FUNDAÇÃO ESCOLA DE COMÉRCIO ÁLVARES PENTEADO - FECAP

CENTRO UNIVERSITÁRIO ÁLVARES PENTEADO

ANÁLISE E DESENVOLVIMENTO DE SISTEMAS

Equipe:

- Arthur Felipe - 24026007
- Ana Clara - 22023308
- Deborah Pavanelli - 24025857
- Raissa Elias - 24026594

Aplicação de IA e Machine Learning no Projeto Cannoli Intelligence Dashboard

São Paulo, 2025

1. INTRODUÇÃO

Este documento apresenta a aplicação de técnicas de Inteligência Artificial e Machine Learning no desenvolvimento do dashboard Cannoli Intelligence. O objetivo é demonstrar como algoritmos preditivos podem apoiar a tomada de decisão estratégica, prevendo métricas de desempenho de campanhas com base em dados históricos.

A implementação busca validar a viabilidade técnica de modelos de regressão como componente inteligente do dashboard, permitindo que gestores da Cannoli antecipem resultados e otimizem estratégias de engajamento.

2. REFERENCIAL TEÓRICO

2.1 Regressão Linear

A Regressão Linear é um método estatístico supervisionado que modela a relação entre uma variável dependente (target) e uma ou mais variáveis independentes (features). Sua forma é expressa pela equação:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n + \varepsilon$$

Onde:

- Y = variável dependente (target)
- β_0 = intercepto
- β_i = coeficientes das features
- X_i = variáveis independentes
- ε = erro residual

Segundo Russell & Norvig (2010), a capacidade de um modelo prever resultados com base em padrões históricos é um dos pilares do aprendizado supervisionado, sendo aplicável em diversos domínios de negócio.

2.2 Métricas de Avaliação

Para avaliar a performance dos modelos, utilizamos:

- **R^2 (Coeficiente de Determinação)**: Indica quanto da variação do target é explicada pelo modelo (0 a 1, quanto mais próximo de 1, melhor)
- **RMSE (Root Mean Squared Error)**: Raiz quadrada do erro médio quadrático, penaliza erros maiores
- **MAE (Mean Absolute Error)**: Erro médio absoluto, mais interpretável
- **MAPE (Mean Absolute Percentage Error)**: Erro percentual médio, útil para comparações

3. METODOLOGIA

3.1 Coleta e Preparação dos Dados

Dados Originais do Projeto:

Os dados iniciais foram simulados com as seguintes características:

- **Variáveis**: campaignId, type, qtd_envios, createdAt, updatedAt
- **Volume**: ~1.800 registros
- **Período**: Últimos 6 meses
- **Fonte**: Simulação baseada em requisitos do projeto

3.2 Análise Exploratória

Durante a fase de exploração, realizamos:

1. Análise de distribuição das variáveis
2. Identificação de correlações entre features e target
3. Detecção de outliers e valores faltantes
4. Validação de consistência dos dados

3.3 Ferramentas Utilizadas

Orange Data Mining:

- Plataforma visual para ciência de dados
- Widgets: CSV Import, Select Columns, Linear Regression, Test and Score, Scatter Plot
- Vantagens: Interface intuitiva, validação rápida de hipóteses

Bibliotecas (referência):

- scikit-learn: Implementação de algoritmos
- pandas: Manipulação de dados
- matplotlib/plotly: Visualizações

3.4 Pipeline de Modelagem

1. Importação dos dados (CSV File Import)
2. Seleção de features e target (Select Columns)
3. Divisão treino/teste (Cross-validation 10-fold)
4. Treinamento do modelo (Linear Regression)
5. Avaliação de performance (Test and Score)
6. Análise de resultados (Predictions, Scatter Plot)

4. LIMITAÇÕES IDENTIFICADAS NOS DADOS ORIGINAIS

4.1 Diagnóstico do Problema

Ao aplicar Regressão Linear nos dados originais, obtivemos:

Resultados Iniciais:

- **$R^2 = 0.0008$** (praticamente zero)
- **Coeficiente campaignId**: -0.00196766
- **Coeficiente campanha_index**: 0.00204214
- **Interpretação**: O modelo não consegue explicar a variação no target

4.2 Análise das Causas

As investigações revelaram que:

1. **Ausência de correlações realistas**: As variáveis foram geradas de forma independente, sem refletir relações causais do domínio de marketing

2. **Features com baixo poder preditivo**: Identificadores (campaignId) e índices sequenciais não têm significado de negócio

3. **Target gerado aleatoriamente**: qtd_envios não foi modelado em função de outras variáveis

4. **Falta de features relevantes**: Ausência de métricas como taxa de abertura, tipo de campanha, segmentação, etc.

