Seleção de cargas agropecuárias para inspeção

Projeto do curso Bootcamp ENAP em Machine Learning (SIGVIG Madeira Importação – Porto de Santos)

MINISTÉRIO DA AGRICULTURA, PECUÁRIA E ABASTECIMENTO

PÁTRIA AMADA BRASIL GOVERNO FEDERAL

Maurício Marinho – MAPA
Daniel Fugisawa - ABIN

Brasília, Dezembro de 2021

Problema/Hipótese

- No Porto de Santos, de 2015 a 2017, das interceptações totais de produtos importados que utilizam suporte ou embalagens de Madeira, **18% foram de pragas quarentenárias ausentes no Brasil** (pragas com potencial de dano econômico para a agricultura).

- Em 2021, até o mês de Outubro, chegaram em média 50 mil contêineres por mês para inspeção. Os fiscais selecionam **discricionariamente** quais cargas podem apresentar maior risco de pragas.

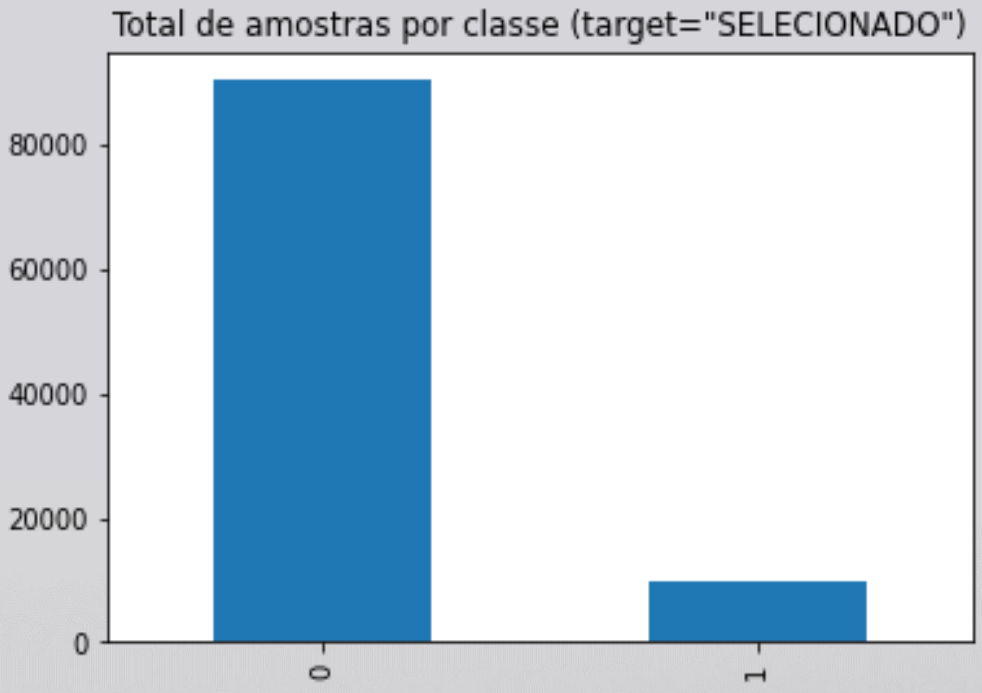
- Com a falta de padronização no gerenciamento de risco em várias unidades Vigiagro do país, há necessidade de critérios de seleção de cargas semelhantes aplicados em todos os portos.

- **Objetivo da área de negócio**: harmonizar e automatizar nacionalmente o gerenciamento de risco para a seleção de cargas a serem inspecionadas.

Abordagem para solução do problema

- **Sistema supervisionado preditivo baseado em Machine Learning para indicar quais cargas deverão ser inspecionadas pelos fiscais.**

Target sem TOM (Termo de ocorrência de Madeira):
(**0** – Liberar Carga , **1** – Selecionar Carga)



0 – 90.350 registros
1 – 9.650 registros

Target com TOM:
(**2** – Alto risco, **1** -Selecionar Carga, **0** – Liberar Carga)

0 – 90.350 registros , **1** – 9.523 registros
2 – 127 registros

Dados (pré-processamento)

- Quantidade separada para treino e teste: últimos **100 mil registros e no final o dataset completo com 583 mil registros**.

