bRpYDgAAQBAJ>. Citado 4 vezes nas páginas 14, 24, 25 e 40.

HAN, Micheline Kamber Jiawei; PEI, Jian. **Data Mining**. 3rd ed. ed. [S.l.]: Elsevier, 2012. ISBN 978-0-12-381479-1. Citado na página 40.

HENRIK, RICHARDS Joseph BRINK; MARK, FETHEROLF. **Real World Machine Learning**. 1. ed. [S.l.]: Manning Publications, 2017. Citado 4 vezes nas páginas 14, 20, 27 e 28.

HO, Tin Kam. **Random Decision Forests**. Washington, DC, USA: IEEE Computer Society, 1995. 278– p. (ICDAR '95). ISBN 0-8186-7128-9. Disponível em: <http://dl.acm.org/citation.cfm?id=844379.844681>. Citado na página 25.

KOHAVI, Ron. **A Study of Cross-validation and Bootstrap for Accuracy Estimation and Model Selection**. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1995. 1137–1143 p. (IJCAI'95). ISBN 1-55860-363-8. Disponível em: <http://dl.acm.org/citation.cfm?id=1643031.1643047>. Citado na página 26.

MITCHELL, Tom M. **Machine Learning**. [S.l.]: McGraw-Hill, 1997. ISBN 0-07-042807-7. Citado na página 20.

MOSTARD, Julien; TEUNTER, Ruud; KOSTER, Rene. **Forecasting demand for single-period products: A case study in the apparel industry**. [S.l.]: ELSEVIER SCIENCE BV, 2011. v. 211. 139–147 p. ISSN 0377-2217. Citado 2 vezes nas páginas 12 e 19.

NENNI, Maria Elena; GIUSTINIANO, Luca; PIROLO, Luca. **Demand Forecasting in the Fashion Industry: A Review**. [s.n.], 2013. v. 5. 37 p. Disponível em: <https://doi.org/10.5772/56840>. Citado 3 vezes nas páginas 13, 18 e 19.

PRUDSYS, Company. Data mining cup 2018. 2018. Disponível em: <https://www.data-mining-cup.com/dmc-2018/>. Citado na página 29.

RASCHKA, Sebastian. **Python Machine Learning**. [S.l.]: Packt Publishing, 2015. ISBN 1783555130, 9781783555130. Citado na página 25.

ROJAS, Raúl. **Neural Networks: A Systematic Introduction**. Berlin, Heidelberg: Springer-Verlag, 1996. ISBN 3-540-60505-3. Citado na página 22.

STOCK, James. **Strategic logistics management**. Boston: McGraw-Hill/Irwin, 2001. ISBN 0256136874. Citado na página 18.

TEAM, Kaggler. **Airbnb New User Bookings, Winner's Interview: 3rd place: Sandro Vega Pons**. 2016. Disponível em: <http://blog.kaggle.com/2016/03/07/airbnb-new-user-bookings-winners-interview-3rd-place-sandro-vega-pons/>. Citado na página 58.

THOMASSEY, Sébastien; FIORDALISO, Antonio. **A Hybrid Sales Forecasting System Based on Clustering and Decision Trees**. Amsterdam, The Netherlands, The Netherlands: Elsevier Science Publishers B. V., 2006. v. 42. 408–421 p. ISSN 0167-9236. Disponível em: <http://dx.doi.org/10.1016/j.dss.2005.01.008>. Citado na página 12.

VEEN, Hendrik Jacob van; DAT, Le Nguyen The; SEGNINI, Armando. **Kaggle Ensembling Guide**. 2015. Disponível em: <https://mlwave.com/kaggle-ensembling-guide/>. Citado na página 25.

XIA, Min et al. **Fashion Retailing Forecasting Based on Extreme Learning Machine with Adaptive Metrics of Inputs**. Amsterdam, The Netherlands, The Netherlands: Elsevier Science Publishers B. V., 2012. v. 36. 253–259 p. ISSN 0950-7051. Disponível em: <http://dx.doi.org/10.1016/j.knosys.2012.07.002>. Citado na página 16.

YANGI, Jile. **Winning Solution to 2016 Data Mining Cup**. 2016. Disponível em: <https://github.com/jlyang1990/Data_Mining_Cup_2016>. Citado 2 vezes nas páginas 56 e 57.

YU, Yong; HUI, Chi-Leung; CHOI, Tsan-Ming. **An Empirical Study of Intelligent Expert Systems on Forecasting of Fashion Color Trend**. Tarrytown, NY, USA: Pergamon Press, Inc., 2012. v. 39. 4383–4389 p. ISSN 0957-4174. Disponível em: <http://dx.doi.org/10.1016/j.eswa.2011.09.153>. Citado na página 12.