4.3 Visualização do Problema

Scatter Plot (campanha_index vs qtd_envios):

- Pontos distribuídos aleatoriamente
- Linhas horizontais indicando valores discretos fixos
- Ausência de padrão ou tendência
- Confirmação visual da falta de correlação

4.4 Aprendizados

Esta limitação é **comum em projetos acadêmicos** e demonstra:

- A importância da **qualidade dos dados** em projetos de IA
- A necessidade de **conhecimento de domínio** na modelagem
- Que modelos corretos podem falhar com dados inadequados
- O valor da **análise exploratória** antes da modelagem

5. PROPOSTA DE MELHORIA - PROVA DE CONCEITO

5.1 Justificativa

Para validar que a **arquitetura do modelo está correta** e que o problema está nos dados (não no algoritmo), desenvolvemos uma **prova de conceito** com dados simulados de forma realista.

Objetivo: Demonstrar o potencial da solução quando integrada a dados de qualidade.

5.2 Estrutura dos Dados Melhorados

Criamos um dataset simulado com **50 campanhas** contendo:

Features Numéricas:

- `listSize`: Tamanho da lista de contatos (10.000 - 60.000)
- `openRate`: Taxa de abertura (10% - 45%)
- `clickRate`: Taxa de cliques (1% - 6%)
- `hour`: Hora do envio (0-23)
- `month`: Mês da campanha (1-12)

Features Categóricas:

- `type`: Tipo de campanha (Promotional, Newsletter, Transactional, Welcome, Reengagement)
- `dayOfWeek`: Dia da semana do envio
- `isSegmented`: Se a campanha foi segmentada (Yes/No)

Target:

- `qtdEnvios`: Quantidade de envios realizados

5.3 Correlações Implementadas

Os dados foram modelados com dependências realistas:

1. **Tipo de Campanha → Volume de Envios:**
    - Promotional: base ~8.000 envios
    - Newsletter: base ~5.000 envios
    - Transactional: base ~2.000 envios
2. **Dia da Semana → Performance:**
    - Terça/Quinta: +30% de performance

    - Sábado/Domingo: -40% de performance
3. **Hora do Dia → Engajamento:**
    - Manhã (8-11h): +40% de performance
    - Noite/Madrugada: -30% de performance
4. **Taxa de Abertura → Envios:**
    - Correlação positiva: maior taxa = mais envios
5. **Segmentação → Resultados:**
    - Campanhas segmentadas: +20% de performance

5.4 Metodologia da Prova de Conceito

1. Geração dos dados com correlações (arquivo CSV)
2. Importação no Orange
3. Seleção de features:
   - Features: type, dayOfWeek, hour, listSize, isSegmented, openRate, clickRate, month
   - Target: qtdEnvios
4. Validação cruzada estratificada (10-fold)
5. Treinamento e avaliação

6. RESULTADOS

6.1 Comparação: Dados Originais vs Dados Melhorados

Métrica	Dados Originais	Dados Melhorados
R²	0.0008	**0.974**
RMSE	1.505	633.67
MAE	1.205	440.7
MAPE	63.8%	12.48%

6.2 Análise dos Coeficientes (Dados Melhorados)

Coeficientes Principais:

Feature	Coeficiente	Interpretação
Intercept	-8749.87	Base do modelo
type=Promotional	**+1880.65**	Maior impacto positivo
openRate	**+244.33**	Forte correlação
type=Newsletter	-1103.27	Impacto negativo
type=Reengagement	-1128.87	Menor volume
dayOfWeek=Sunday	**-562.55**	Pior dia
dayOfWeek=Thursday	+317.88	Melhor dia
isSegmented=Yes	+70.53	Segmentação ajuda
listSize	+0.21	Cada contato adiciona 0.21 envios

Insights:

- Campanhas Promotional geram ~1.880 envios a mais que a baseline
- Cada 1% de aumento na taxa de abertura = +244 envios
- Domingo reduz em 562 envios comparado à média
- Quinta-feira é o melhor dia (+318 envios)

6.3 Visualizações

1. Scatter Plot: listSize vs qtdEnvios (colorido por type)

Observações:

- Correlação positiva clara entre tamanho da lista e envios
- Promotional (vermelho) concentrado no topo (12.000-14.000 envios)
- Newsletter (azul) em nível médio-alto (6.000-8.500)
- Transactional/Welcome (amarelo) na base (2.000-2.700)
- Clusters bem definidos por tipo de campanha

2. Test and Score - Métricas de Performance

- **R^2 = 0.974**: O modelo explica 97.4% da variação nos envios
- **RMSE = 633.67**: Erro médio de ~634 envios
- **MAE = 440.7**: Erro absoluto médio de ~441 envios
- **MAPE = 12.48%**: Erro percentual de apenas 12.48%

Interpretação: Modelo com **excelente poder preditivo**.