- Features/Colunas/Variáveis independentes:

1. País de origem da carga.
2. NCM (Nomenclatura Comum do Mercosul).
   - Reduzida para dois dígitos (atualmente usa 4 dígitos)
3. GR (Gerenciamento de risco) - Classificações baseadas em percentual de temos de ocorrência de madeira (TOM) em período anual.
4. Quantidade de cargas selecionadas (anual).
5. Quantidade de cargas recebidas (anual).
6. Indicação de que o CNPJ não foi inspecionado (anual).
7. IN_49_2009 (situação de indicação para inspeção).
8. Percentual não selecionado do CNPJ (anual).
9. Diferença entre total recebido e total inspecionados do CNPJ (anual)

Dados (pré-processamento)

- **Codificação de variáveis categóricas (GR, NCM e País)**:
  - **CatBoostEncoder**
    - Não cria novas colunas.
    - Proporcionou menor tempo de execução do treinamento do modelo.
  - **OnHotEncoder**
    - Aumentou substancialmente o tempo de execução dos modelos.
    - Apresentou melhor recall que o CatBoost porém piorou muito no balanced acuracy (ou seja, muita ocorrência de falso positivo).

- **Desbalanceamento**: O Undersampling da classe majoritária foi feito mas não gerou bons resultados. A estratégia utilizada foi alterar os parâmetros do modelo para indicar esta situação.

- **Normalização**: Não ocasionou melhora substancial no modelo de Regressão Logística.

- Os modelos foram treinados e testados após busca de hiperparâmetros com **GridSearchCV** utilizando **validação cruzada 5 folds** com proporção**: 80% treino e 20% teste, com estratificação.**

Treinamento e avaliação do modelo

Target sem TOM (Termo de ocorrência de Madeira**):** (**0** – Liberar Carga , **1** – Selecionar Carga)

Métrica principal: **Recall** (VP / VP + FN) onde VP = número de verdadeiros positivos e FN = falsos negativos

Modelo	Melhores hiperparâmetros	Recall/balanced acuracy score (100K)	Recall (583K)	F1 (583K)
Regressão logística	C: 10 class_weight: balanced max_iter: 1000 penalty: l2	0.76	0.60	0.84
KNN	n_neighbors: 2 p: 2	0.48	0.49	0.93
Random Forest	n_estimators: 500, max_depth: None, min_samples_split: 100, min_samples_leaf: 10, class_weight: **balanced**	**0.79**	**0.71**	0.88
XGBoost	eta: 0.01, gamma: 0.5, max_depth: 25, min_child_weight: 3, reg_lambda: 1, subsample: 1	0.45	----	----
Ensemble	Regressão logística, Random Forest e KNN	0.66	----	----

Treinamento e avaliação do modelo

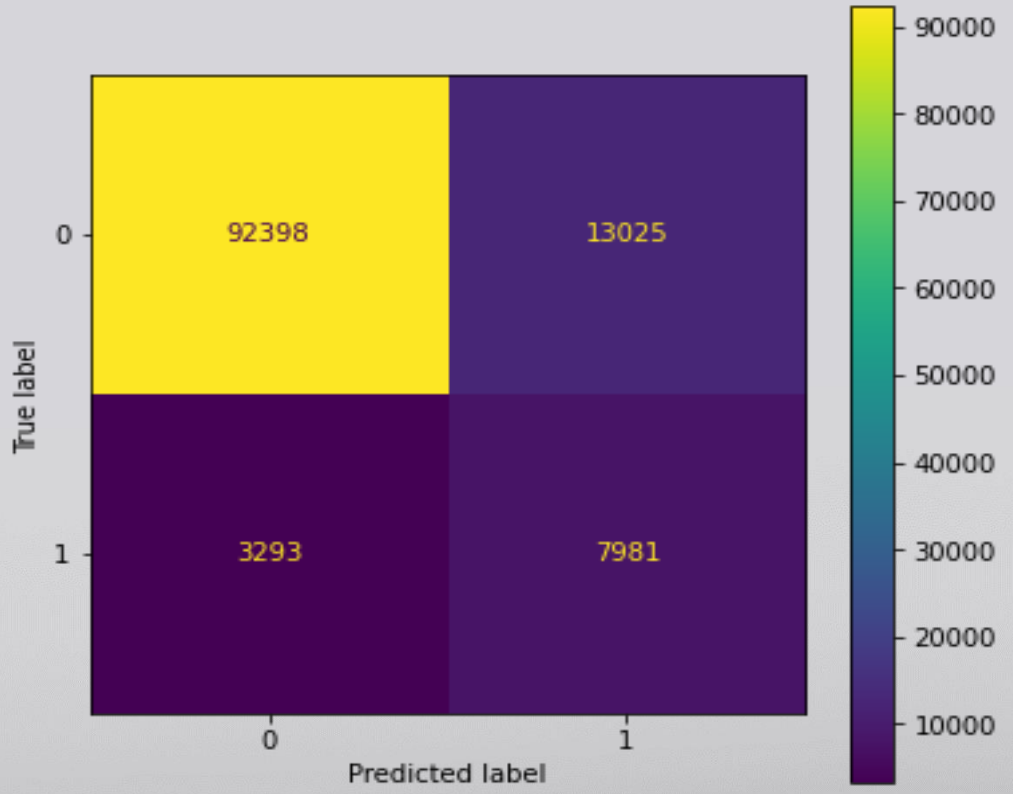
Target com TOM: (**2** – Alto risco, **1** -Selecionar Carga, **0** – Liberar Carga)

