PARTE II

PRÁTICA

11

APLICAÇÃO 1: RECLAMAÇÕES POR PERDAS

Neste capítulo, veremos como aplicar os conhecimentos obtidos ao conjunto de dados chamado *TSA Claims Data* (Dados de Reclamações TSA). A sigla TSA se refere à agência americana *Transportation Security Administration* (Administração de Segurança de Transporte), responsável por supervisionar o sistema de transporte nos Estados Unidos. Esse conjunto de dados, que pode ser obtido no link `https://www.dhs.gov/tsa-claims-data` do próprio governo norte-americano, foca nas reclamações de clientes em aeroportos. Analisaremos as reclamações registradas no ano de 2015 – no site, o arquivo está nomeado como *Claims Data 2015 (as of Feb. 9, 2016).xlsx*. O objetivo dessa análise não é determinar qual é a melhor ou pior companhia aérea (ou aeroporto), mas sim extrair informações dos dados e recomendar tomadas de decisão para clientes. Começaremos realizando análises descritivas para entender os dados. Em seguida, aplicaremos técnicas de pré-processamento de dados para garantir que a análise seja pertinente e que não haja dados discrepantes.

11.1 Análise Descritiva

O conjunto de dados possui, ao todo, 8667 registros de reclamações descritas por 11 variáveis, sendo uma quantitativa e as demais qualitativas. O nome, a descrição, os possíveis valores e a quantidade de

ocorrências de cada valor que uma variável pode assumir estão listados a seguir:

- *Claim Number*: o número da reclamação, funciona como um identificador único, portanto todos os valores são diferentes.
- *Date Received*: a data que a reclamação foi registrada.
- *Incident D*: data na qual foi relatado o incidente.
- *Airport Code*: código do aeroporto onde aconteceu o incidente. Ao total, foram registrados incidentes em 301 aeroportos.
- *Airport Name*: nome do aeroporto.
- *Airline Name*: nome da companhia aérea relacionada à reclamação. O conjunto de dados contém informações de 119 companhias aéreas.
- *Claim Type*: tipo de reclamação. São 7 tipos de reclamações registradas: *Passenger Property Loss* (Perda de propriedade do passageiro, com 4551 registros), *Property Damage* (Danos materiais, com 3888 registros), *Personal Injury* (Danos pessoais, com 122 registros), *Motor Vehicle* (Veículo, com 35 registros), "−" (indefinido, com 34 registros), *Complaint* (Reclamação, com 28 registros) e *Employee Loss - MPCECA* (Perda de funcionário, com 9 resgistros)
- *Claim Site*: local onde aconteceu o incidente apontado pela reclamação. Ao todo, são 6 locais: *Checked Baggage* (Bagagem despachada, 6261), *Checkpoint* (Ponto de verificação, 2293), *Motor Vehicle* (Veículo, 49), "−" (indefinido, 39) *Other* (outro, 24), *Bus Station* (Estação de ônibus, 1)
- *Item Category*: classificação do item que o cliente relatou na reclamação. Ao total, são 567 tipos de itens registrados. Os mais frequentes são: *Baggage/Cases/Purses* (bagagens/estojos/bolsas, 1004), *Computer e Accessories* (Computador e acessórios, 736), *Clothing* (vestuário, 723)
- *Close Amount*: quantia aproximada do valor do item relatado na reclamação.

- *Disposition*: decisão final após a avaliação da reclamação do cliente. Existem 4 possíveis decisões no conjunto de dados: *Deny* (negar, 3574), "−" (indefinido, 2066), *Approve in Full* (aprovar na íntegra, 1958), *Settle* (pagar o valor devido, 1069).

Nesse conjunto de dados, a variável *Close Amount* pode ser uma possível variável alvo, assim como *Disposition*. A Tabela 11.1 mostra 4 estatísticas: média, primeiro quartil (Q1), mediana (Q2) e o terceiro quartil (Q3) para a variável *Close Amount*, separando o conjunto de dados de acordo com a variável *Disposition*: o conjunto completo de dados, todas as reclamações, exceto as que foram negadas (*Deny*), todas as reclamações completamente aprovadas e as reclamações com valores pagos.