Apêndices

APÊNDICE A – COMPETIÇÕES

A.1 Data Mining Cup 2016

DMC é uma competição internacional organizada pela empresa alemã *Prudsys AG*. Em 2016, teve como desafio a predição da taxa de devolução de artigos de moda de uma distribuidora, tendo sido fornecida uma base de dados com $2,33$ milhões de registros e 14 atributos: pedido, data, artigo, cor, tamanho, grupo, quantidade, preço real e de referência, cupom, usuário, dispositivo, método de pagamento e a quantidade de devolução (atributo-alvo). A descrição e faixa de valores dos atributos são apresentados na Figura 23.

A equipe vencedora do evento foi composta por seis estudantes de estatística de doutorado da Universidade da Califórnia, dentre 120 equipes de 30 diferentes países. A melhor solução foi disponibilizada no *GitHub* (YANGI, 2016) e é analisada a seguir. Na descrição do projeto, a equipe destaca que o trabalho de FE foi o mais importante e crucial para obtenção de sucesso. Diferentes perspectivas foram utilizadas:

Figura 23 – Atributos do conjunto de dados do desafio DMC 2016

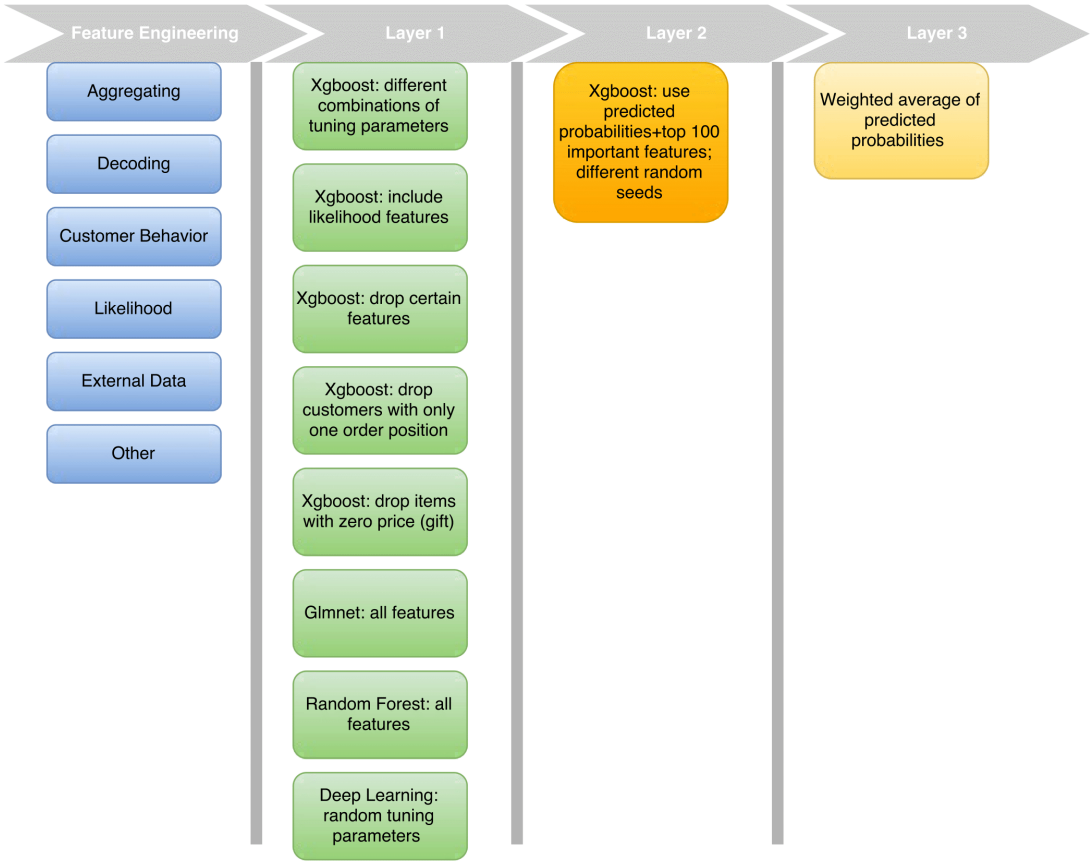
ATRIBUTOS - DMC 2016		
Nome	**Descrição**	**Faixa de valores**
ordem	Número da compra	Número natural precedido de "a"
data	Data da compra	Data no formato AAAA-MM-DD
artigo	Número do aritgo	Número natural precedido de "i"
cor	Código de cor do artigo	Número Natural
tamanho	Tamanho do artigo	Número ou string de letras
grupo	Grupo do produto	Número Natural
quantidade	Quantidade comprada	Número Natural incluindo 0
preço	Valor total da ordem	Número real positivo
rrp	preço recomendado	Número real positivo
voucher	número do cupom	Número natural precedido de "v0"
quant_voucher	Valor do cupom por ordem	Número real positivo
consumidor	Número do consumidor	Número natural precedido por "c"
dispositivo	Dispositivo utilizado	{1,2,3,4,5}