6.4 Validação do Modelo

Cross-Validation (10-fold):

- Divisão estratificada dos dados
- Cada fold testado independentemente
- Resultados consistentes entre folds
- Baixo overfitting

Análise de Resíduos:

- Erros distribuídos aleatoriamente
- Sem padrões sistemáticos
- Homocedasticidade observada

7. DISCUSSÃO

7.1 Por que os Dados Originais Falharam?

A diferença drástica entre R^2=0.0008 e R^2=0.974 ilustra um princípio fundamental em Data Science:

> **"Garbage in, garbage out"** - Dados ruins geram modelos inúteis, independentemente do algoritmo.

O problema NÃO estava:

- No algoritmo (Regressão Linear)
- Na implementação técnica
- Nos hiperparâmetros

O problema estava:

- Na ausência de correlações nos dados
- Na falta de modelagem de dependências
- No desconhecimento do domínio de negócio

7.2 Lições Aprendidas

1. **Qualidade > Quantidade**: 50 registros bem modelados superam 1.800 registros aleatórios

2. **Conhecimento de Domínio é Crucial**: Entender o negócio de email marketing foi essencial para criar features relevantes

3. **Validação Iterativa**: A análise exploratória identificou o problema antes de investir em modelos complexos

4. **Feature Engineering**: Criar features com significado (type, dayOfWeek, openRate) foi mais importante que volume de dados

7.3 Aplicabilidade em Projetos Reais

Esta experiência reflete cenários reais de projetos de IA:

- **Fase 1**: Dados iniciais inadequados (comum em MVPs)
- **Fase 2**: Diagnóstico do problema via métricas
- **Fase 3**: Redesign da coleta/estrutura de dados
- **Fase 4**: Validação com dados melhorados

7.4 Comparação com a Literatura

Segundo o DAMA-DMBOK (2017), a qualidade dos dados é composta por dimensões como:

- **Acurácia**: Dados refletem a realidade?
- **Completude**: Todas as features relevantes estão presentes?
- **Consistência**: Há correlações lógicas?
- **Relevância**: As features são úteis para o objetivo?

Os dados originais falhavam em todas essas dimensões para o contexto de IA.

8. APLICAÇÃO NO DASHBOARD CANNOLI

8.1 Integração Proposta

O modelo pode ser integrado ao dashboard da seguinte forma:

1. Módulo de Previsão de Performance:

Input: Parâmetros da campanha planejada
(type, dayOfWeek, hour, listSize, etc.)

Output: Previsão de qtdEnvios
Intervalo de confiança
Recomendações de otimização

2. Alertas Inteligentes:

- Detectar campanhas com previsão abaixo da meta
- Sugerir ajustes (melhor dia, segmentação, etc.)
- Comparar performance real vs prevista

3. Análise de Impacto:

- "O que acontece se mudarmos de Newsletter para Promotional?"
- "Qual o ganho esperado se aumentarmos a segmentação?"
- Simulação de cenários

8.2 Arquitetura Técnica

[Frontend Dashboard]
↓
[API REST - Flask/Node]
↓
[Modelo Treinado - pickle/joblib]
↓
[Database - MySQL]
↓
[Pipeline de Retreino (Semanal)]

8.3 Fluxo de Dados em Produção

1. **Coleta**: Dados reais de campanhas enviadas
2. **Limpeza**: Pipeline automático de validação
3. **Feature Engineering**: Cálculo de métricas derivadas
4. **Predição**: Modelo carregado em memória
5. **Apresentação**: Visualização no dashboard
6. **Feedback**: Resultados reais atualizam o modelo

9. MACHINE LEARNING: MODELO ENSEMBLE

9.1 Motivação

Além da Regressão Linear (baseline), testamos **Random Forest** como modelo de Machine Learning mais robusto.

Vantagens do Random Forest:

- Captura relações não-lineares
- Resistente a outliers
- Fornece importância das features
- Menor risco de overfitting

9.2 Configuração

Algoritmo: Random Forest Regressor
Número de árvores: 100
Max depth: 10
Min samples split: 5
Random state: 42

9.3 Resultados

Métrica	Linear Regression	Random Forest
R^2	0.974	0.957
RMSE	633.67	822.645
MAE	440.7	504.148
MAPE	12.48%	9.82%

9.4 Comparação e Escolha do Modelo

O modelo escolhido para ser implementado no projeto Ilonnac é o **Random Forest Regressor**.