Tabela 11.1: Estatísticas sobre o conjunto de dados completo.

Valores	**Média**	**Q1**	**Q2**	**Q3**
Todos	87,50	0,00	0,00	89,23
Todos, exceto *Deny*	191,25	45,00	100,00	211,98
Approve in Full	159,37	34,93	80,56	181,89
Approve in Settle	254,97	65,00	150,00	300,00

Podemos notar que a média de todos os valores pagos é relativamente baixa quando comparada com outras decisões. Por outro lado, para decisões do tipo *Settle* (pagar o valor devido), todas as quatro estatísticas apresentaram valores mais altos. Isso mostra que, provavelmente, você não receberá nada pela reclamação registrada. Entretanto, se sua reclamação for acatada e houver uma decisão de pagar o valor devido (*Settle*), é provável que você receba, em média, 254,97 dólares. Além disso, apenas 25% das reclamações foram acatadas e resultaram em valores pagos acima de 300 dólares.

A Figura 11.1 apresenta boxplots que mostram o comportamento dos valores de *Close Amount* para cada um dos possíveis tipos de reclamação, conforme a variável *Disposition*.

É possível perceber que, independentemente da decisão tomada após a reclamação do cliente, valores acima de 500 dólares são considerados fora do padrão (ou *outliers*). Valores próximos de 5000 são

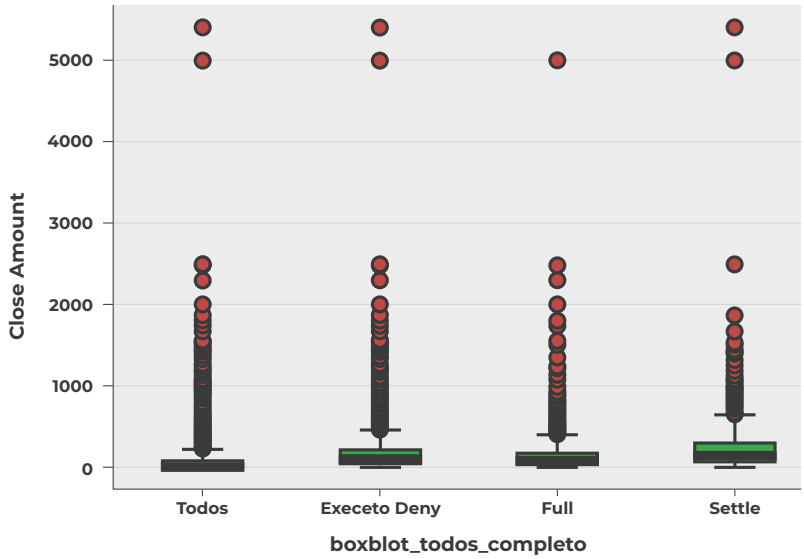

Figura 11.1: Boxplot para cada um dos possíveis tipos de reclamação.

possíveis de encontrar, mas são improváveis. Pode-se afirmar que valores entre 0 e 650 dólares (limite superior aproximado do boxplot para *Settle*) são mais prováveis de serem registrados em uma reclamação de perda de itens nesses aeroportos.

11.2 Pré-processamento

Ao analisar os valores que as variáveis podem assumir, é possível notar que alguns desses valores estão registrados como "−", o que indica a ausência de informações. Assim, para que a extração de dados seja mais eficiente, todas as reclamações que tinham o valor "−" em algumas das variáveis foram removidas. Mais especificamente, as seguintes variáveis apresentaram esse valor: *Close Amount*, *Claim Type*, *Claim Site* e *Airline Name*. Após essa etapa, o conjunto de dados passou a conter 6231 reclamações.

