

**PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS**  
**NÚCLEO DE EDUCAÇÃO A DISTÂNCIA**  
**Pós-graduação *Lato Sensu* em Ciência de Dados e Big Data**

**Pablo Henrique Ramos da Silva**

**O IMPACTO DAS AMOSTRAS DESBALANCEADAS  
NOS PROBLEMAS DE CLASSIFICAÇÃO**

Belo Horizonte  
2020

**Pablo Henrique Ramos da Silva**

**O IMPACTO DAS AMOSTRAS DESBALANCEADAS  
NOS PROBLEMAS DE CLASSIFICAÇÃO**

Trabalho de Conclusão de Curso apresentado  
ao Curso de Especialização em Ciência de  
Dados e Big Data como requisito parcial à  
obtenção do título de especialista.

Belo Horizonte  
2020

## **AGRADECIMENTO**

Agradeço a Deus pela vida.

A minha esposa Monica e ao meu filho Eric pela parceria e paciência nos momentos de ausência e estudo.

Ao amigo e professor Dr. Nélío Machado, por me incentivar, revisar e me orientar pacientemente no desenvolvimento do TCC e pelo conhecimento e aprendizado que tive a oportunidade de experimentar durante todo o processo.

Aos professores e a PUC minas por todo o conhecimento compartilhado durante todo o curso.

## LISTA DE FIGURAS

Figura 1- Quadrante Gartner 2018 .....	9
Figura 2 - KNIME Workflow   Tratamento da variável Company .....	10
Figura 3 - KNIME Workflow   Enriquecimento de dados - Country .....	11
Figura 4 - KNIME Workflow   Tratamento de datas – Arrival e Departure.....	11
Figura 5 - KNIME Workflow   Criação de nova variável Qty_days.....	12
Figura 6 - Imagem representativa de um outlier .....	12
Figura 7 - Imagem representativa de um "Interquartile range" .....	13
Figura 8 - KNIME Workflow   Tratamento e calculo da mediana para a variável Distance....	13
Figura 9 - KNIME Workflow   Processo de "binning" das variáveis Distance e Qty_days.....	14
Figura 10 - KNIME Workflow   Tratamento de missing values para variável Device.....	14
Figura 11 - KNIME Workflow   Objeto One to Many .....	14
Figura 12 - Análise de dados   WOE e IV da variável GDS.....	17
Figura 13 - Análise de dados   WOE e IV da variável Adults .....	17
Figura 14 - Análise de dados   WOE e IV da variável Children .....	18
Figura 15 - Análise de dados   WOE e IV da variável Infants.....	18
Figura 16 - Análise de dados   WOE e IV da variável Train .....	18
Figura 17 - Análise de dados   WOE e IV da variável Haul_type .....	19
Figura 18 - Análise de dados   WOE e IV da variável Distance .....	19
Figura 19 - Análise de dados   WOE e IV da variável Device.....	19
Figura 20 - Análise de dados   WOE e IV da variável Trip_type .....	20
Figura 21 - Análise de dados   WOE e IV da variável Product.....	20
Figura 22 - Análise de dados   WOE e IV da variável SMS.....	20
Figura 23 - Análise de dados   WOE e IV da variável NO_GDS .....	21
Figura 24 - Análise de dados   WOE e IV da variável Company.....	21
Figura 25 - Análise de dados   WOE e IV da variável Qty_days.....	21
Figura 26 - Análise de dados   WOE e IV da variável Country .....	22
Figura 27 - Análise de dados   Tabela de classificação de variáveis.....	22
Figura 28 - KNIME Workflow   Modelos preditivos .....	23
Figura 29 - Gráfico e tabela de balanceamento da amostra.....	25
Figura 30 - Representação gráfica de amostra desbalanceada.....	26
Figura 31 - Representação gráfica de amostra balancead.....	27
Figura 32 - KNIME Workflow   Configuração do objeto "Equal Size Sampling" .....	27
Figura 33 - Quantitativo de instancias desbalanceadas .....	28
Figura 34 - Quantitativo de instancias balanceadas .....	28
Figura 35 - KNIME Workflow   Visão geral do modelo (Parte 1).....	29
Figura 36 - KNIME Workflow   Visão geral do modelo (Parte 2).....	29
Figura 37 - Matriz Confusão modelo desbalanceado (Randon Forest).....	31
Figura 38 - Matriz Confusão modelo balanceado (XGBooster) .....	33
Figura 39 - Fórmula para cálculo da Taxa Positiva Verdadeira (ROC).....	34
Figura 40 - Fórmula para cálculo da Taxa Falso Positivo (ROC).....	34
Figura 41 - Curva ROC do modelo preditivo final .....	34

## SUMÁRIO

<b>1. Introdução .....</b>	<b>6</b>
<b>1.1. Contextualização .....</b>	<b>6</b>
<b>1.2. O problema proposto .....</b>	<b>6</b>
<b>2. Coleta de Dados.....</b>	<b>7</b>
<b>3. Ferramenta .....</b>	<b>8</b>
<b>4. Processamento e tratamento de Dados.....</b>	<b>10</b>
<b>5. Análise e Exploração dos Dados .....</b>	<b>15</b>
<b>6. Construção de Modelos de Machine Learning.....</b>	<b>23</b>
<b>7. Apresentação dos Resultados .....</b>	<b>30</b>
<b>8. Links .....</b>	<b>36</b>
<b>9. Referências .....</b>	<b>37</b>

## 1. Introdução

### 1.1. Contextualização

Tanto os dados utilizados na análise e desenvolvimento dos modelos quanto o nome real da empresa foram anonimizados para que a empresa não seja identificada. A empresa aqui em questão foi denominada como eFly.

A eFly é uma agência de viagens online, presente em diversos países e com a combinação mais vasta de produtos no mercado, que disponibiliza mais de 155.000 rotas de voos de mais de 660 companhias aéreas e mais de 1.700.000 de hotéis em 40.000 destinos. A empresa desenvolve e utiliza as ferramentas mais avançadas para pesquisar milhões de combinações de voos e hotéis, garantindo desta forma o melhor preço e a maior conveniência aos seus clientes.

A eFly está sempre buscando maneiras de melhorar a satisfação de seus clientes através de uma oferta precisa de produtos e serviços de acordo com a necessidade individual de cada cliente. Com este objetivo, gostaríamos de prever se os clientes estão interessados em contratar bagagem de porão adicional para acelerar o processo de reserva e check-in, além de estimativa de receitas adicionais com bagagens.

### 1.2. O problema proposto

A eFly precisa identificar os potenciais compradores de bagagem extra para otimizar e direcionar seus esforços de venda. O custo de realizar uma ação de venda não direcionada além de elevado costuma ser pouco efetivo e por este motivo a proposta consiste em analisar dados históricos de vendas e desenvolver um modelo preditivo capaz de otimizar e identificar eficientemente os potenciais compradores de bagagens adicionais, de forma a auxiliar e direcionar as campanhas de marketing.

Para facilitar o entendimento do problema e da solução a ser proposta utilizamos a técnica do 5WS, que consiste em responder as seguintes perguntas:

**Why?** Auxiliar a equipe de vendas e marketing a direcionar seus esforços a clientes/passageiros cujo potencial de compra seja elevado.

**Who?** Equipe de marketing e vendas da eFly.

**What?** Potenciais compradores de bagagem extra através da análise de dados históricos da eFly que, posteriormente serão utilizados para desenvolver o modelo preditivo.

**Where?** Nos pontos de venda da eFly (físico e online).

**When?** No momento da interação do cliente/comprador com a eFly. Com relação aos dados analisados, foi utilizado um *subset* relativo ao período de aproximadamente 1 ano.

## 2. Coleta de Dados

Os dados utilizados na pesquisa são reais e foram cedidos por uma grande empresa do setor de viagens com a condição de que a empresa não fosse identificada e que seus dados fossem anonimizados de forma a preservar a individualidade e estratégia da empresa.

Optamos por utilizar estes dados por ser um caso real de um problema conhecido e amplamente abordado e tratado por cientistas de dados e profissionais de *analytics* em todo o mundo.

Os dados de treinamento totalizam 50.000 instâncias que deverão ser divididos entre amostra de treinamento e amostra de teste para desenvolvimento e validação do modelo preditivo.

O arquivo no formato CSV (*Comma-separated Values*) contem a seguinte estrutura de variáveis:

Coluna	Tipo	Descrição
TIMESTAMP	Date	Data em que a reserva foi realizada.
WEBSITE	String	Website em que a viagem foi comprada. É composto por um prefixo que representa o site ("ED", "OP", "GO") e um sufixo para o país (por exemplo: ES = Espanha)
GDS	Integer	Número de voos comprados através do GDS (Sistema de Distribuição Global)
No. GDS	Integer	Número de voos comprados por outros canais.
DEPARTURE	Date	Data de partida
ARRIVAL	Date	Data de chegada

ADULTS	Integer	Número de adultos
CHILDREN	Integer	Número de filhos
INFANTS	Integer	Número de bebês
TRAIN	Boolean	Indica se a reserva contém bilhetes de trem ou não
DISTANCE	Float	Distância total percorrida
DEVICE	String	Dispositivo usado para a compra
HAUL TYPE	String	Indica o tipo de transporte, se foi "doméstica", "continental" ou "intercontinental".
TRIP TYPE	String	As viagens podem ser "Ida", "Ida e Volta" ou "Vários destinos"
PRODUCT	String	As reservas podem conter apenas viagens ("Viagem") ou viagens + hotel ("Dynpack").
SMS	Boolean	Indica se o cliente selecionou uma confirmação por SMS
<b>EXTRA BAGGAGE</b>	<b>Boolean</b>	Variável resposta a ser estimada. Indica se o cliente/passageiro comprou ou não bagagem adicional para sua viagem.

### 3. Ferramenta

Optamos pela utilização da ferramenta KNIME ([www.knime.com](http://www.knime.com)) por ser uma ferramenta analítica de fácil utilização e boa performance tanto para as fases de tratamento e transformação dos dados quanto para a fase de desenvolvimento dos modelos preditivos.

KNIME ( / n aɪ m / ), o *Konstanz Information Miner*, é uma livre e open-source de análise de dados, relatórios e plataforma de integração. KNIME integra vários componentes para aprendizado de máquina e mineração de dados através do seu conceito de *data pipelining* modular. A interface gráfica de utilizador e utilização de JDBC permite a montagem de nodos de mistura diferentes fontes de dados, incluindo o pré-processamento ( ETL: Extração, transformação, Carregamento ), para a modelagem, análise de dados e de visualização, sem, ou com apenas o mínimo, de programação (*low-code*).