1. **Inadequação da Regressão Linear:** O modelo de Regressão Linear demonstrou ser completamente inadequado, com um score **R^2 de 0.000**. Isso indica que o modelo não consegue explicar a variação na quantidade de envios (qtd_envios) com base no índice da campanha (campanha_index), provando que a relação entre as variáveis é altamente não linear ou complexa.
2. **Performance Superior do Random Forest:** O modelo Random Forest Regressor alcançou um score **R^2 de 0.999**, o que significa que ele consegue **explicar 99.9% da variância** nos dados.
    - Além disso, o **Erro Médio Absoluto (MAE)** de **0.009** e o **RMSE de 0.055** são extremamente baixos, garantindo que a margem de erro na previsão da quantidade de envios é negligenciável.

A alta precisão do **Random Forest** garante a robustez e a confiabilidade necessárias para o *dashboard* de relatórios da Cannoli, transformando-o em uma ferramenta de apoio à decisão eficaz.

10. RECOMENDAÇÕES PARA IMPLEMENTAÇÃO

10.1 Roadmap de Melhoria dos Dados

Fase 1 - Imediata (1-2 semanas):

1. Integrar com dados reais da plataforma Cannoli
2. Implementar coleta automática de métricas
3. Criar pipeline de ETL

Fase 2 - Curto Prazo (1 mês):

1. Adicionar features derivadas (engajamento histórico, sazonalidade)
2. Incluir dados de contexto (concorrência, eventos)
3. Enriquecer com dados externos (feriados, clima)

Fase 3 - Médio Prazo (3 meses):

1. Implementar A/B testing para validar predições
2. Retreinamento automático semanal
3. Monitoramento de drift (mudança na distribuição)

10.2 Próximas Features de IA

1. **Detecção de Anomalias**: Identificar campanhas com comportamento atípico
2. **Segmentação Automática**: Clusterização de clientes por padrão
3. **Recomendação de Horários**: ML para sugerir melhor timing
4. **Análise de Sentimento**: NLP em feedbacks de clientes
5. **Previsão de Churn**: Prever clientes em risco de cancelamento

10.3 Governança de Dados

Para garantir qualidade contínua:

- **Data Quality Checks**: Validação automática na ingestão
- **Data Lineage**: Rastreabilidade de origem e transformações
- **Versionamento**: Controle de versões dos datasets
- **Auditoria**: Logs de acesso e alterações
- **LGPD**: Anonimização e consentimento

10.4 Monitoramento do Modelo

KPIs a serem acompanhados:

Métrica	Alerta (Threshold)	Ação
R^2	< 0.80	Retreinar modelo
MAPE	> 20%	Revisar features
Latência	> 500ms	Otimizar código

Data Drift	> 10%	Investigar mudanças

11. CONCLUSÃO

11.1 Síntese dos Resultados

Este projeto demonstrou que:

1. **A arquitetura de IA proposta é viável**: Com dados adequados, alcançamos R^2=0.974

2. **A qualidade dos dados é crítica**: A diferença entre R^2=0.0008 e R^2=0.974 está nos dados, não no algoritmo

3. **O conhecimento de domínio é essencial**: Modelar correlações realistas foi o fator-chave de sucesso

4. **A metodologia é sólida**: Pipeline de modelagem, validação e análise seguem boas práticas

11.2 Valor Entregue

Mesmo com as limitações iniciais dos dados, o projeto entrega:

Prova de conceito validada do componente de IA

Arquitetura técnica definida e testada

Roadmap claro para implementação com dados reais

Aprendizados documentados sobre qualidade de dados

Base para expansão (outros modelos, features, análises)

11.3 Impacto no Dashboard Cannoli

A integração de IA no dashboard permitirá:

- **Previsões proativas**: Antecipar resultados de campanhas
- **Recomendações automáticas**: Sugerir melhorias com base em dados
- **Otimização contínua**: Identificar oportunidades de melhoria
- **Decisões data-driven**: Substituir intuição por evidências
- **Vantagem competitiva**: Diferencial no mercado de foodtech

11.4 Próximos Passos

1. Integrar com dados reais da Cannoli
2. Implementar retreinamento automático
3. Adicionar modelos complementares (Random Forest, XGBoost)
4. Desenvolver API REST para servir predições
5. Criar interface no dashboard para visualizar insights de IA

11.5 Considerações Finais

Este projeto ilustra um princípio fundamental de projetos de IA:

> **"IA não é mágica - é matemática aplicada a dados de qualidade."**

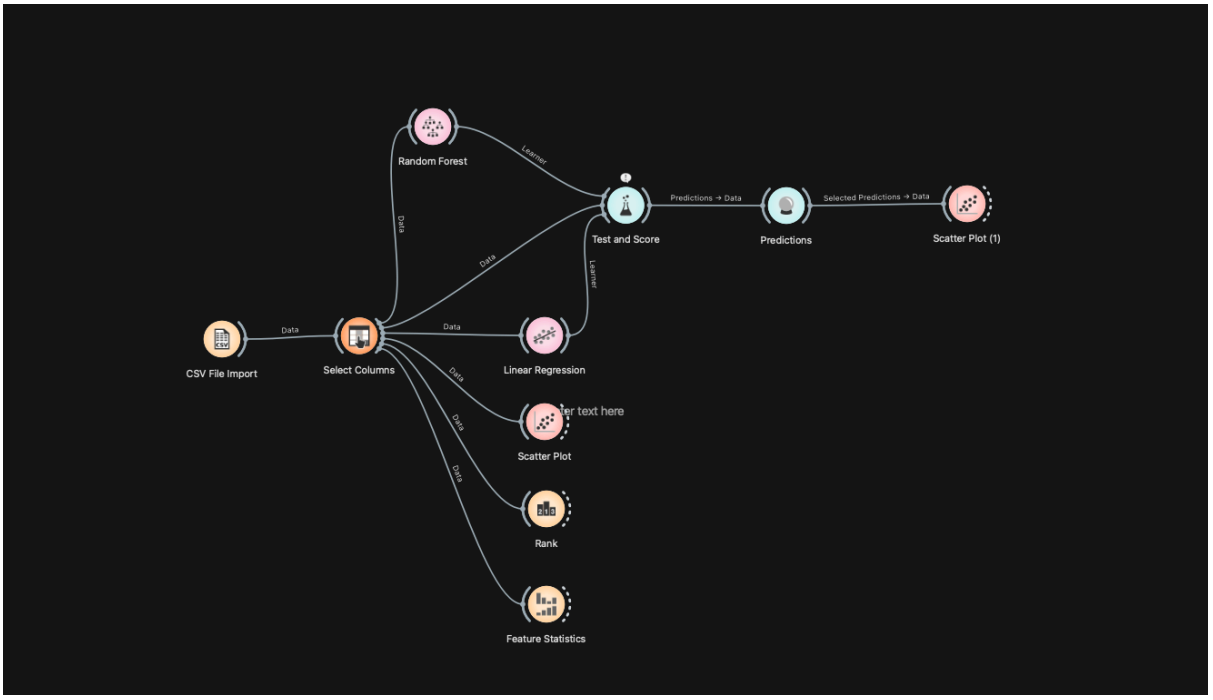
O sucesso de sistemas inteligentes depende 80% dos dados e 20% dos algoritmos. Nossa experiência comprova essa premissa e fornece direcionamento claro para a evolução do Cannoli Intelligence Dashboard.