Outra etapa realizada no pré-processamento foi a seleção das cinco companhias aéreas mais listadas no conjunto de dados. O objetivo foi focar no estudo das companhias que potencialmente poderiam gerar alguma insatisfação do serviço prestado, conforme relatado pelos cli-

entes. As companhias aéreas selecionadas e o número de reclamações correspondentes presentes no conjunto de dados são: Southwest Airlines (1213), Delta Air Lines (1103), American Airlines (867), UAL (858) e USAir (501). Assim, após essa etapa, o conjunto de dados passou a conter 4542 reclamações. Isso significa que as companhias aéreas selecionadas representam 72,89% do total de reclamações registradas.

A Tabela 11.2 mostra as estatísticas de média, primeiro quartil (Q1), mediana (Q2) e terceiro quartil (Q3) para a variável *Close Amount*, separando o conjunto de dados de acordo com a variável *Disposition*, após a etapa de pré-processamento.

Tabela 11.2: Estatísticas sobre o conjunto de dados após pré-processamento.

Valores	**Média**	**Q1**	**Q2**	**Q3**
Todos	40,65	0,00	0,00	56,15
Exceto *Deny*	98,53	37,00	77,55	149,99
Approve in Full	89,70	29,75	68,75	132,71
Approve in Settle	119,15	50,00	100,00	175,00

11.3 Análise das Reclamações

Com essa análise, buscamos identificar quais características das reclamações (variáveis) mais contribuem para um cliente ter sua queixa aceita ou não. Para isso, vamos dividir nosso conjunto de reclamações em duas grandes classes: *Deny* e *Not-deny* (todas as reclamações, exceto *Deny*). Além disso, removeremos as variáveis que não trazem nenhuma contribuição para a extração de informação (*Claim Number*, por exemplo) ou variáveis com semânticas semelhantes (*Airport Code* e *Airport Name*). Ao final, temos as seguintes variáveis: *Airline Name*, *Claim Type*, *Claim Site* e *Disposition*.

Para extrair as contribuições das variáveis para a classificação, utilizaremos os algoritmos de Aprendizagem de Máquina: Regressão Logística, Árvore de Decisão e Floresta Aleatória. Como esses algoritmos usam variáveis numéricas e as variáveis do conjunto de dados são categóricas, faremos uma transformação dessas variáveis usando a função

"pandas.get_dummies" em Python. O conjunto de dados foi dividido em treinamento e teste, na proporção de $70 - 30\%$, respectivamente.

Após a construção de um modelo usando o conjunto de treinamento, os resultados da execução do modelo no conjunto de teste são obtidos. A Figura 11.2 mostra a importância das variáveis identificada pela Regressão Logística.

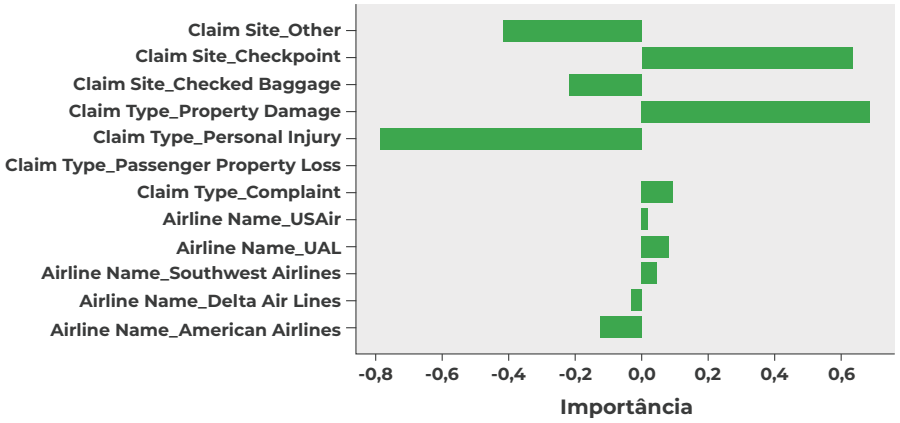

Figura 11.2: Importância das variáveis na Regressão Logística.