O Desenvolvimento da KNIME foi iniciado em janeiro de 2004 por uma equipe de engenheiros de software na Universidade de Konstanz como um produto



proprietário. A equipe de desenvolvimento original, dirigido por Michael Berthold, veio de uma empresa no Vale do Silício fornecendo software para a indústria farmacêutica. O objetivo inicial era criar uma plataforma modular, altamente escalável e aberta de processamento de dados que permitisse a fácil integração de diferentes carregamentos de dados, processamento, transformação, análise e módulos de exploração visual, sem o foco em qualquer área de aplicação particular. A plataforma pretendia não apenas ser uma plataforma de colaboração e pesquisa, mas também servir como uma plataforma de integração de vários outros projetos de análise de dados.

Em 2006, a primeira versão do KNIME foi lançada e várias empresas farmacêuticas começaram a usar KNIME e um número de fornecedores de software de ciências da vida começou a integrar suas ferramentas em KNIME. A partir de 2012, KNIME é utilizado por mais de 15.000 utilizadores reais (isto é, sem contar *downloads*, apenas usuários que regularmente realizam atualizações) não só nas ciências da vida, mas também em bancos, editoras, fabricante de automóveis, empresas de telecomunicações, empresas de consultoria e várias outras indústrias, mas também em um grande número de grupos de pesquisa em todo o mundo. Atualizações mais recentes para KNIME Server e KNIME Dados Extensões grandes, para fornecer suporte para Apache 2.3.

No quadrante do Gartner de 2018, o KNIME aparece entre os 4 líderes da categoria “Data Science and Machine Learning Platforms”.

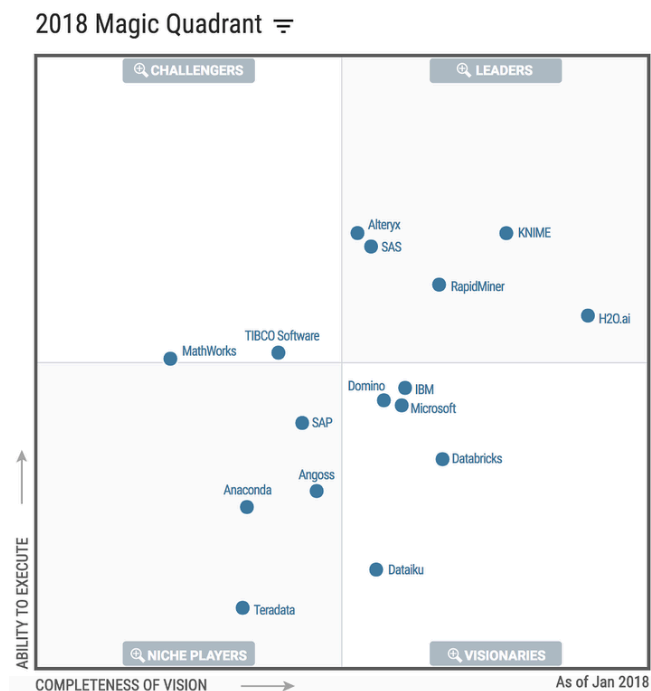


Figura 1- Quadrante Gartner 2018

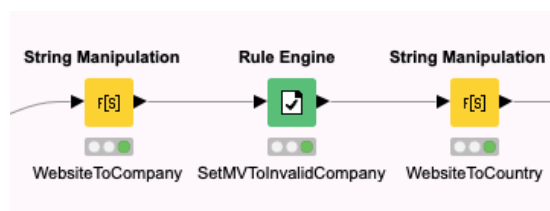
#### 4. Processamento e tratamento de Dados

O dataframe importado possui 50.000 instâncias, aos quais procedemos a análise, limpeza e tratamento dos dados. Nesta fase identificamos diversas inconsistências que foram devidamente tratadas, conforme sugere a literatura.

Importante ressaltar que todas as alterações realizadas foram submetidas a testes, simulações e validações para garantir que o resultado final não fosse impactado por manipulações indevidas dos dados.

Assim sendo, os dois primeiros caracteres da variável **WEBSITE** se referem ao site ou portal em que a passagem foi comprada e os caracteres restantes correspondem ao País da compra. Decidimos criar então duas novas variáveis: **COMPANY** e **COUNTRY**, para receber as novas informações que foram derivadas da variável **WEBSITE**.

O novo campo **COMPANY** possui apenas 3 valores possíveis, ED, OP e GO. No entanto, durante o processo de entendimento dos dados, foi identificada uma inconsistência: haviam algumas instâncias com valor “TL”, que não conseguimos associar a nenhuma empresa do metadata. Por este motivo, as instâncias com valor “TL” foram substituídas por “MV” (Missing Value).



*Figura 2 - KNIME Workflow | Tratamento da variável Company*

O novo campo **COUNTRY** foi confrontado com uma lista oficial países (posteriormente retirada do modelo) para garantir 100% de consistência. Os valores a seguir foram considerados inconsistentes por não terem sido encontrados na lista oficial e foram substituídos pela sigla do País correto. Uma análise manual foi realizada nos valores não encontrados para tratamento e enriquecimento dos dados de forma a evitar perda de informação, conforme mencionado pelos autores Y. Wang, H. Tercan, T. Thiele, T. Meisen, S. Jeschke e W. Schulz, no artigo "Advanced data enrichment and data analysis in manufacturing industry by an example of laser drilling process".

Foram substituídos os valores conforme a tabela abaixo:

Código Antigo	Código Novo	Nome País
PLC	PL	Polônia
FRC	FR	França
DEC	DE	Alemanha
DKC	DK	Dinamarca
UK	GB	Reino Unido

Os países que não puderam ser identificados tiveram suas respectivas siglas substituídas por “MV (Missing Value)”.

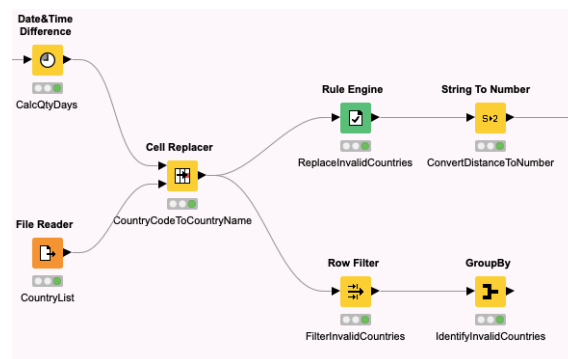


Figura 3 - KNIME Workflow | Enriquecimento de dados - Country

Os campos **DEPARTURE** e **ARRIVAL** apesar de serem do tipo *date* não estão corretamente formatados (DD/MON) e precisam receber o ano ao final para que a ferramenta entenda e trate os dados com o formato correto. Como não se sabe o ano em que ocorreram as viagens, foi atribuído o ano de 2018 para todas as “departures”. Para os “arrivals” o valor 2018 também foi atribuído, exceto para instâncias em que “arrival” era menor que “departure”. Para estes casos, o ano de 2019 foi atribuído para “arrival”.

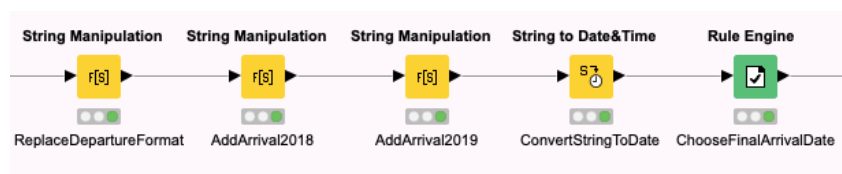


Figura 4 - KNIME Workflow | Tratamento de datas – Arrival e Departure

Como *Feature Engineering*, criamos uma nova variável denominada **QTY\_DAYS**, criada a partir da subtração entre as variáveis **ARRIVAL** e **DEPARTURE** ( $ARRIVAL - DEPARTURE$ ), para armazenar a quantidade de dias de cada viagem realizada.



Figura 5 - KNIME Workflow | Criação de nova variável Qty\_days

Em função da discrepância identificada entre as variáveis **QTY\_DAYS** e **DISTANCE** quando relacionado ao tipo de transporte (doméstico, continental ou intercontinental) e ao tipo de viagem (ida, volta ou múltiplos destinos), identificou-se a necessidade de tratamento dos outliers para estas variáveis.

Conforme recomendado pelo autor Aggarwal C. (2015), um outlier é uma observação que se desvia tanto das outras observações que suscita suspeitas de que foi gerada por um mecanismo diferente. Se o ponto de dados estiver acima do limite superior ou abaixo do limite inferior, ele poderá ser considerado um desvio.

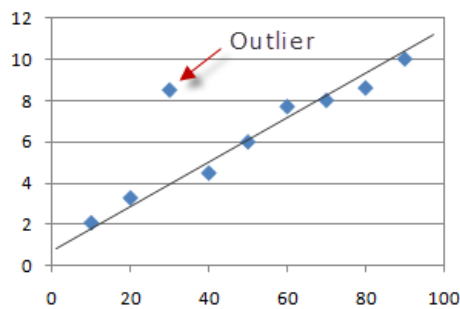


Figura 6 - Imagem representativa de um outlier

Na prática calculamos o limite superior de forma que o ponto/observação será considerado outlier toda vez que ultrapassa esse limite calculado. O limite superior será dado por  $Q3 + 1.5 * IQR$ , onde IQR ( $Q3 - Q1$ ) é a distância Interquartilica. Abaixo apresentamos uma figura que demonstra o cálculo do IQR e dos quartis Q1 e Q3.

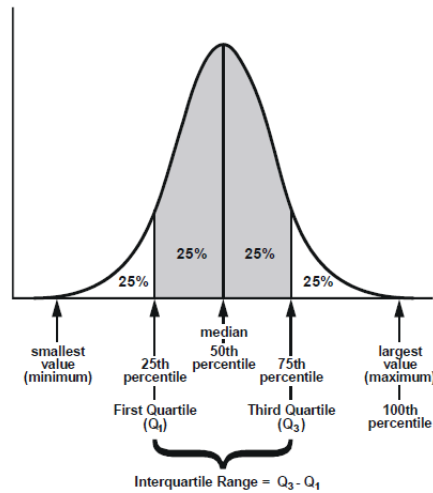


Figura 7 - Imagem representativa de um "Interquartile range"

O campo **DISTANCE** também possuía alguns valores zerados aos quais consideramos como sendo *Missing Values* ("MV"). Neste caso, decidimos substituir os valores zerados pela mediana, considerando a agregação pelo tipo de transporte e pelo tipo de viagem (**HAUL\_TYPE** e **TRIP\_TYPE**, respectivamente).