Fonte: Autoria própria

- Agregação: A partir do agrupamento de variáveis, foram aplicadas diferentes funções de agregações como mínimo, máximo, média, desvio, soma, moda, número de elementos, número de elementos únicos, de forma a gerar novas características expandindo a base de dados original.
- Decodificação: Busca pela compreensão de significados implícitos, de modo a simplificar características, homogeneizar unidades, classificar grupos com características semelhantes.
- Comportamento do consumidor: Análises identificaram que certos consumidores retornam 15 vezes mais peças. Classificou-se esses em três categorias (*Wardrobe, try it one, fitting roomer*) através da similaridade entre itens de um mesmo pedido. Quanto maior o número de itens comprados, maior a probabilidade de retorno.

A estratégia de modelagem adotada baseia-se em arquitetura multi-camadas, apresentada na Figura 24, cuja predição de várias bases de aprendizado são combinadas de modo a melhorar a eficácia da predição.

Figura 24 – Modelagem da equipe vencedora da competição DMC 2016



Fonte: (YANGI, 2016)

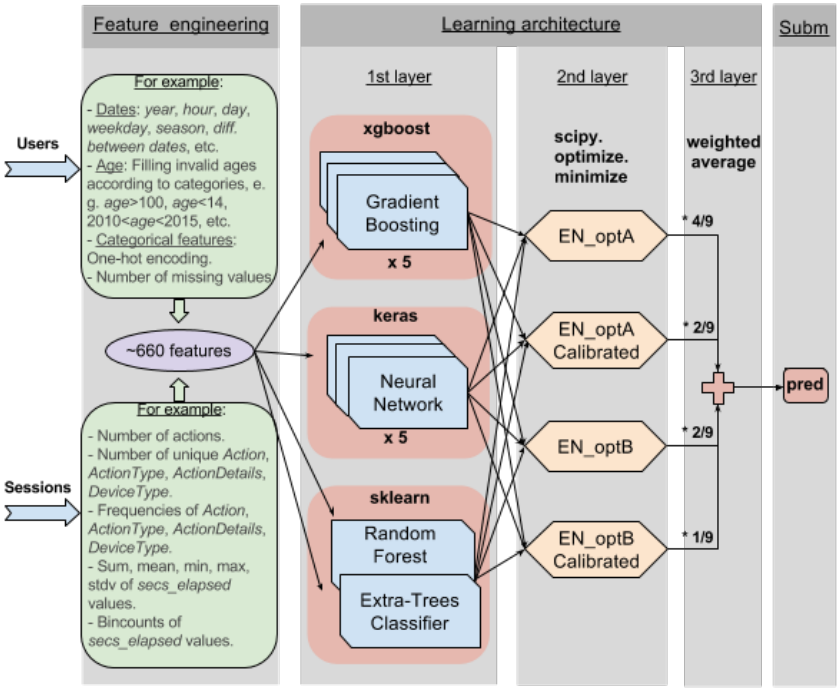
A.2 AirBnB: Reservas de novos usuários

Competição realizada pela plataforma *Kaggle* em 2016, na qual a empresa de hospedagem *AirBnB* lançou o desafio: Em qual país um novo usuário irá realizar a sua primeira reserva?

Sandro Vegas Pons, pesquisador de pós doutorado na Universidade de Trento na Itália, obteve a 3ª Colocação diante de 1462 competidores. A estratégia e modelagem adotada foi compartilhada na plataforma *Kaggle*, a qual é analisada a seguir (TEAM, 2016).

A solução baseia-se na aplicação de FE e em uma arquitetura de 3 camadas combinando diferentes técnicas de aprendizado, apresentadas na Figura 25.

Figura 25 – Modelagem vencedora da competição AirBnB 2016



Fonte: (TEAM, 2016)

Na parte da FE, foram utilizados: técnica de *one-hot-enconding*, preenchimento de valores faltantes com diferentes método, agrupamento e agregação de características baseado em contadores, frequência, elementos únicos, etc.

A arquitetura de aprendizado é composta de modelos diversos. Na primeira camada foram utilizados os seguintes classificadores: cinco versões de *Gradient Boosting*, cinco versões de Redes Neurais, uma versão de RF e uma versão de DT.