Podemos observar que o local "checkpoint" e o tipo de reclamação "property damage" são as variáveis que mais contribuem para não negar a quantia declarada na reclamação. Por outro lado, o local "other" e o tipo de reclamação "personal injury" contribuem para a negação. A Árvore de Decisão confirma a importância das variáveis "checkpoint" e "property damage".

Além das variáveis "checkpoint" e "property damage", Floresta Aleatória também considerou como importantes as variáveis "checked baggage" e "passenger property loss".

Além da análise da importância das variáveis, outra ferramenta interessante para o cientista de dados é chamada Shap. O principal objetivo dessa abordagem é explicar a saída de qualquer modelo de Aprendizagem de Máquina, fazendo uma relação entre as variáveis de entrada e o valor predito pelo modelo. Isso é importante para compreender o aprendizado obtido por um modelo, mesmo que ele seja do tipo caixa-preta. A relação entre as variáveis e o valor predito deixa

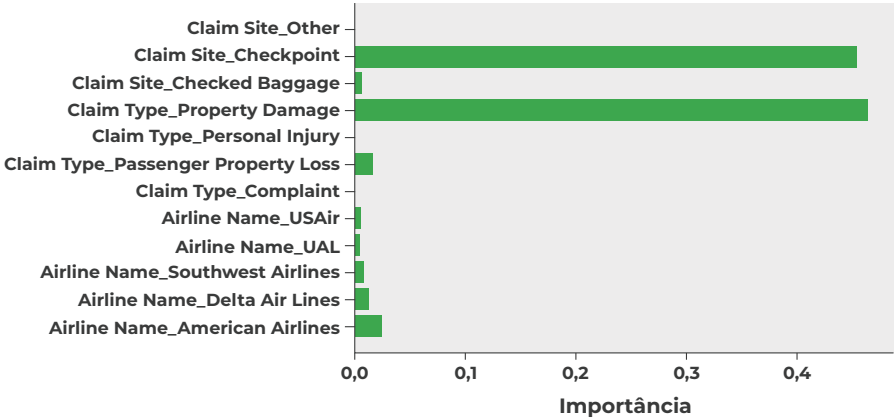

Figura 11.3: Importância das variáveis na Árvore de Decisão.

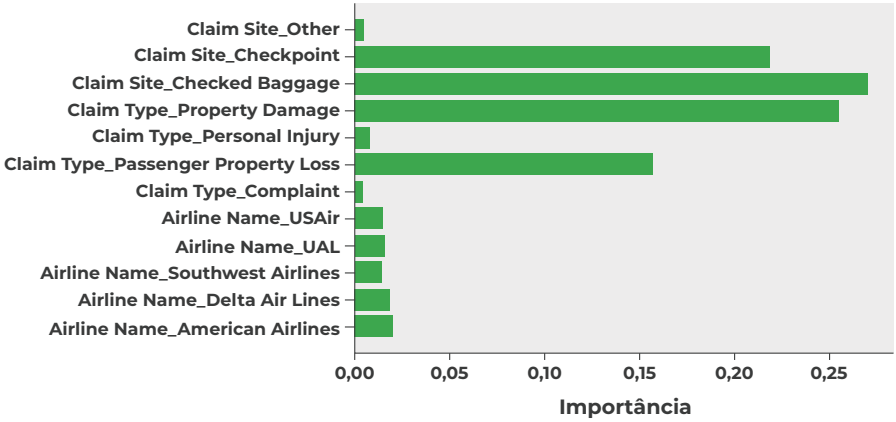

Figura 11.4: Importância das variáveis na Floresta Aleatória.

mais clara a correlação capturada pelos modelos (lembre-se de que correlação é diferente de causalidade e, portanto, devemos usar essa informação com responsabilidade). O gráfico chamado Shap Beeswarm tem como objetivo mostrar essa relação instância por instância do conjunto de dados, onde cada ponto no gráfico representa uma instância. As figuras 11.5, 11.6 e 11.7 ilustram a representação Shap Beeswarm para os modelos obtidos pela Regressão Logística, Árvore de Decisão e Floresta Aleatória, respectivamente.