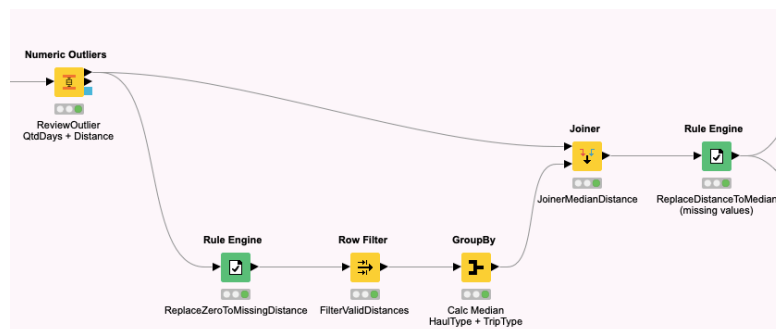


Figura 8 - KNIME Workflow | Tratamento e calculo da mediana para a variável Distance

Valores faltantes (*missing values*) é um problema recorrente nos dados das empresas e o devido tratamento deve ocorrer durante a fase de pré-processamento dos dados, que precede a criação de um modelo analítico. As causas são originadas por diferentes motivos tais como falha no equipamento que transmite e armazena os dados, falha do manipulador, falha de integrações de sistemas, falha de quem fornece a informação, e outros. Essa situação pode tornar os dados inconsistentes e inaptos de serem analisados, conduzindo a conclusões que não condizem com a realidade. Face a isso, decidimos realizar o tratamento de *missing values* conforme recomenda o autor Larose et. al (2019). Inclusive, o autor apresenta em sua pesquisa os impactos dos *missing values*, pois os mesmos aumentam a incerteza em um conjunto de dados.

Como sabemos, maior incerteza no conjunto de dados nos leva a modelos preditivos menos acurados e ineficientes.

Após todos os tratamentos dados a variável **DISTANCE**, notou-se uma variedade muito grande de valores, o que dificulta a utilização da variável no modelo. Para resolver este problema, o autor Zeng, G. (2014) recomenda como boas práticas a técnica de “*binning*”, que consiste em construir  $k$  classes para as variáveis numéricas. Em outras palavras, a técnica consiste em categorizar através da criação de classes (também conhecida como *ranges*) para as variáveis numéricas.

Para resolver este problema, recomenda-se a técnica de “*binning*”, que consiste na utilização de algoritmos próprios que identificam e formam os  $n$  agrupamentos (*clustering*), de acordo com o número informado no parâmetro de entrada da função. Após testes de performance do modelo final chegou-se no número de 10 grupos. O mesmo tratamento foi aplicado com a variável **QTY\_DAYS**.

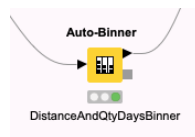


Figura 9 - KNIME Workflow | Processo de "binning" das variáveis Distance e Qty\_days

O campo **DEVICE** possuía alguns valores nulos e estes foram substituídos por MV (Missing Values).



Figura 10 - KNIME Workflow | Tratamento de missing values para variável Device

De forma a melhorar a performance nos modelos preditivos, utilizamos o objeto “One to Many” do *KNIME*, que consiste em construir variáveis *dummies* para cada uma das variáveis categóricas, conforme exemplo abaixo:

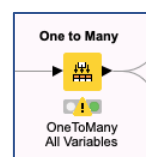
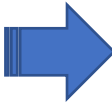


Figura 11 - KNIME Workflow | Objeto One to Many

ROW	DEVICE		ROW	SMARTPHONE_DEVICE	COMPUTER_DEVICE
1	Smartphone		1	1	0
2	Computer		2	0	1
3	Smartphone		3	1	0
4	Smartphone		4	1	0

## 5. Análise e Exploração dos Dados

Análise exploratória de dados (AED) é uma etapa muito importante que se realiza depois da engenharia de características [*feature engineering*, ou seja, o pré-processamento dos dados para técnicas de *machine learning*, aprendizado de máquinas] e da coleta de dados. Deve-se colocá-la em prática antes de qualquer tipo de modelagem em si. Isso acontece pois é essencial que o cientista de dados seja capaz de entender a natureza dos dados sem fazer suposições.

O objetivo da AED é utilizar síntese estatística e técnicas de visualização para entender melhor os dados e identificar *insights* sobre tendências e a qualidade dos dados, bem como para formular hipóteses e fazer suposições nas análises. Análise exploratória de dados NÃO SE TRATA de elaborar visualizações sofisticadas ou mesmo esteticamente agradáveis. O objetivo é fazer testes e encontrar respostas com os dados. Seu objetivo, enquanto cientista de dados, deveria ser criar um gráfico, no qual qualquer um que o olhasse por alguns segundos pudesse entender o que se passa. Caso contrário, a visualização é muito complicada (ou sofisticada) e algum similar simplificado deveria ser utilizado.

As métricas estatísticas que serão utilizadas a seguir servem para nos auxiliar na identificação das variáveis mais importantes do conjunto de dados, ou seja, nos ajuda na identificação das variáveis com maior poder preditivo:

- WOE – A métrica WOE (*Weight Of Evidence*) ou, em português, peso da evidência, informa o poder preditivo de uma variável independente em relação à variável dependente. Em outras palavras, o WOE identifica o peso da variável com relação a ocorrência e não ocorrência do evento (variável resposta). WOE positivo significa maior ocorrência do evento do que o não evento ou, para nosso caso, maior ocorrência de compras de bagagens extra do que a não ocorrência. O cálculo do WOE se dá pela fórmula descrita abaixo:

$$WOE = \ln \left( \frac{\% \text{ of non-events}}{\% \text{ of events}} \right)$$

Onde:

% of events - % de compra de bagagens adicionais para um determinado grupo

% of non-events - % de compra de bagagens adicionais para um determinado grupo

- IV – A métrica IV (Information Value) ou, em português, valor da informação é uma das técnicas mais úteis para selecionar variáveis importantes em um modelo preditivo. Ajuda a classificar as variáveis com base em sua importância. O IV é calculado usando a seguinte fórmula:

$$IV = \sum (\% \text{ of non-events} - \% \text{ of events}) * WOE$$

A interpretação do IV deve seguir a seguinte tabela:

Information Value	Preditividade da Variável
< 0,02	Sem utilidade para o modelo
0,02 - 0,1	Baixo poder preditivo
0,1 - 0,3	Médio poder preditivo
0,3 - 0,5	Alto poder preditivo
> 0,5	Poder preditivo suspeito

Ou seja, se a estatística IV do preditor for:

- ⇒ Menos que 0,02, o preditor não é útil para modelagem;
- ⇒ De 0,02 a 0,1, o preditor tem apenas uma relação fraca com a razão de chances de variável resposta positiva;
- ⇒ De 0,1 a 0,3, o preditor tem uma relação de força média com a razão de chances variável resposta positiva;
- ⇒ De 0,3 a 0,5, o preditor tem uma forte relação com a razão de chances de variável resposta positiva;
- ⇒ Mais que 0,5, relacionamento suspeito (deve-se verificar/validar caso a caso);

Para cada uma das classes/grupos encontrados em cada variável contida no conjunto de dados, calculamos a distribuição de ocorrências de compra de bagagens adicionais (EXTRA\_BAGGAGE\_1) e não compra de bagagens



adicionais (EXTRA\_BAGGAGE\_0). Calculamos também o WOE e IV, conforme detalhado abaixo:

GDS	EXTRA_BAGGAGE_0	EXTRA_BAGGAGE_1	DISTR_0	DISTR_1	WOE	IV
4	2	0	0,00%	0,00%	-	-
3	41	9	0,10%	0,09%	-0,1047	0,0000
2	2216	342	5,51%	3,49%	-0,4570	0,0092
1	23077	3770	57,40%	38,47%	-0,4001	0,0758
0	14865	5678	36,98%	57,94%	0,4492	0,0942
Total	40201	9799				0,1792

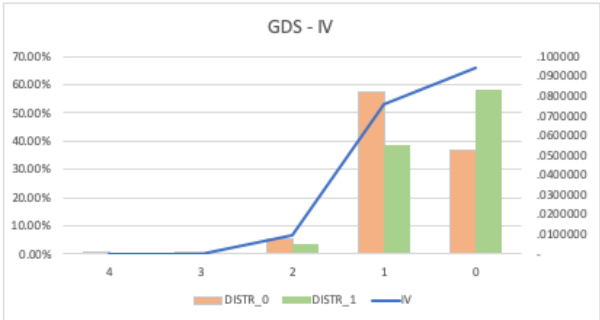


Figura 12 - Análise de dados | WOE e IV da variável GDS

GDS

As classes de GDS isoladas apresentam um baixo IV, como podemos observar na tabela ao lado. Mas se olharmos para o IV total, a variável possui uma média capacidade preditiva (pertencente majoritariamente às classes 1 e 0). Variável pode e deve ser utilizada como parte do modelo preditivo.

ADULTS

Semelhante a variável GDS, as classes de ADULTS isoladas apresentam um baixo IV, mas se olharmos para o IV total a variável possui uma capacidade preditiva média (data majoritariamente pelos valores 1 e 2). Variável pode e deve ser utilizada como parte do modelo preditivo.

Insight: Mais de 2 adultos viajando juntos tendem a não comprar bagagens extras

ADULTS	EXTRA_BAGGAGE_0	EXTRA_BAGGAGE_1	DISTR_0	DISTR_1	WOE	IV
0	10	0	0,02%	0,00%	-	-
8	24	6	0,06%	0,06%	0,0253	0,0000
9	17	6	0,04%	0,06%	0,3702	0,0001
7	25	11	0,06%	0,11%	0,5906	0,0003
5	213	84	0,53%	0,86%	0,4811	0,0016
6	93	54	0,23%	0,55%	0,8680	0,0028
3	1654	579	4,11%	5,91%	0,3620	0,0065
4	838	367	2,08%	3,75%	0,5860	0,0097
1	27221	4990	67,71%	50,92%	-0,2849	0,0478
2	10106	3702	25,14%	37,78%	0,4074	0,0515
Total	40201	9799				0,1203

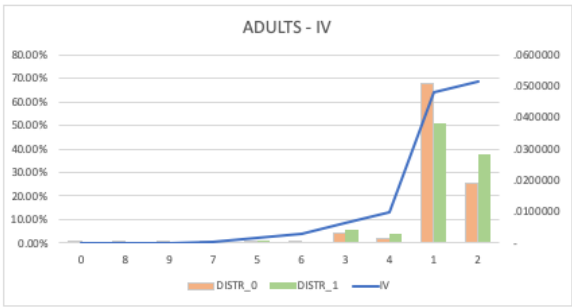


Figura 13 - Análise de dados | WOE e IV da variável Adults

CHILDREN	EXTRA_BAGGAGE_0	EXTRA_BAGGAGE_1	DISTR_0	DISTR_1	WOE	IV
5		1	0,00%	0,01%	-	-
4	23	3	0,06%	0,03%	-0,6253	0,0002
3	114	43	0,28%	0,44%	0,4366	0,0007
0	37512	8849	93,31%	90,31%	-0,0327	0,0010
2	661	259	1,64%	2,64%	0,4747	0,0047
1	1891	644	4,70%	6,57%	0,3344	0,0062
Total	40201	9799				0,0128

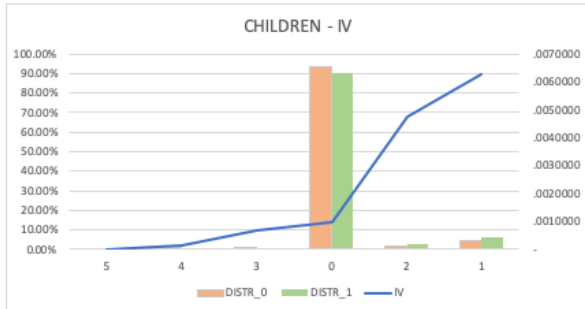


Figura 14 - Análise de dados | WOE e IV da variável Children

## CHILDREN

Variável de baixa IV, pouco relevante para o modelo e sem capacidade preditiva. Não deve ser levada em consideração para construção do modelo preditivo.