Na segunda camada as predições anteriores foram concatenadas para criar um novo

conjunto de treinamento, que foram aplicados a dois modelos *ensemble*, gerando novas predições.

Na terceira camada, as quatro predições anteriores foram linearmente combinadas com pesos fixos, obtendo assim a predição final.

APÊNDICE B – AGREGAÇÃO DO ATRIBUTO *SIZE*

A seguir são apresentados os valores agregados do atributo *size*:

- “XS”, “XS (30-34)”, “XS (32-34)”, “XS (32)”, “XS/S”, “XS (116-128)”, “S (36)”, “XS (30-34)”, “XS (32/34)”: **“XS”**
- “S”, “S (34-38)”, “S (34/36)”, “S (36)”, “S (128-140)”, “XS (32)”, “S (34-38)”, “S (34/36)”, “5 (46-48)”: **“S”**
- “M”, “M (38/40)”, “M (38/42)”, “M (40)”, “38/40 (M/L)”, “M (38-42)”, “M/L”, “102 (M)”, “24 (M)”, “M (140-152)”, “M (38 - 42)”, “38/40 (M / L)”, “M (38-42)”: **“M”**
- “L”, “L (40/42)”, “L (42/46)”, “L (43 - 46)”, “L/XL (39-47)”, “L (42-47)”, “L (44)”, “L/K”, “L/T”, “L (152-158)”, “L (42-46)”, “7 (L)”, “L (43 - 46)”: **“L”**
- “XL”, “XL (44/46)”, “XL (46-50)”, “XL (46-48,5)”, “L/XL (39-47)”, “XL (158-170)”, “XL/T”, “8 (XL)”: **“XL”**
- “2XL”, “2XL/T”, “3XL”, “3XL/T”, “28 (3XL)”: **“XXL”**
- “4XL”, “30 (5XL)”: **“XXXL”**
- “0 (Bambini)”, “1 (Junior)”, “01 Junior”, “YSM 125-135”, “YM 135-147,5”, “YLG 147,5-157,5”, “YXL 157,5-167,5”, “1 (Junior)”: **“Kids”**
- “2 (Senior)”, “02 Senior”: **“Senior”**
- “31”, “31,5”: **“31”**
- “33”, “33,5”: **“33”**
- “35”, “35,5”: **“35”**
- “36”, “36 2/3”, “36,5”: **“36”**
- “37”, “37 1/3”, “37,5”: **“37”**
- “38”, “38 2/3”, “38,5”, “19 (38)”: **“38”**
- “39”, “39 1/3”, “39,5”: **“39”**
- “40”, “40 2/3”, “40,5”: **“40”**
- “41”, “41 1/3”, “41,5”: **“41”**
- “42”, “42 2/3”, “42,5”: **“42”**
- “43”, “43 1/3”, “43,5”: **“43”**
- “44”, “44 2/3”, “44,5”: **“44”**
- “45”, “45 1/3”, “45,5”: **“45”**
- “46”, “46 2/3”, “46,5”: **“46”**
- “47”, “47 1/3”, “47,5”: **“47”**
- “48”, “48 2/3”, “48,5”: **“48”**
- “128”, “128 0 (128)”, “0 (128)”: **“128”**

- “140”, “1 (140)”, “10 (140)”: **“140”**
- “152”, “2 (152)”: **“152”**
- “164”, “14 (164)”, “3 (164)”: **“164”**
- “176”, “16 (176)”: **“176”**
- “1 (25-30)”, “00 (27-30)”: **“25-30”**
- “0 (31-33)”, “1 (31-34)”, “2 (31-34)”: **“31-34”**
- “1 (33-36)”, “1 (34-36)”, “35 - 38”, “35/38”, “3 (35-38)”, “2 (35-38)”, “2 (37-39)”, “10 (35-40)”, “34 - 40”, “2 (37-40)”, “10 (36-40)”: **“35-40”**
- “39 - 42”, “39-42”, “39/42”, “3 (39-42)”, “4 (39-42)”, “3 (40-42)”, “3 (41-43)”, “41 - 44”, “12 (41-45)”, “4 (43-45)”: **“41-45”**
- “43 - 46”, “43-46”, “4 (43-46)”, “5 (43-46)”, “43/46”, “4 (44-46)”, “45 - 47”: **“44-46”**
- “45-48”, “14 (46-48)”, “5 (47-49)”, “47/49”, “47 - 50”, “6 (47-50)”: **“47-50”**
- “116-122”, “6/8 (116-128)”, “116/128”: **“116-128”**
- “10/12 (140-152)”, “140/152”: **“140-152”**
- “14/16 (164-176)”, “164/176”: **“164-176”**