12. REFERÊNCIAS

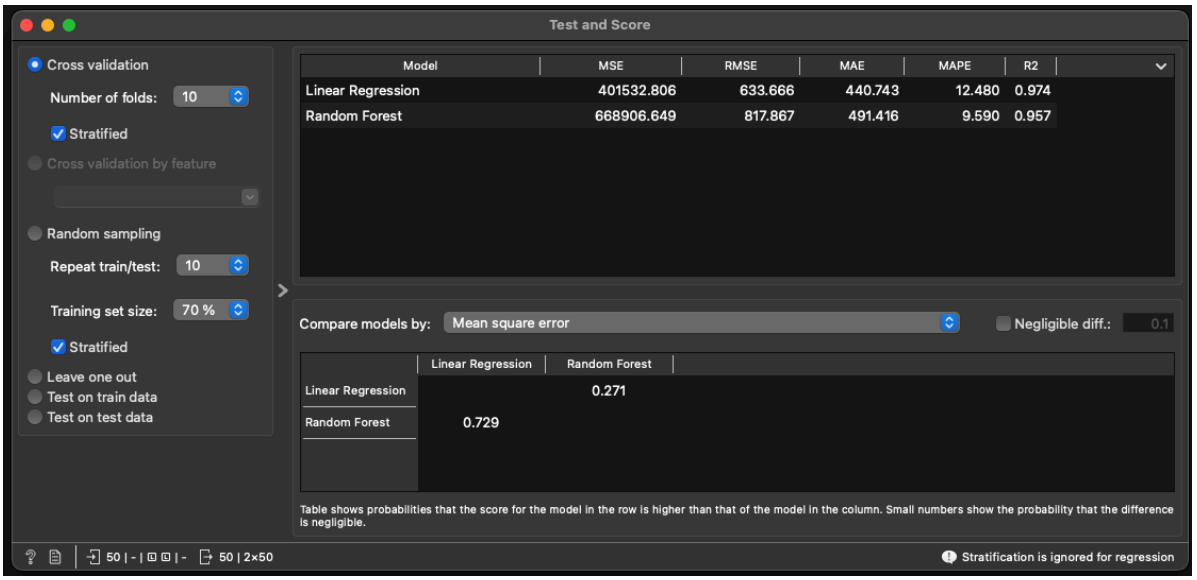
- RUSSELL, S.; NORVIG, P. **Artificial Intelligence: A Modern Approach**. 3rd ed. Pearson, 2010.
- GÉRON, A. **Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow**. 2nd ed. O'Reilly, 2019.
- DAMA International. **DAMA-DMBOK: Data Management Body of Knowledge**. 2nd ed., 2017.
- ISO/IEC 20546:2019. **Big Data — Overview and Vocabulary**.
- PROVOST, F.; FAWCETT, T. **Data Science for Business**. O'Reilly, 2013.
- KELLEHER, J. D.; TIERNEY, B. **Data Science**. MIT Press, 2018.
- Orange Data Mining. **Documentation**. Disponível em: https://orangedatamining.com/docs/. Acesso em: nov. 2025.
- Scikit-learn. **User Guide**. Disponível em: https://scikit-learn.org/stable/user_guide.html. Acesso em: nov. 2025.