## INFANTS

Variável de baixa IV, pouco relevante para o modelo e sem capacidade preditiva. Não deve ser levada em consideração para construção do modelo preditivo.

INFANTS	EXTRA_BAGGAGE_0	EXTRA_BAGGAGE_1	DISTR_0	DISTR_1	WOE	IV
2	15		0,04%	0,00%	-	-
0	39531	9576	98,33%	97,72%	-0,0062	0,0000
1	655	223	1,63%	2,28%	0,3341	0,0022
Total	40201	9799				0,0022

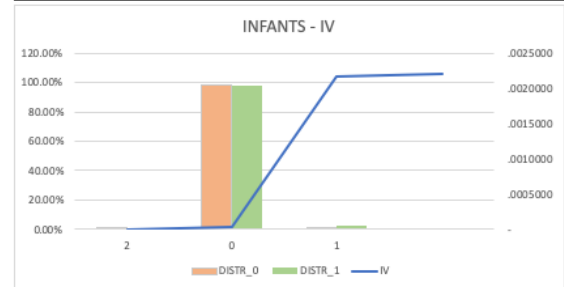


Figura 15 - Análise de dados | WOE e IV da variável Infants

TRAIN	EXTRA_BAGGAGE_0	EXTRA_BAGGAGE_1	DISTR_0	DISTR_1	WOE	IV
0	39933	9798	99,33%	99,99%	0,0066	0,0000
1	268	1	0,67%	0,01%	-4,1794	0,0274
Total	40201	9799				0,0275



Figura 16 - Análise de dados | WOE e IV da variável Train

## TRAIN

A variável TRAIN possui um baixo valor de IV. Se observarmos os valores contidos na variável podemos aferir que o valor 1 possui uma baixa capacidade preditiva, porém este representa menos de 1% do total da amostra, o que torna esta variável sem poder preditivo algum. Variável pode ser utilizada como parte do modelo preditivo, mas não se deve esperar nenhuma alteração nos resultados.

## HAUL\_TYPE

As classes da variável HAUL\_TYPE apresentam média capacidade preditiva. Junto com a NO\_GDS, a variável HAUL\_TYPE é a variável com maior preditividade no conjunto de dados. Insight: Nota-se que o valor “DOMESTIC” detêm um baixo poder preditivo, porém contempla aproximadamente 70% dos casos de passageiros que compraram bagagem adicional, quando o esperado era bagagens adicionais requeridas para voos mais longos.

HAUL_TYPE	EXTRA_BAGGAGE_0	EXTRA_BAGGAGE_1	DISTR_0	DISTR_1	WOE	IV
CONTINENTAL	8960	2093	22,29%	21,36%	-0,0426	0,0004
DOMESTIC	19237	6641	47,85%	67,77%	0,3480	0,0693
INTERCONTINENTAL	12004	1065	29,86%	10,87%	-1,0107	0,1919
Total	40201	9799				0,2617

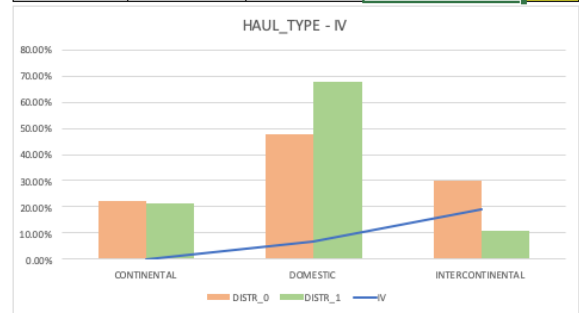


Figura 17 - Análise de dados | WOE e IV da variável Haul\_type

DISTANCE_BIN	EXTRA_BAGGAGE_0	EXTRA_BAGGAGE_1	DISTR_0	DISTR_1	WOE	IV
[364,100322]	4084	1020	10,16%	10,41%	0,0243	0,0001
(800026,899984]	2947	695	7,33%	7,09%	-0,0330	0,0001
(200280,300238]	4566	1154	11,36%	11,78%	0,0362	0,0002
(700068,800026]	2337	507	5,81%	5,17%	-0,1165	0,0007
(600111,700068]	2644	558	6,58%	5,69%	-0,1441	0,0013
(500153,600111]	2460	492	6,12%	5,02%	-0,1978	0,0022
(899984,999941]	2565	497	6,38%	5,07%	-0,2295	0,0030
(300238,400195]	2165	398	5,39%	4,06%	-0,2821	0,0037
(400195,500153]	2873	524	7,15%	5,35%	-0,2900	0,0052
(100322,200280]	13560	3954	33,73%	40,35%	0,1792	0,0119
Total	40201	9799				0,0283

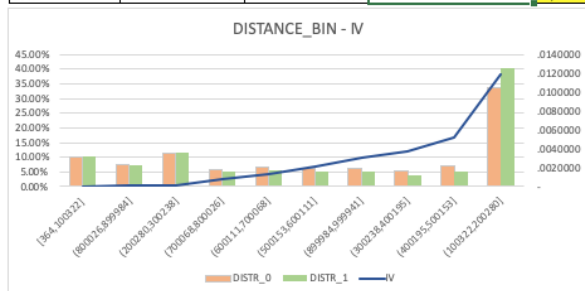


Figura 18 - Análise de dados | WOE e IV da variável Distance

## DISTANCE\_BIN

As classes de DISTANCE foram binarizadas e ainda assim apresentam baixa capacidade preditiva.

Insight: Observe que a classe com maior valor de IV é a classe "(100322,200280]", com 0,119. Novamente, o razoável seria para maiores distâncias maior probabilidade de se obter bagagens adicionais pois, em teoria, maiores distâncias demandam mais dias em viagem.

## DEVICE

Variável de baixa IV, pouco relevante para o modelo e sem capacidade preditiva. Não deve ser levada em consideração para construção do modelo preditivo.

DEVICE	EXTRA_BAGGAGE_0	EXTRA_BAGGAGE_1	DISTR_0	DISTR_1	WOE	IV
MV	133	0	0,33%	0,00%	-	-
TABLET	2508	644	6,24%	6,57%	0,0521	0,0002
OTHER	722	220	1,80%	2,25%	0,2232	0,0010
COMPUTER	27094	6970	67,40%	71,13%	0,0539	0,0020
SMARTPHONE	9744	1965	24,24%	20,05%	-0,1895	0,0079
Total	40201	9799				0,0111

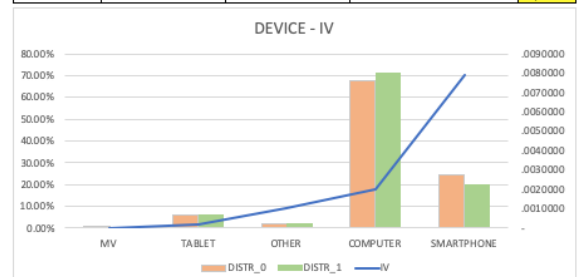


Figura 19 - Análise de dados | WOE e IV da variável Device

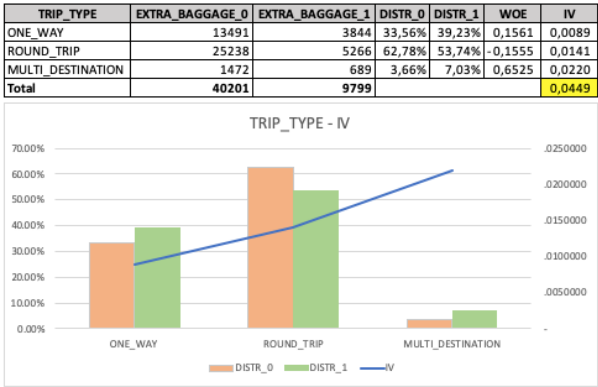


Figura 20 - Análise de dados | WOE e IV da variável Trip\_type

### TRIP\_TYPE

A soma dos valores da variável TRIP\_TYPE possuem um baixo poder preditivo, o que denota uma relação fraca com a variável resposta. No entanto, ainda é capaz de agregar algum valor ao modelo. Variável pode ser utilizada como parte do modelo preditivo.

Insight: Para esta variável a intuição se aplicou ao dizermos que MULTI\_DESTINATION requerem mais dias de viagem e, consequentemente, bagagens adicionais.

### PRODUCT

Variável de baixa IV, pouco relevante para o modelo e sem capacidade preditiva. Não deve ser levada em consideração para construção do modelo preditivo.

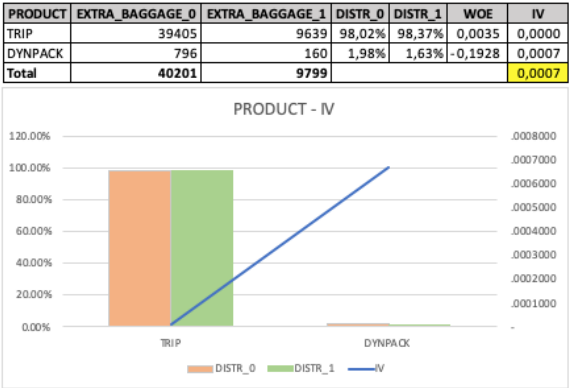


Figura 21 - Análise de dados | WOE e IV da variável Product

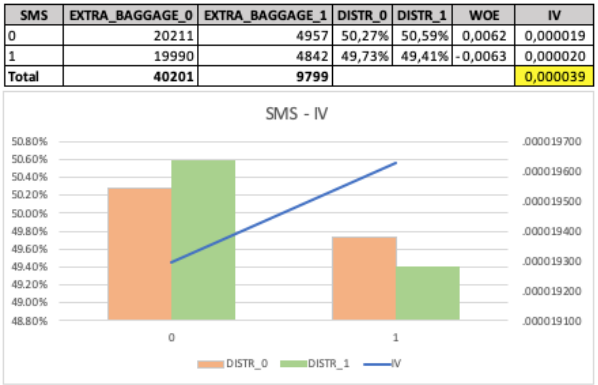


Figura 22 - Análise de dados | WOE e IV da variável SMS

### SMS

Variável de baixa IV, pouco relevante para o modelo e sem capacidade preditiva. Não deve ser levada em consideração para construção do modelo preditivo.