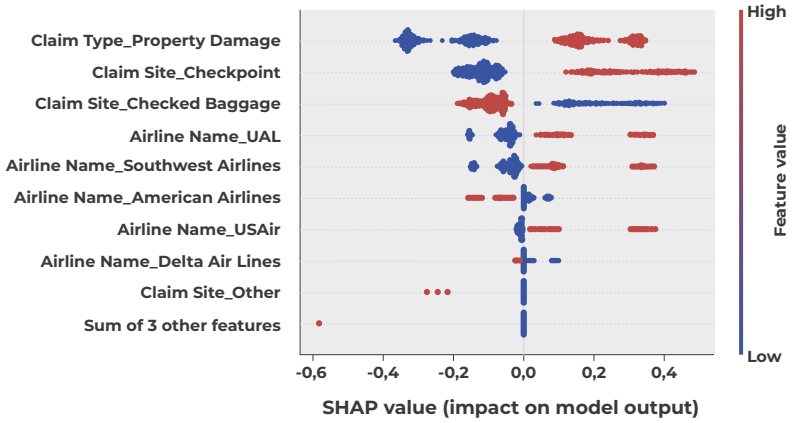

Figura 11.5: Shap Beeswarm na Regressão Logística.

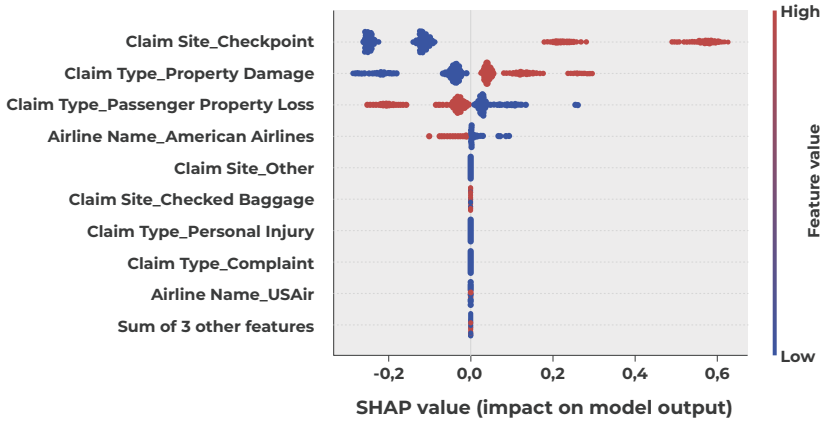

Figura 11.6: Shap Beeswarm na Árvore de Decisão.

Podemos notar, no modelo Regressão Linear, que o valor "checkpoint" tem de fato uma relevância para decidir se um cliente terá seu pedido aceito ou negado. Os valores "property damage" e "passenger property loss" também contribuem positivamente. Na Árvore de Decisão, as variáveis "checkpoint" e "property demage" são as que mais influenciam na decisão. Já na Floresta Aleatória, embora "property demage" e "checkpoint" tenham uma influência positiva, a variável "checked baggage" tem uma influência forte e negativa na decisão. Em resumo, o cliente que alegar que teve uma perda ou dano à sua

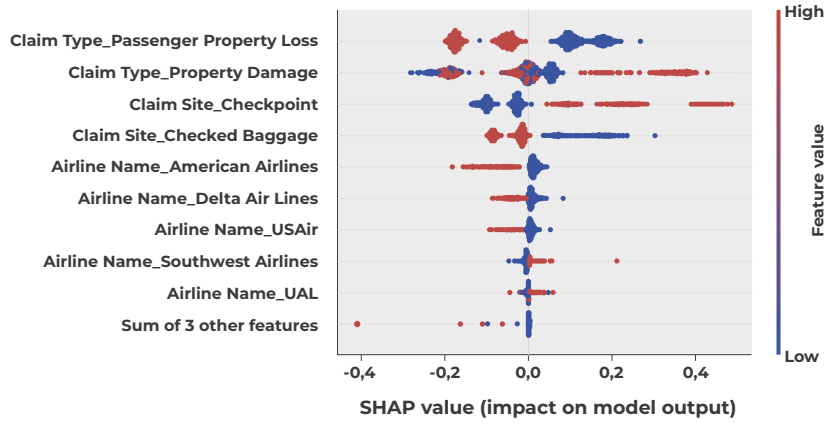

Figura 11.7: Shap Beeswarm na Floresta Aleatória.