Modelo	Melhores hiperparâmetros	Recall (100K)	Recall (583K)	F1 score (583K)
Random Forest	n_estimators: 500, max_depth: None, min_samples_split: 100, min_samples_leaf: 10, class_weight: balanced	1 – 0.76 2 – 1.00	1 – 0,70 **2 – 0.88**	0.88

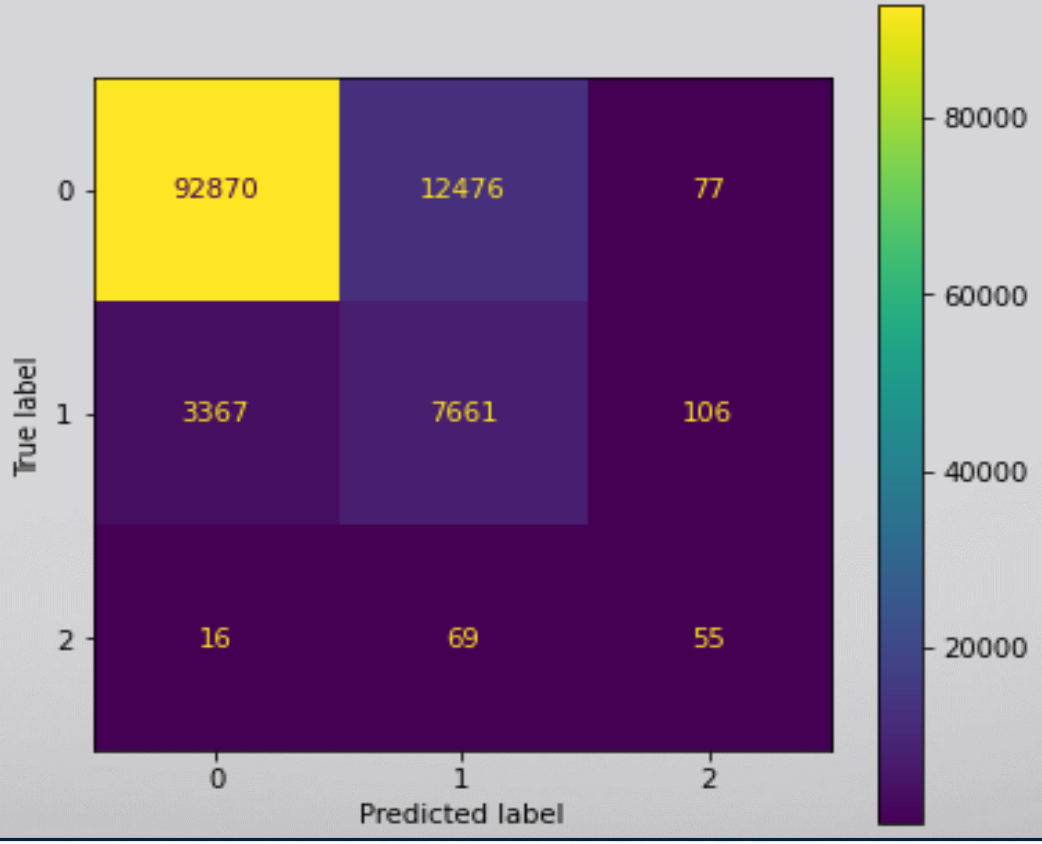
Interpretação do modelo

- **Melhor modelo testado:** Random Forest Classifier
- Abaixo as matrizes de confusão do teste na amostra total de 2021 até Outubro (20% de 583K = 117K)

Target sem TOM (Termo de ocorrência de Madeira):
(0 – Liberar Carga , **1** – Selecionar Carga)