## NO\_GDS

As classes de NO\_GDS apresentam médio poder preditivo. NO\_GDS é a variável com maior preditividade no conjunto de dados. No entanto, nota-se que o valor “0” contempla aproximadamente 60% dos casos de passageiros que compraram bagagem adicional.

NO_GDS	EXTRA_BAGGAGE_0	EXTRA_BAGGAGE_1	DISTR_0	DISTR_1	WOE	IV
4	0	2	0,00%	0,02%	-	-
3	86	42	0,21%	0,43%	0,6949	0,0015
2	2824	1053	7,02%	10,75%	0,4251	0,0158
1	15551	5870	38,68%	59,90%	0,4373	0,0928
0	21740	2832	54,08%	28,90%	-0,6266	0,1578
<b>Total</b>	<b>40201</b>	<b>9799</b>				<b>0,2679</b>

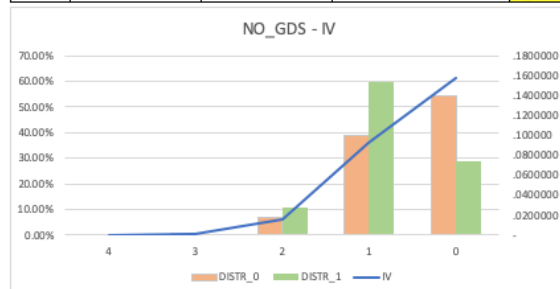


Figura 23 - Análise de dados | WOE e IV da variável NO\_GDS

## COMPANY

COMPANY	EXTRA_BAGGAGE_0	EXTRA_BAGGAGE_1	DISTR_0	DISTR_1	WOE	IV
ED	22878	5490	56,91%	56,03%	-0,0156	0,0001
GO	4876	1138	12,13%	11,61%	-0,0434	0,0002
OP	11763	2952	29,26%	30,13%	0,0291	0,0003
MV	684	219	1,70%	2,23%	0,2727	0,0015
<b>Total</b>	<b>40201</b>	<b>9799</b>				<b>0,0021</b>

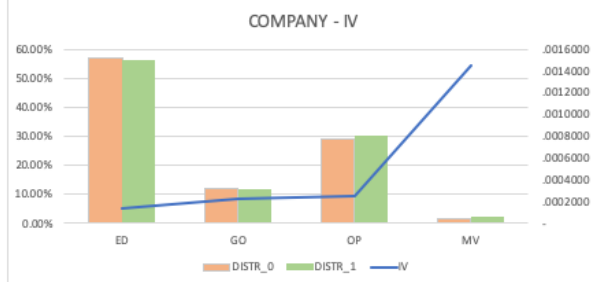


Figura 24 - Análise de dados | WOE e IV da variável Company

Variável de baixa IV, pouco relevante para o modelo e sem capacidade preditiva. Não deve ser levada em consideração para construção do modelo preditivo.

QTY_DAYS_BIN	EXTRA_BAGGAGE_0	EXTRA_BAGGAGE_1	DISTR_0	DISTR_1	WOE	IV
(46,52]	140	23	0,35%	0,23%	-0,3945	0,0004
(35,40]	185	19	0,46%	0,19%	-0,8643	0,0023
[0,6]	26953	6049	67,05%	61,73%	-0,0826	0,0044
(18,23]	927	131	2,31%	1,34%	-0,5451	0,0053
(29,35]	471	40	1,17%	0,41%	-1,0544	0,0080
(40,46]	232	10	0,58%	0,10%	-1,7325	0,0082
(12,18]	3893	1362	9,68%	13,90%	0,3614	0,0152
(23,29]	734	54	1,83%	0,55%	-1,1979	0,0153
(6,12]	5822	2084	14,48%	21,27%	0,3843	0,0261
(52,57]	844	27	2,10%	0,28%	-2,0307	0,0370
<b>Total</b>	<b>40201</b>	<b>9799</b>				<b>0,1223</b>

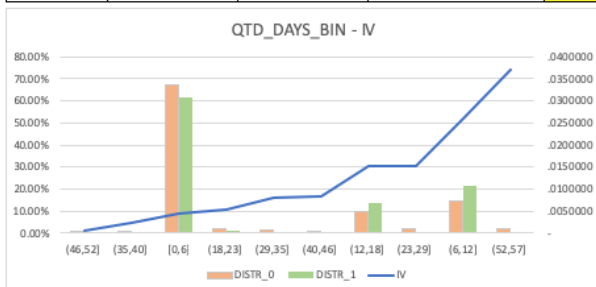


Figura 25 - Análise de dados | WOE e IV da variável Qty\_days

## QTY\_DAYS\_BIN

A soma dos valores da variável QTY\_DAYS\_BIN fazem com que esta obtenha uma capacidade preditiva com poder médio, principalmente pelos valores [0,6], [6,12] e [12,18]. Variável pode e deve ser utilizada como parte do modelo preditivo.

## COUNTRY

A variável COUNTRY possui um numero demasiado de classes, o que dificulta a utilização desta variável pelo modelo preditivo. Apesar da soma dos IV's obterem um valor que representa uma baixa capacidade preditiva, ainda acho que ao analisar as ocorrências, as classes possuem uma performance ainda pior (vide algumas classes com IV 0). Uma prática comum para reduzir a complexidade de modelos preditivos é agrupar classes com a mesma classificação de IV. O agrupamento de classes possui dois objetivos: O primeiro é reduzir classes pouco influentes (no exemplo de country poderíamos criar uma classe denominada “Outros Países” e substituímos as ocorrências dos países com IV “0” por esta classe); e o segundo é dar relevância a classes pouco relevantes, como por exemplo unificar duas classes com IV “0,1” e outra com “0,2” passando assim de duas classes com média capacidade preditiva para uma única classe com alta capacidade preditiva. Variável pode ser utilizada com cautela no modelo preditivo.

COUNTRY	EXTRA_BAGGAGE_0	EXTRA_BAGGAGE_1	DISTR_0	DISTR_1	WOE	IV
CN	3	0	0,01%	0,00%	-	-
PL	15	0	0,04%	0,00%	-	-
VE	12	0	0,03%	0,00%	-	-
IT	3879	944	9,65%	9,63%	-0,0016	0,0000
AT	71	17	0,18%	0,17%	-0,0179	0,0000
IN	5	1	0,01%	0,01%	-0,1978	0,0000
PE	5	1	0,01%	0,01%	-0,1978	0,0000
GR	6	2	0,01%	0,02%	0,3130	0,0000
RU	11	2	0,03%	0,02%	-0,2931	0,0000
CL	43	9	0,11%	0,09%	-0,1524	0,0000
SG	73	20	0,18%	0,20%	0,1169	0,0000
HK	53	11	0,13%	0,11%	-0,1608	0,0000
TH	30	9	0,07%	0,09%	0,2076	0,0000
CA	102	28	0,25%	0,29%	0,1188	0,0000
NL	163	45	0,41%	0,46%	0,1245	0,0001
MA	16	6	0,04%	0,06%	0,4308	0,0001
BR	139	27	0,35%	0,28%	-0,2270	0,0002
AE	151	29	0,38%	0,30%	-0,2384	0,0002
MX	209	41	0,52%	0,42%	-0,2172	0,0002
FR	12266	2900	30,51%	29,59%	-0,0305	0,0003
CH	278	84	0,69%	0,86%	0,2148	0,0004
EG	2	3	0,00%	0,03%	1,8171	0,0005
PH	10	7	0,02%	0,07%	1,0549	0,0005
ZA	69	9	0,17%	0,09%	-0,6253	0,0005
DE	5044	1312	12,55%	13,39%	0,0650	0,0005
CO	82	11	0,20%	0,11%	-0,5972	0,0005
NZ	44	20	0,11%	0,20%	0,6232	0,0006
ES	6287	1438	15,64%	14,67%	-0,0636	0,0006
DK	174	62	0,43%	0,63%	0,3797	0,0008
US	1787	375	4,45%	3,83%	-0,1498	0,0009
ID	4	6	0,01%	0,06%	1,8171	0,0009
TR	90	10	0,22%	0,10%	-0,7856	0,0010
SE	159	62	0,40%	0,63%	0,4698	0,0011
GB	6503	1737	16,18%	17,73%	0,0915	0,0014
AU	337	126	0,84%	1,29%	0,4278	0,0019
PT	1263	231	3,14%	2,36%	-0,2872	0,0023
NO	199	94	0,50%	0,96%	0,6616	0,0031
AR	180	96	0,45%	0,98%	0,7830	0,0042
JP	285	23	0,71%	0,23%	-1,1054	0,0052
FI	152	1	0,38%	0,01%	-3,6123	0,0133
Total	40201	9799				0,0414

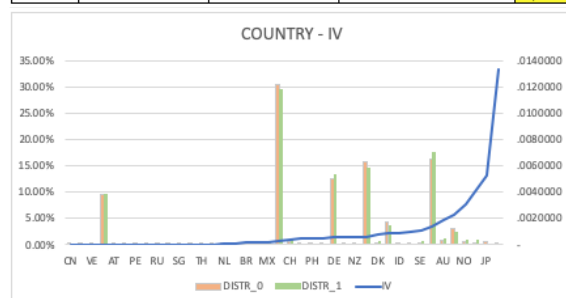


Figura 26 - Análise de dados | WOE e IV da variável Country

Ao final desta análise temos uma tabela classificatória que nos informa, em ordem de importância, a capacidade preditora das variáveis contidas no conjunto de dados, como demonstrado a seguir:

Variáveis	IV	Classificação
NO_GDS	0,26787	Médio poder preditivo
HAUL_TYPE	0,26166	
GDS	0,17919	
QTY_DAYS (BIN)	0,12232	
ADULTS	0,12027	Baixo poder preditivo
TRIP_TYPE	0,04489	
COUNTRY	0,04136	
DISTANCE (BIN)	0,02830	
TRAIN	0,02748	Sem capacidade preditiva
CHILDREN	0,01282	
DEVICE	0,01112	
INFANTS	0,00220	
COMPANY	0,00207	
PRODUCT	0,00068	
SMS	0,00004	

Durante a fase de construção dos modelos as variáveis sem capacidade preditiva serão retiradas do modelo.

Figura 27 - Análise de dados | Tabela de classificação de variáveis

## 6. Construção de Modelos de Machine Learning

Após a fase de pré-processamento e entendimento dos dados, iniciamos a construção dos modelos preditivos. Todas as fases do projeto foram realizadas na ferramenta KNIME e serão descritas mais a frente.