propriedade, e que esse fato aconteceu no ponto de verificação, existe uma grande chance de sua reclamação ser atendida.

11.4 Análise dos Valores Recebidos

Na seção anterior, realizamos a análise baseada na decisão de aceitar ou negar o pedido de consideração da reclamação feita pelo cliente. Após as análises, sabemos quais tipos de reclamações e locais são mais propensos a serem aceitos ou negados, mas não sabemos quanto o cliente pode receber caso seu pedido seja considerado. Esta seção analisará essa segunda questão. Para isso, removeremos as reclamações negadas, ou seja, aquelas cujos valores das decisões (representados pela variável "Disposition") sejam diferentes de 0. Além disso, com base na Tabela 11.1, os valores devidos pagos ("Approved in Settle") tem o terceiro quartil (Q3) igual a $300, ou seja, apenas 25% das reclamações tem valores acima de $300. Assim, consideraremos nesta análise apenas os valores abaixo de $300, com o objetivo de considerar apenas os valores mais prováveis e remover possíveis *outliers*. A Figura 11.8 mostra um boxplot para cada tipo de decisão, considerando apenas os valores menores que $300.

Também removeremos as variáveis "Claim Number", "Date Received", "Incident D", "Airport Code", "Disposition", "Item Category" e

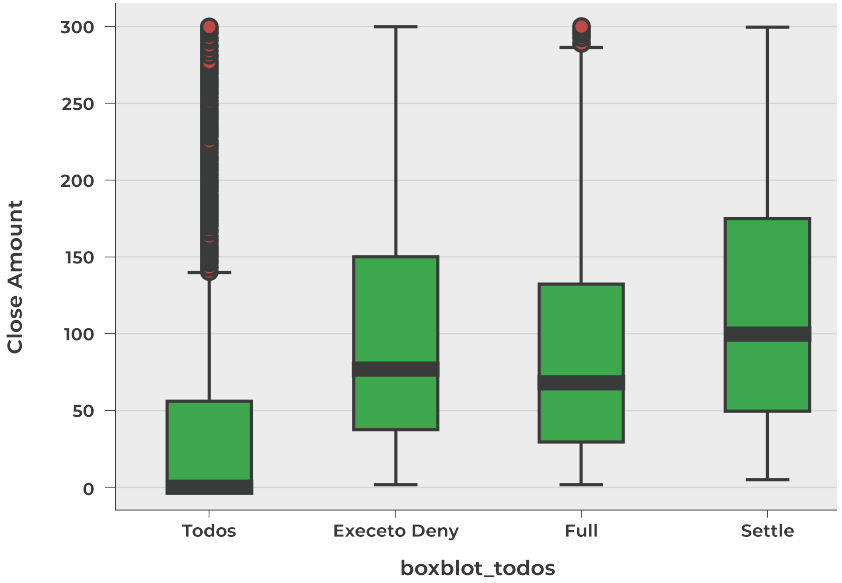

Figura 11.8: Boxplot para cada um dos possíveis tipos de reclamações considerando valores recebidos menores que $300.

"Airport Name", pois elas não contribuem para a extração da informação que queremos analisar. Assim, temos um problema de regressão, onde o objetivo é predizer o valor recebido (variável "Close Amount") com base nas variáveis independentes que caracterizam as reclamações. Consideraremos o algoritmo Floresta Aleatória para gerar um modelo de regressão para o nosso conjunto de dados, que será dividido em conjunto de treinamento e de teste com a proporção $70 - 30\%$.