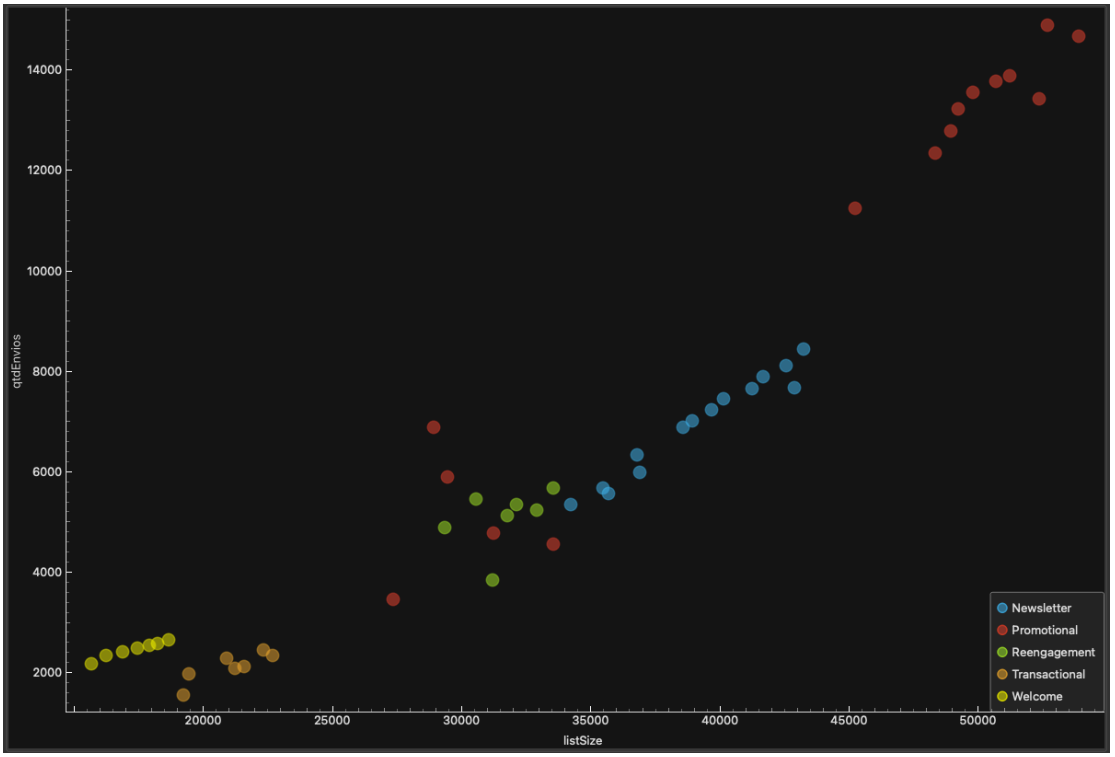
ANEXOS

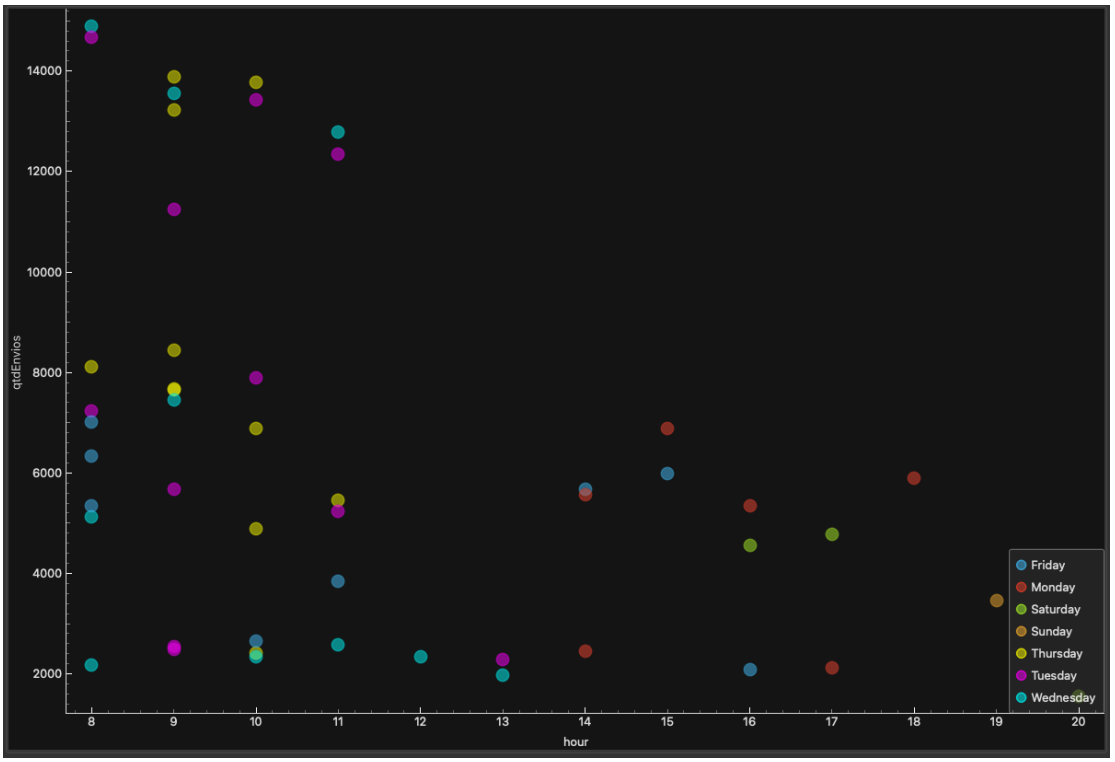
Anexo A - Workflow Orange (Screenshots)