Modelos preditivos, de maneira geral, servem para identificar padrões ocultos, *insights* e tendências de acordo com os dados analisados de forma a prever algum fato relevante e responder perguntas relacionadas ao negócio. Existem dois tipos de modelos preditivos: os supervisionados e os não supervisionados. Nos modelos preditivos supervisionados a variável resposta está presente em todas as instâncias da base de dados de treinamento e durante a fase de treinamento o modelo preditivo aprende os padrões contidos nas variáveis preditoras. Ao contrário dos modelos preditivos supervisionados, os modelos não supervisionados recebem apenas variáveis preditoras como entrada e a variável resposta deve ser estimada. A escolha do modelo é importante e deve considerar estes fatores durante o processo. Se, por exemplo, não houver condições de associar o dado de entrada ao resultado, uma alternativa é usar o modelo não supervisionado.

Olhando para os nossos dados, a primeira decisão tomada é a de utilizar modelos preditivos supervisionados, uma vez que temos a variável resposta (dados de saída) no conjunto de dados, representada pela variável EXTRA\_BAGGAGE.

Foram testados 6 diferentes tipos de modelos na busca do melhor resultado.

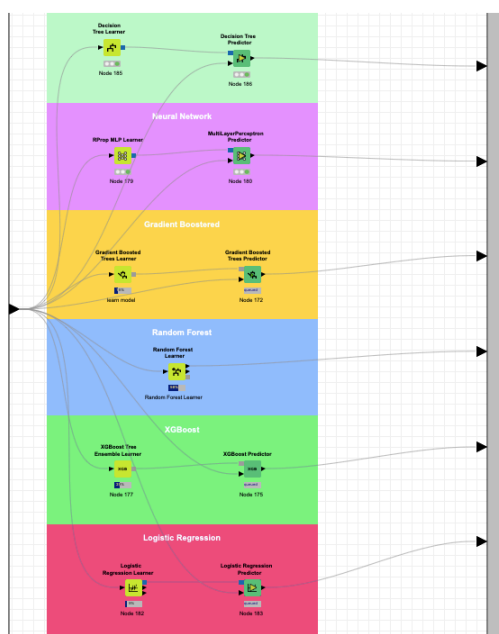


Figura 28 - KNIME Workflow | Modelos preditivos

Conforme demonstrado no quadro abaixo, vimos que os seis modelos utilizados apresentam resultados semelhantes (em torno de 80%), tomando a métrica Acurácia para medir a performance dos modelos.

Model	Accuracy
Random Forest	81,213%
Logistic Regression	80,984%
MultiLayer Predictor (Neural Network)	81,472%
Gradient Boosted	80,896%
XGBooster	81,973%
Decision Tree	81,789%

Acurácia é a soma do verdadeiro positivo e do verdadeiro negativo dividido pelo total de instâncias. Para cada linha calcula-se a precisão potencial com base nas possibilidades de nossas matrizes de confusão.

$$Acurácia = \frac{Verdadeiros\ Positivos\ (TP) + Verdadeiros\ Negativos\ (VN)}{Total}$$

Interpretamos que todos os modelos acima apresentam ótimos resultados de acurácia. No entanto, devemos desconfiar de acurácias tão altas pura e simplesmente.

Pegamos a matriz confusão do modelo Random Forest, apenas para usarmos como exemplo e demonstrarmos a necessidade de se utilizar métricas complementares para avaliar a performance real dos modelos.

EXTRA BAGAGE	FALSE	TRUE
FALSE	29700	486
TRUE	6592	722

Importante relembrarmos a pergunta a ser respondida pelo modelo, que é identificar os passageiros com alto potencial de compra de bagagem adicional. Se olharmos novamente para a matriz de confusão acima podemos identificar que a alta acurácia se deu pela célula FF, ou seja, o modelo previu - majoritariamente - os passageiros que NÃO compraram bagagem adicional. A taxa de acerto real dos TT's ficou por volta dos 9%, conforme detalhado nos cálculos abaixo:



Comprou bagagem extra = TF + TT = 6592 + 722 = 7314

NÃO comprou bagagem extra = FF + FT = 29700 + 486 = 30186

Acerto FF = FF / Total = 29700 / 30186 = 0,983 ou 98,38%

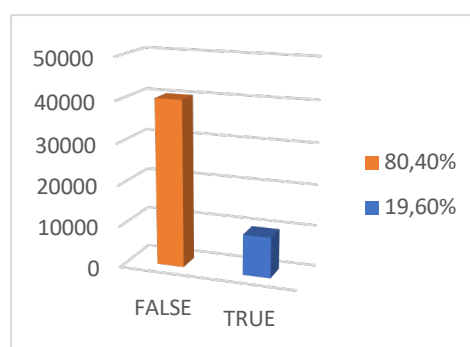
Acerto TT = TT / Total = 722 / 7314 = 0,0987 ou 9,87%

Mais detalhes a respeito dos cálculos realizados acima serão disponibilizados mais a frente.

Observe os valores da célula da categoria FF e perceberá uma grande quantidade de dados nessa categoria e poucos dados na categoria TT, que é a categoria de maior interesse neste trabalho. Reforçando esta afirmação, podemos verificar no gráfico e tabela abaixo, que a quantidade de passageiros que não compraram bagagem adicional é de aproximadamente 4 vezes maior dos passageiros que compraram. Com isso, concluímos que a nossa amostra está desbalanceada.

Row ID	S	EXTRA_BAGGAGE	I	Count...	D	Percent...
Row0	0			40201		80.402
Row1	1			9799		19.598

Figura 29 - Gráfico e tabela de balanceamento da amostra



Percebe-se com a tabela abaixo que apesar da acurácia elevada o índice real de acerto, ou seja, o percentual de acerto de passageiros que comprariam bagagem, variou entre 9,8% a 16,2% nos modelos preditivos desenvolvidos anteriormente.

Model	Accuracy	Qty True (Real)	Qty True (Predicted)	Accuracy (True)
Random Forest	81,213%	7300	716	9,808%
Logistic Regression	80,984%	7300	743	10,178%
MultiLayer Predictor (Neural Network)	81,472%	7300	825	11,301%
Gradient Boosted	80,896%	7300	1040	14,247%
XGBooster	81,973%	7300	1112	15,233%
Decision Tree	81,789%	7300	1177	16,123%

Estamos diante de um dos maiores problemas/desafios de *Machine Learning*, que é desenvolver soluções baseadas em amostras desbalanceadas. Os

especialistas em *machine learning* recomendam e sugerem diversas técnicas para balanceamento de amostras.

Amostras não balanceadas representam um problema para criação de modelos analíticos, de acordo com Dal Pozzolo, A., Caelen, O., Johnson, R. A., & Bontempi, G. (Dezembro, 2015), que diz que uma classificação desbalanceada é um exemplo de problema de classificação em que a distribuição de instâncias entre as classes conhecidas é enviesada ou distorcida. A distribuição pode variar de um pequeno viés a um grave desequilíbrio, onde existe uma instância na classe minoritária para centenas, milhares ou milhões de instâncias na(s) classe(s) majoritária(s).

Amostras desbalanceadas representam um desafio para a modelagem preditiva, pois a maioria dos algoritmos de aprendizado de máquina usados para classificação foram projetados com base no pressuposto de um número igual de instâncias para cada classe. Isso resulta em modelos com baixo desempenho preditivo, especificamente para a classe minoritária. Esse é um problema porque, normalmente, a classe minoritária é mais importante e, portanto, o problema é mais sensível aos erros de classificação da classe minoritária do que a classe majoritária. Um exemplo deste fenômeno são os modelos de fraude para seguradoras, onde deseja-se identificar a classe minoritária, que são os fraudadores.

O problema de aprender a partir de conjuntos de dados desbalanceados tem sido estudado por vários autores (Pazzani e Brink, 1994)(Ling e Li, 1998)(Kubat e Matwin, 1997)(Fawcett e Provost, 1997)(Weiss, 2004)(Han e Mao, 2005). As diversas abordagens estudadas nestes trabalhos estão divididas em duas linhas de pesquisa: Pré processamento de dados e adaptação de algoritmos. Para resolver problemas relacionados a amostras desbalanceadas, o autor recomenda utilizarmos uma técnica que consiste em balancear as amostras de forma a garantir a mesma quantidade de instâncias para os valores da variável resposta.

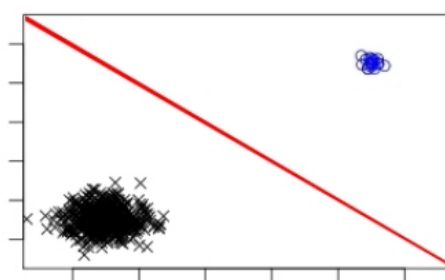


Figura 30 - Representação gráfica de amostra desbalanceada

O balanceamento realizado no pré-processamento de dados consiste no equilíbrio na distribuição das classes no conjunto de dados por meio de mecanismos que

não alterem a distribuição original. O método utilizado consiste em remover elementos da classe majoritária (no caso EXTRA\_BAGGAGE=0) de forma a promover o balanceamento e, consequentemente, o seu equilíbrio. Observe na figura a seguir que, após o balanceamento dos dados, a classe minoritária (representada pelo ponto preto) possui o mesmo formato / tamanho da classe majoritária (representada pelo ponto azul). Outro ponto importante é que a localização das classes no gráfico permaneceu inalterada de forma a garantir a distribuição dos dados.

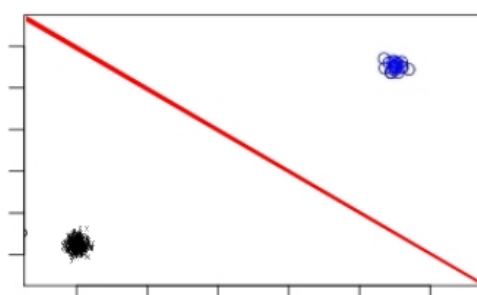


Figura 31 - Representação gráfica de amostra balanceada

Para a implementação desta técnica no KNIME, utilizamos o objeto “*Equal Size Sampling*”, que tem por objetivo remover – de forma aleatória - linhas do conjunto de dados de treinamento para que os valores da variável resposta sejam igualmente distribuídos, ou seja, possuam a mesma quantidade de instâncias. As linhas retornadas por este nó conterão todos os registros da classe minoritária e uma amostra aleatória de cada uma das classes majoritárias, na qual cada amostra contém tantos objetos quanto a classe minoritária.