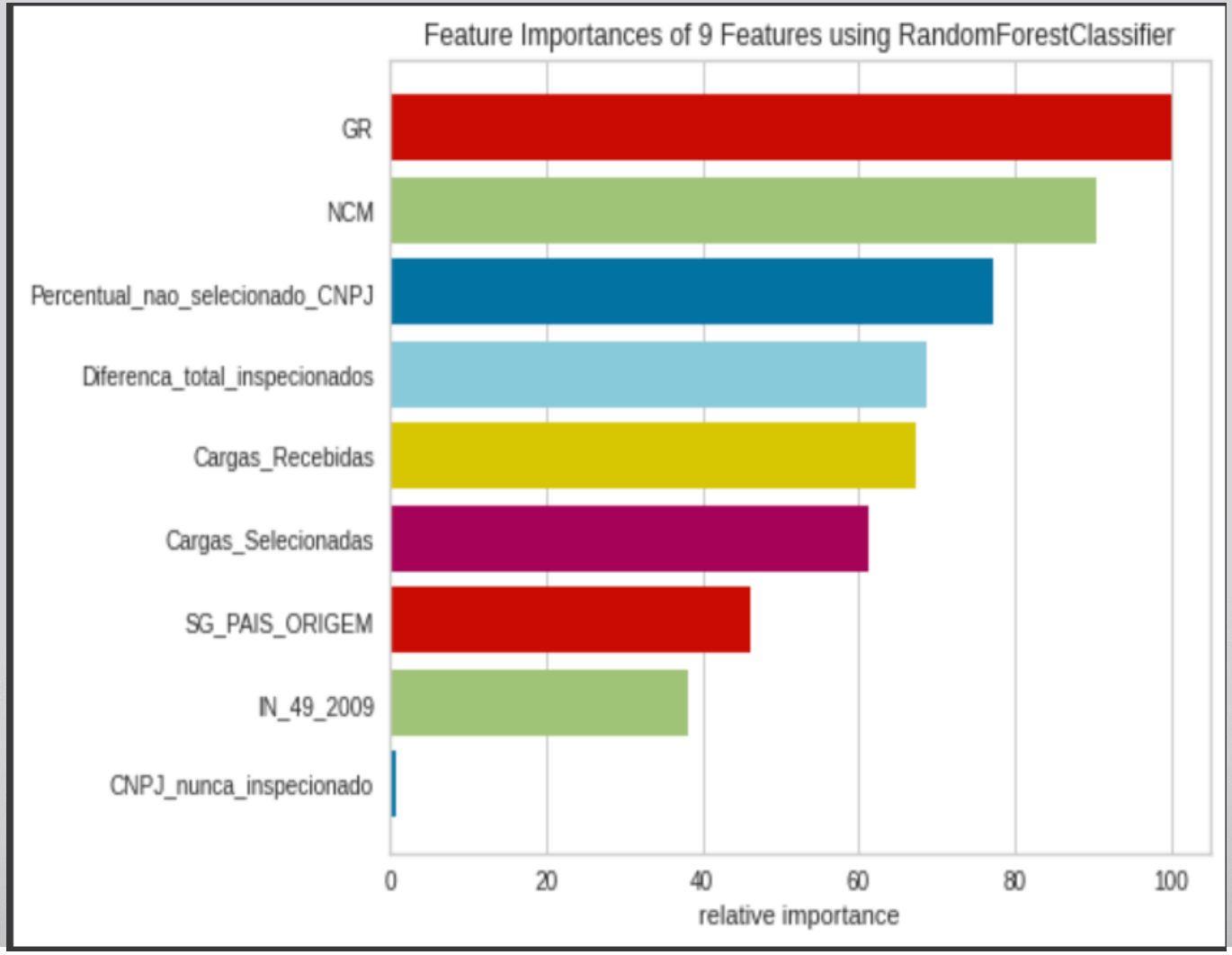
Target com TOM:
(**2** – Alto risco, **1** -Selecionar Carga, **0** – Liberar Carga)