A Figura 11.9 mostra a importância das variáveis para predizer o valor recebido.

Podemos perceber que as variáveis "Claim Type" e "Claim Site" (tipo e lugar registrados na reclamação) contribuem com um maior peso para decidir o valor final a ser recebido pelo cliente. Vamos analisar o Shap Beeswarm para a predição dos valores recebidos usando o algoritmo Floresta Aleatória, mostrado na Figura 11.10.

O Shap mostra que, apesar do tipo de reclamação ser uma variável importante, o valor recebido pode variar dependendo do valor dessa variável. Mais especificamente, quando o tipo da reclamação é "property damage", o valor recebido varia positivamente. Por outro lado,

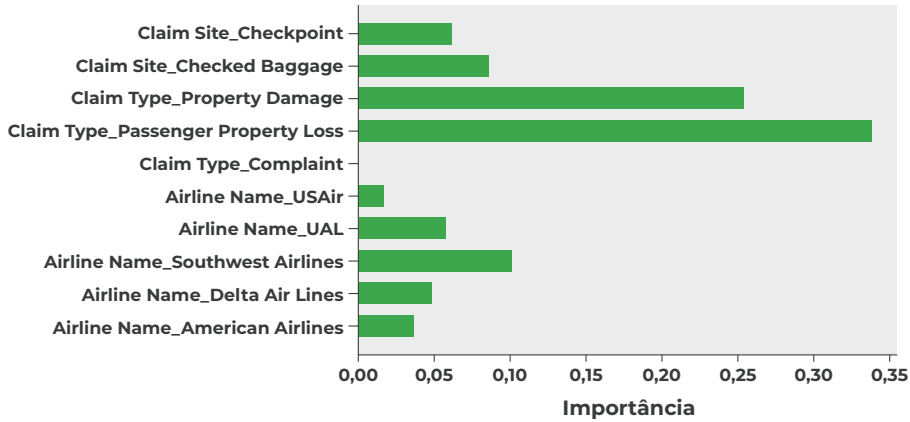

Figura 11.9: Importância das variáveis na Floresta Aleatória na predição do valor recebido.

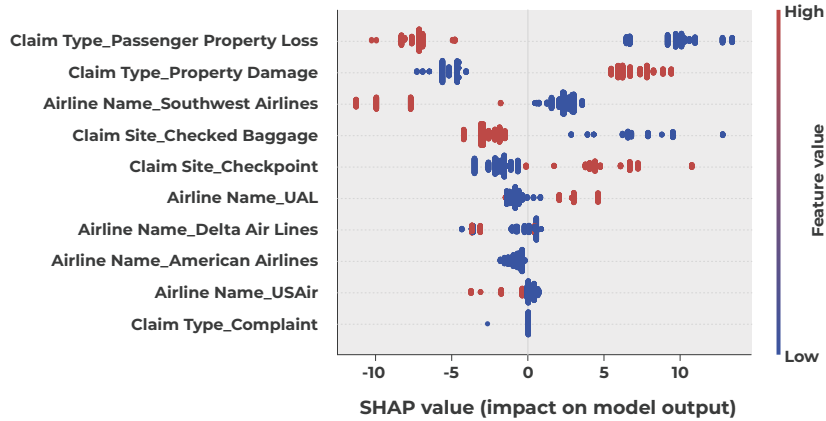

Figura 11.10: Shap Beeswarm para Floresta Aleatória na predição de valores recebidos.

quando o tipo da reclamação é "passenger property loss", o valor recebido varia negativamente. Isso quer dizer que um dano à propriedade pode resultar em o cliente receber um valor maior do que se a reclamação fosse sobre perda da propriedade.

De forma semelhante, pela figura da importância das variáveis obtida pela Floresta Aleatória, o local é uma característica da reclamação que influencia o valor recebido. Entretanto, com base no Shap, o valor recebido pode variar dependendo do local. Se o local for "checkpoint",