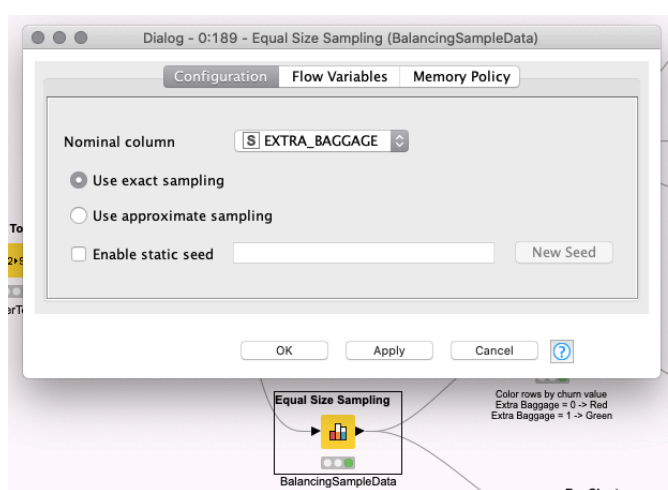


Figura 32 - KNIME Workflow | Configuração do objeto "Equal Size Sampling"

O resultado comparativo da amostra balanceada v.s. amostra desbalanceada com relação a variável resposta foi:

## Antes

Row ID	S	EXTRA_BAGGAGE	I	Count...	D	Percent...
Row0	0			40201		80.402
Row1	1			9799		19.598

Figura 33 - Quantitativo de instancias desbalanceadas

## Depois

Row ID	S	EXTRA_BAGGAGE	I	Count...	D	Perce...
Row0	0			9799		50
Row1	1			9799		50

Figura 34 - Quantitativo de instancias balanceadas

Em resumo, amostra balanceada permite que modelo analítico seja treinado corretamente levando em consideração a mesma quantidade de amostras para cada valor da variável resposta.

Assim, com o devido tratamento no desbalanceamento da amostra, prosseguimos com o desenvolvimento do modelo. Reaplicamos os modelos desenvolvidos anteriormente à amostra balanceada e o modelo que apresentou melhor performance foi o XGBoost Predictor. Sendo assim, seguimos apenas com este modelo, conforme pode-se notar na figura do modelo completo desenvolvido no KNIME, demonstrado a seguir:

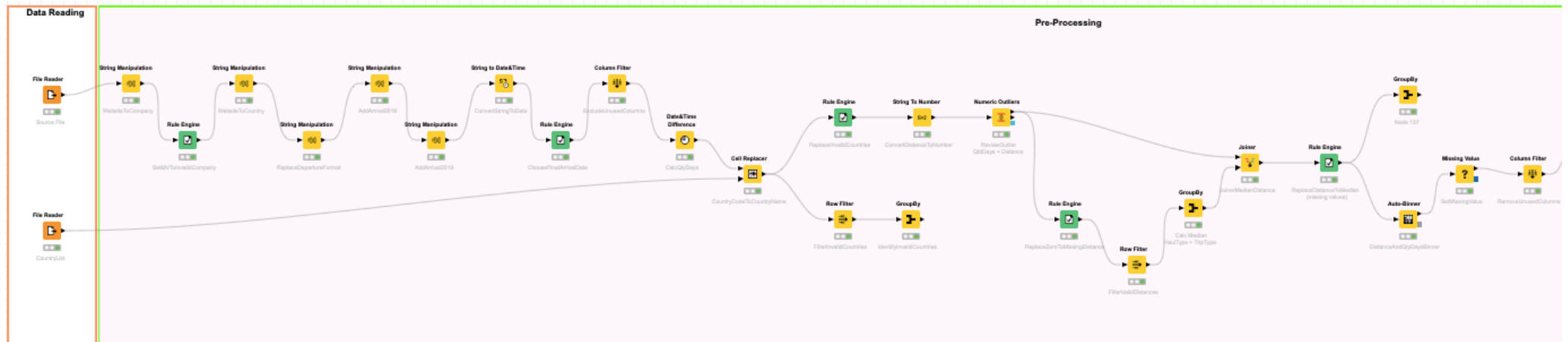


Figura 35 - KNIME Workflow | Visão geral do modelo (Parte 1)

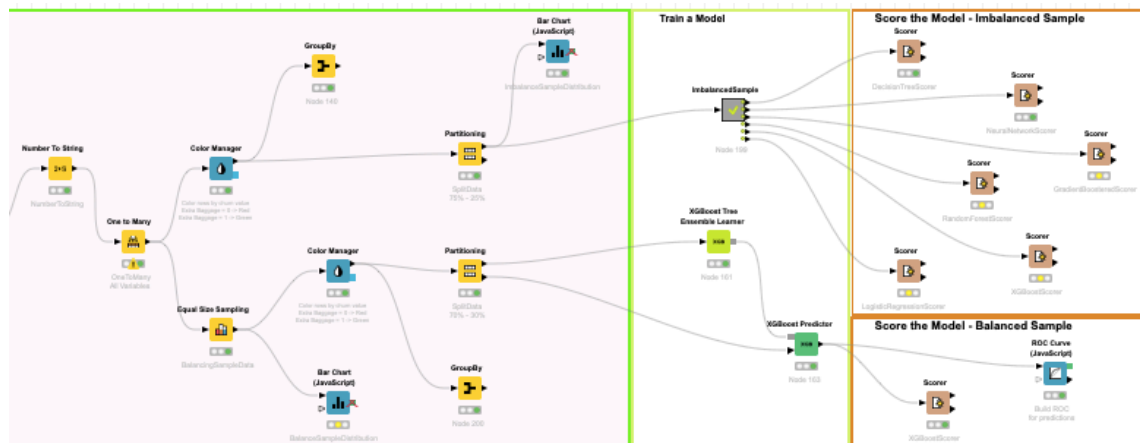


Figura 36 - KNIME Workflow | Visão geral do modelo (Parte 2)

## 7. Apresentação dos Resultados

Conforme vimos anteriormente, as amostras desbalanceadas têm um impacto negativo nos modelos preditivos e para evitarmos estes problemas, procedemos ao balanceamento da amostra. Esta é a principal contribuição deste trabalho, que é mostrar o impacto negativo das amostras desbalanceadas na construção de modelos preditivos e nos resultados distorcidos que podem ser – erroneamente – utilizados para tomada de decisão.

Este é um problema recorrente e que muitos cientistas de dados negligenciam no processo de desenvolvimento de modelos preditivos, causando resultados falso-positivos.

O desbalanceamento de dados é muito mais frequente do que se imagina. A escolha do cientista de dados pela técnica correta para tratamento das amostras pode ser uma tarefa difícil em alguns casos, especialmente quando se quer identificar padrões de comportamento dos dados para eventos cuja ocorrência seja baixa com relação ao total da amostra. É preciso ser cuidadoso no balanceamento de forma a garantir que os padrões e a distribuição das ocorrências das variáveis preditoras não seja alterada pois desta forma estaríamos criando novos cenários que dariam origem a previsões incompletas ou equivocadas.

Para evitar incorrer nesse tipo de erro, os autores JM. Lobo, A. Jiménez-Valverde, e R. Real 2008 recomendam usar outras métricas como, por exemplo, f1-score pois são mais apropriadas para avaliar a qualidade ou acuracidade real dos modelos preditivos com maior segurança e confiabilidade do que a acurácia pura e simplesmente, que somente pode ser interpretado diante de amostras balanceadas.

Não podemos falar de f1-score sem falar das métricas Precisão e Recall pois estas estão interligadas, como veremos mais a frente.

É importante entendermos o cálculo e objetivo da métrica Precisão para evitarmos que ela seja confundida com Acurácia, o que não é raro de acontecer. A acurácia pretende responder o quão frequente o classificador está correto e a precisão mostra daqueles que foram classificados corretamente, quais efetivamente eram corretos.

Através da matriz de confusão do modelo de pior performance obtido com as amostras desbalanceadas, podemos diferenciar facilmente as duas métricas.

EXTRA_BAGGAGE...	0	1
0	29700	486
1	6592	722

Figura 37 - Matriz Confusão modelo desbalanceado (Random Forest)

Ou seja:

$$\text{Acurácia} = \frac{\text{Verdadeiros Positivos (TP)} + \text{Verdadeiros Negativos (VN)}}{\text{Total}}$$

$$\text{Acurácia} = (722 + 29700) / 37500 = 0,811 \text{ ou } 81,1\%$$

$$\text{Precisão} = \frac{\text{Verdadeiros Positivos (TP)}}{\text{Verdadeiros Positivos (TP)} + \text{Falsos Positivos (FP)}}$$

$$\text{Precisão} = 722 / (722 + 486) = 0,597 \text{ ou } 59,7\%$$

Já o Recall é a frequência em que o seu classificador encontra os exemplos de uma classe, ou seja, “quando realmente é da classe X, o quão frequente você classifica como X?”. Traduzindo para fórmula:

$$\text{Recall} = \frac{\text{Verdadeiros Positivos (TP)}}{\text{Verdadeiros Positivos (TP)} + \text{Falsos Negativos (FN)}}$$

Ainda utilizando o exemplo acima, podemos calcular o recall da seguinte forma:

$$\text{Recall} = 722 / (722 + 6592) = 0,0987 \text{ ou } 9,87\%$$

Note que as métricas de Acurácia, Recall ou Precisão – se utilizadas sozinhas ou descontextualizadas - não apenas mostram um resultado equivocado como também permitem que o cientista de dados afira *insights* e premissas que podem leva-lo a conclusões ineficientes. Em outras palavras, as métricas precisam ser adequadas ao contexto da pesquisa, principalmente quando utilizadas em modelos desbalanceados. Tão importante quando saber escolher o melhor modelo, é saber escolher a métrica mais adequada para medir a qualidade do resultado obtido.

Já o *f1-score* combina precisão e recall de modo a trazer um valor único que indique a qualidade geral do modelo e se aplica – inclusive – a instâncias com classes desbalanceadas. A fórmula que define o *f1-score* é a seguinte (quanto maior melhor o modelo):

$$F1 = \frac{2 * \text{precisão} * \text{recall}}{\text{precisão} + \text{recall}}$$

Utilizando o exemplo anterior teríamos:

$$F1 = (2 * 0,597 * 0,0987) / (0,597 + 0,0987) = 0,1178 / 0,6957 = 0,169 \text{ ou } 16,9\%$$

Conforme já foi dito anteriormente, recall e precisão levam em consideração os acertos (TT ou FF) e por isso estas medidas representam melhor a realidade para qualquer tipo de amostra: balanceada ou desbalanceada. O valor *f1-score* igual a 0,1693 mostra que o modelo é pobre ou, em outras palavras, acerta aproximadamente 16% das classes FF e TT. Isso mostra a importância de se utilizar a métrica correta para cada tipo de modelo.

Já para as amostras balanceadas, obtivemos acurácia de 69,55% que, apesar de inferior ao resultado obtido com as amostras desbalanceadas, apresenta taxa de acerto de 74,88% (acurácia real dos TT), muito superior aos resultados obtidos anteriormente. A tabela abaixo apresenta a acurácia obtida nos modelos desbalanceados em comparação com o modelo final, balanceado.