Interpretação do modelo

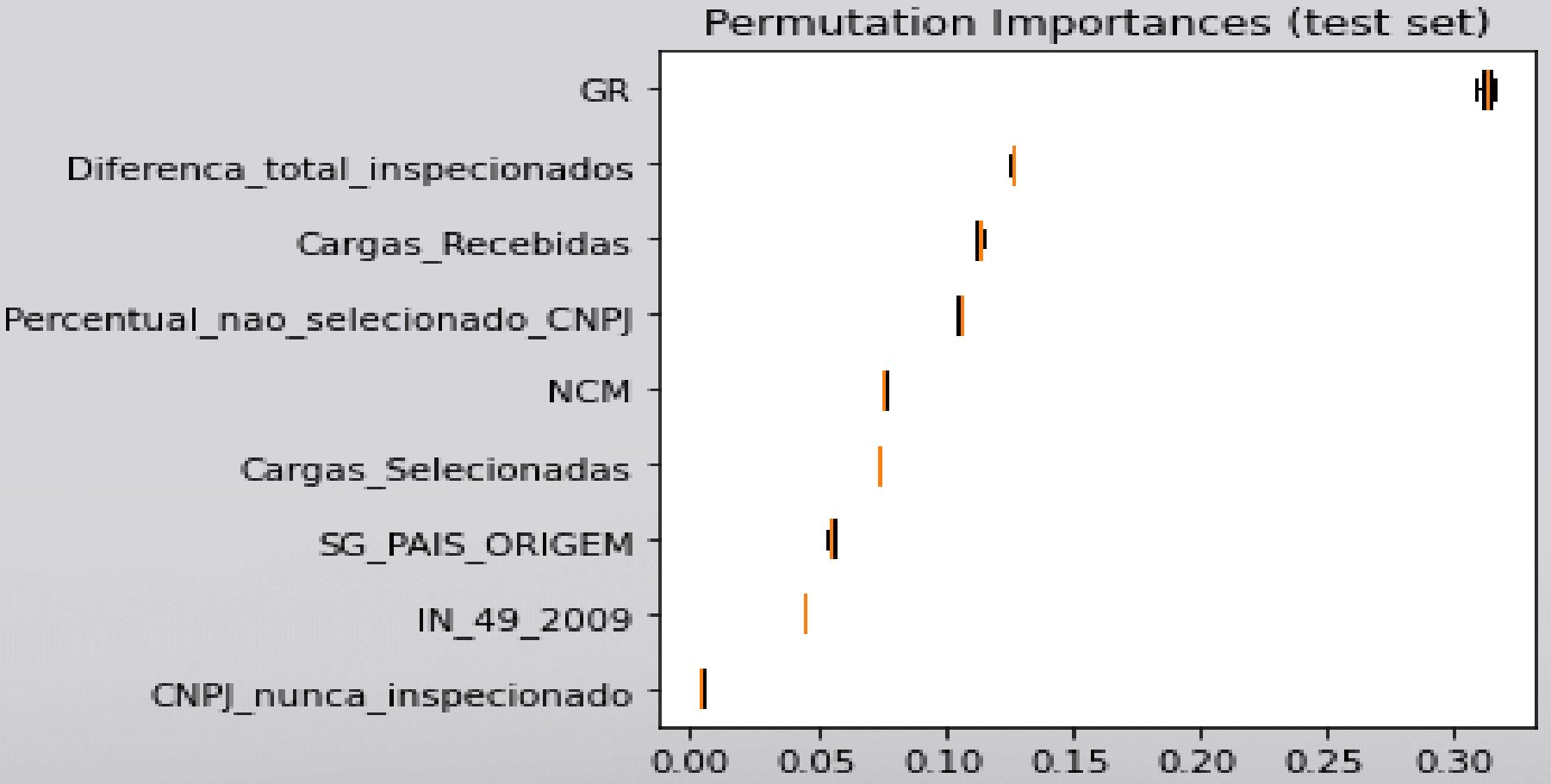
- **Abaixo são apresentadas a importância percentual e a importância relativa das features para o melhor modelo (RandomForest):**

	importance
GR	0.181849
NCM	0.164168
Percentual_nao_selecionado_CNPJ	0.140169
Diferenca_total_inspecionados	0.124847
Cargas_Recebidas	0.122458
Cargas_Selecionadas	0.111396
SG_PAIS_ORIGEM	0.083713
IN_49_2009	0.069664
CNPJ_nunca_inspecionado	0.001737



Interpretação do modelo

- **Resultados (permutation_importance) da queda no desempenho do modelo ao embaralhar cada uma das variáveis.**



Insigths e próximos passos

- **Analisar falso positivo** pois o modelo pode estar indicando situações com risco e que não estão sendo consideradas pelos fiscais.
- **Analisar falso negativo** para verificar se o modelo deixou de indicar casos importantes ou se eram importações aparentemente tranquilas e que foram escolhidas pelos fiscais por algum critério mais subjetivo não aprendido pelo modelo.
- **Testar outros modelos** não usados neste projeto.
- Incluir **novas features** a exemplo da taxa de TOM (com eventual categorização) das empresas exportadoras.
- Sugerir à área negocial que o modelo atual treinado com os devidos ajustes poderia ser utilizado como um **apoio e reforço no gerenciamento de riscos em Santos**. Para isso, seria criada uma coluna no formulário com a sugestão do algoritmo de ML para cada carga com indicação de probabilidade da categoria sugerida pelo modelo.
- Com os devidos ajustes, o modelo pode ser **utilizado em outros portos** que não possuem gerenciamento de risco tão organizado quanto o porto de Santos.