Model	Accuracy	Qty True (Real)	Qty True (Predicted)	Accuracy (True)
Random Forest	81,213%	7300	716	9,808%
Logistic Regression	80,984%	7300	743	10,178%
MultiLayer Predictor (Neural Network)	81,472%	7300	825	11,301%
Gradient Boosted	80,896%	7300	1040	14,247%
XGBooster	81,973%	7300	1112	15,233%
Decision Tree	81,789%	7300	1177	16,123%
<b>XGBooster (Balanced)</b>	<b>69,558%</b>	<b>2919</b>	<b>2186</b>	<b>74,889%</b>

Nota-se que apesar de uma acurácia inicial menor, a taxa de acerto de passageiros que obtiveram bagagem adicional foi muito superior a todos os demais modelos. Isso mostra que o modelo foi treinado corretamente quando a amostra foi balanceada e que o resultado obtido é um bom resultado, ou seja, o modelo consegue prever em quase 75% passageiros que contrataram bagagens extras, o que



representa um bom número em termos percentuais e pode representar um bom número em termos financeiros para a empresa.

Portanto, a matriz confusão do modelo XGBooster com a amostra balanceada ficou:

Row ID	0	1
0	1950	1011
1	779	2140

Figura 38 - Matriz Confusão modelo balanceado (XGBooster)

Se olharmos para as demais métricas do modelo balanceado teremos:

#### Precisão

$$\text{Precisão} = \frac{\text{Verdadeiros Positivos (TP)}}{\text{Verdadeiros Positivos (TP)} + \text{Falsos Positivos (FP)}}$$

$$\text{Precisão} = 2140 / (2140 + 1011) = 0,679 \text{ ou } 67,9\%$$

#### Recall

$$\text{Recall} = \frac{\text{Verdadeiros Positivos (TP)}}{\text{Verdadeiros Positivos (TP)} + \text{Falsos Negativos (FN)}}$$

$$\text{Recall} = 2140 / (2140 + 779) = 0,733 \text{ ou } 73,3\%$$

#### F1-Score

$$F1 = \frac{2 * \text{precisão} * \text{recall}}{\text{precisão} + \text{recall}}$$

$$F1 = (2 * 0,679 * 0,733) / (0,679 + 0,733) = 0,995 / 1,412 = 0,7046 \text{ ou } 70,46\%$$

#### Curva AUC-ROC

A curva ROC (oriunda do termo em inglês “*receiver operating characteristic curve*”) mostra o desempenho de um modelo de classificação em todos os limiares de classificação. Essa curva plota dois parâmetros: Taxa positiva verdadeira e taxa de falsos positivos. A curva ROC, de acordo com o artigo “*Comprehension of the AUC-ROC curve*” escrito por Kurtis Pykes, é uma representação gráfica capaz de ilustrar e

discriminar as classes em vários limites, plotando a taxa positiva verdadeira (TPR) contra a taxa positiva falsa (FPR) nas várias configurações de limite entre 0 e 1, que é exatamente a curva ROC. A Taxa Positiva Verdadeira é uma medida do número de positivos reais que são corretamente identificados como tais nas previsões do nosso modelo. Isso é importante porque, quando um classificador faz previsões queremos saber a proporção de previsões corretas feitas com todos os rótulos corretos reais presentes em nossos dados. Em contrapartida, a quantidade de “alarmes falsos” que classificamos é conhecida como Taxa de Falso Positivo. A leitura da curva ROC é bastante simples, quanto mais próximo de 1 melhor é a ROC.

$$\text{TPR} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

Figura 39 - Fórmula para cálculo da Taxa Positiva Verdadeira (ROC)

$$\text{FPR} = \frac{\text{False Positive}}{\text{False Positive} + \text{True Negative}}$$

Figura 40 - Fórmula para cálculo da Taxa Falso Positivo (ROC)

O modelo XGBooster com as amostras balanceadas resultou em um ROC de 0,729 para EXTRA\_BAGGAGE =1 (gerado diretamente no KNIME), conforme pode ser interpretado pelo gráfico abaixo:

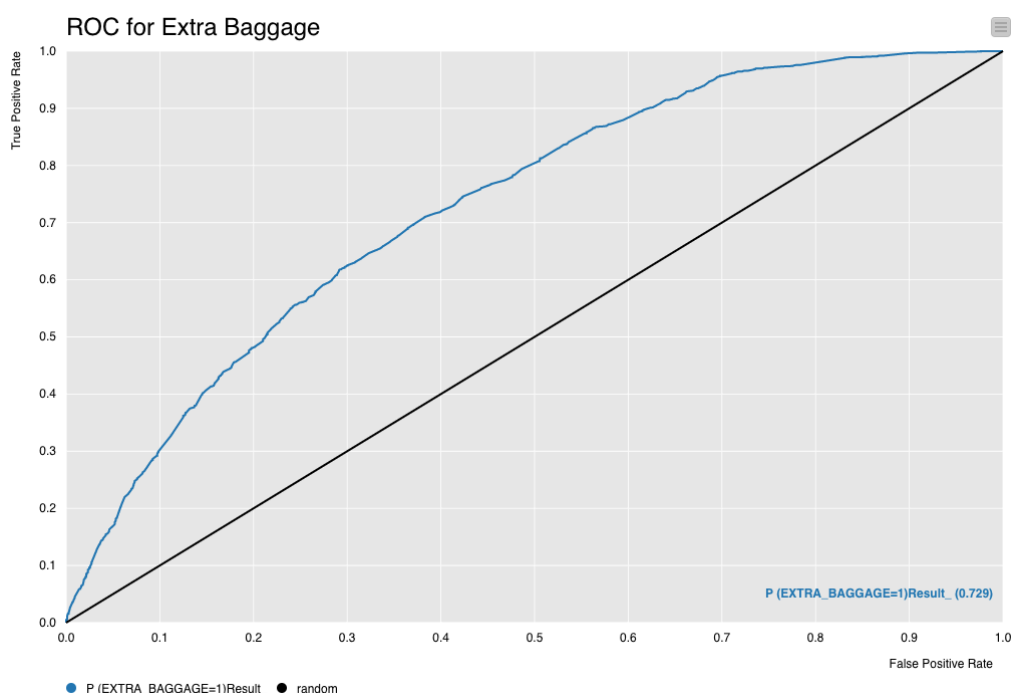


Figura 41 - Curva ROC do modelo preditivo final

Se realizarmos uma análise global sobre os resultados alcançados, após o balanceamento e construção do modelo obtivemos números satisfatórios que refletem uma boa capacidade preditiva do modelo. Rawat Shubhankar em seu artigo “*Is Accuracy EVERYTHING?*” corrobora com a análise deste trabalho ao dizer que a acurácia sozinha não determina a qualidade do modelo preditivo, principalmente se este for construído com base em amostras desbalanceadas. O autor reforça em seu artigo que as métricas de precisão, recall e f1-score são as mais adequadas para avaliar a qualidade dos modelos preditivos.

## 8. Links

Abaixo estão listados todos os arquivos utilizados no processo de desenvolvimento deste projeto. Os arquivos estão disponíveis em:

[https://github.com/phramos/PUCMG\\_TCC](https://github.com/phramos/PUCMG_TCC)

Arquivo	Descrição
PREDICTING BAGGAGE LIKELIHOOD 5.zip	Arquivo compactado contendo todos os objetos KNIME utilizados para desenvolvimento do modelo.
AED.xlsx	Arquivo Excel utilizado para desenvolvimento da Análise e Exploração dos Dados
country.csv	Lista de siglas e nomes dos Países
train.csv	Arquivo contém as 50.000 instâncias utilizadas no projeto.
TCC_Breafing.mov	Video de apresentação do projeto

Em função do tamanho do vídeo de breafing não foi possível – até o momento – publicá-lo no GitHub. Neste caso, o vídeo pode ser encontrado em:

[https://www.icloud.com/iclouddrive/03uecJTFe-mhKcVz4qG1wtU7A#TCC\\_Breafing](https://www.icloud.com/iclouddrive/03uecJTFe-mhKcVz4qG1wtU7A#TCC_Breafing)

## 9. Referências

Aggarwal C. (2015) Outlier Analysis. In: Data Mining. Springer, Cham

Larose, C., Dey, D. K., & Harel, O. (2019). The impact of missing values on different measures of uncertainty. *Statistica Sinica*, 29(2), 551-566.

Zeng, G. (2014). A necessary condition for a good binning algorithm in credit scoring. *Applied Mathematical Sciences*, 8(65), 3229-3242.

Dal Pozzolo, A., Caelen, O., Johnson, R. A., & Bontempi, G. (2015, December). Calibrating probability with undersampling for unbalanced classification. In *2015 IEEE Symposium Series on Computational Intelligence* (pp. 159-166). IEEE.

Y. Wang, H. Tercan, T. Thiele, T. Meisen, S. Jeschke and W. Schulz, "Advanced data enrichment and data analysis in manufacturing industry by an example of laser drilling process," 2017 ITU Kaleidoscope: Challenges for a Data-Driven Society (ITU K), Nanjing, 2017, pp. 1-5.

JM. Lobo, A. Jiménez-Valverde, and R. Real 2008 - AUC: a misleading measure of the performance of predictive distribution models

<https://link.medium.com/whCzJmPzH4>

[https://link.springer.com/chapter/10.1007/978-3-319-14142-8\\_8](https://link.springer.com/chapter/10.1007/978-3-319-14142-8_8)

<https://machinelearningmastery.com/bagging-and-random-forest-for-imbalanced-classification/>

<https://machinelearningmastery.com/one-class-classification-algorithms/>

<https://machinelearningmastery.com/what-is-imbalanced-classification/>

<https://medium.com/@SankaraJ/how-we-handle-missing-values-affects-the-accuracy-of-machine-learning-models-c120310141b5>

<https://medium.com/@swethalakshmanan14/outlier-detection-and-treatment-a-beginners-guide-c44af0699754>

<https://pl.wikipedia.org/wiki/KNIME>

<https://pt-pt.khanacademy.org/math/statistics-probability/summarizing-quantitative-data/box-whisker-plots/a/identifying-outliers-iqr-rule>

<https://towardsdatascience.com/a-deep-dive-into-imbalanced-data-over-sampling-f1167ed74b5>

<https://towardsdatascience.com/attribute-relevance-analysis-in-python-iv-and-woe-b5651443fc04>

<https://towardsdatascience.com/dealing-with-imbalanced-classes-in-machine-learning-d43d6fa19d2>

<https://towardsdatascience.com/from-zero-to-hero-in-xgboost-tuning-e48b59bfaf58>

<https://towardsdatascience.com/how-to-calibrate-undersampled-model-scores-8f3319c1ea5b>

<https://towardsdatascience.com/probability-calibration-for-imbalanced-dataset-64af3730eaab>

<https://towardsdatascience.com/is-accuracy-everything-96da9afd540d>

<https://towardsdatascience.com/comprehension-of-the-auc-roc-curve-e876191280f9>