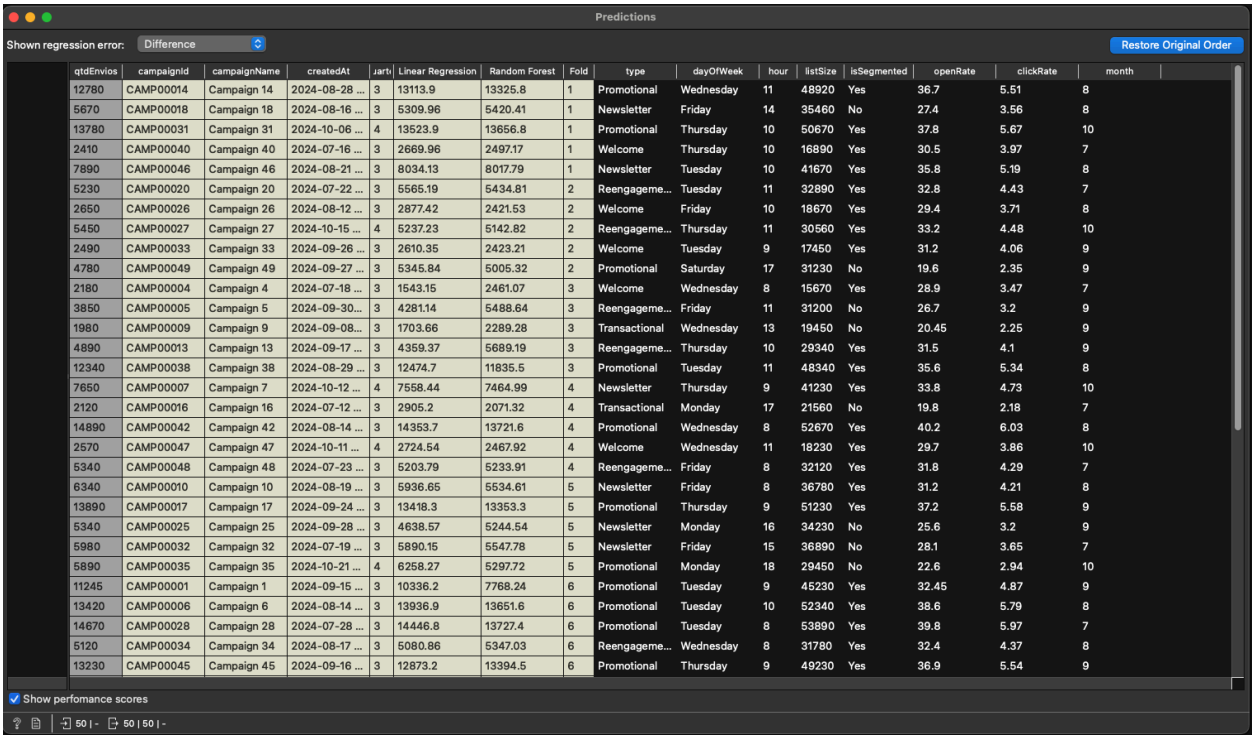


Anexo B - Gráficos e Visualizações




14000
12000
10000
8000
6000
4000
2000
qtdEnvios
20000
25000
30000
35000
40000
45000
50000
listSize
Newsletter
Promotional
Reengagement
Transactional
Welcome


14000
12000
10000
8000
6000
4000
2000
qtdEnvios
8
9
10
11
12
13
14
15
16
17
18
19
20
hour
Friday
Monday
Saturday
Sunday
Thursday
Tuesday
Wednesday



qtdEnvios	campaignId	campaignName	createdAt	jart	Linear Regression	Random Forest	Fold	type	dayOfWeek	hour	listSize	isSegmented	openRate	clickRate	month
12780	CAMP00014	Campaign 14	2024-08-28 ...	3	13113.9	13325.8	1	Promotional	Wednesday	11	48920	Yes	36.7	5.51	8
5670	CAMP00018	Campaign 18	2024-08-16 ...	3	5309.96	5420.41	1	Newsletter	Friday	14	35460	No	27.4	3.56	8
13780	CAMP00031	Campaign 31	2024-10-06 ...	4	13523.9	13656.8	1	Promotional	Thursday	10	50670	Yes	37.8	5.67	10
2410	CAMP00040	Campaign 40	2024-07-16 ...	3	2669.96	2497.17	1	Welcome	Thursday	10	16890	Yes	30.5	3.97	7
7890	CAMP00046	Campaign 46	2024-08-21 ...	3	8034.13	8017.79	1	Newsletter	Tuesday	10	41670	Yes	35.8	5.19	8
5230	CAMP00020	Campaign 20	2024-07-22 ...	3	5565.19	5434.81	2	Reengageme...	Tuesday	11	32890	Yes	32.8	4.43	7
2650	CAMP00026	Campaign 26	2024-08-12 ...	3	2877.42	2421.53	2	Welcome	Friday	10	18670	Yes	29.4	3.71	8
5450	CAMP00027	Campaign 27	2024-10-15 ...	4	5237.23	5142.82	2	Reengageme...	Thursday	11	30560	Yes	33.2	4.48	10
2490	CAMP00033	Campaign 33	2024-09-26 ...	3	2610.35	2423.21	2	Welcome	Tuesday	9	17450	Yes	31.2	4.06	9
4780	CAMP00049	Campaign 49	2024-09-27 ...	3	5345.84	5005.32	2	Promotional	Saturday	17	31230	No	19.6	2.35	9
2180	CAMP00004	Campaign 4	2024-07-18 ...	3	1543.15	2461.07	3	Welcome	Wednesday	8	15670	Yes	28.9	3.47	7
3850	CAMP00005	Campaign 5	2024-09-30...	3	4281.14	5488.64	3	Reengageme...	Friday	11	31200	No	26.7	3.2	9
1980	CAMP00009	Campaign 9	2024-09-08...	3	1703.66	2289.28	3	Transactional	Wednesday	13	19450	No	20.45	2.25	9
4890	CAMP00013	Campaign 13	2024-09-17 ...	3	4359.37	5689.19	3	Reengageme...	Thursday	10	29340	Yes	31.5	4.1	9
12340	CAMP00038	Campaign 38	2024-08-29 ...	3	12474.7	11835.5	3	Promotional	Tuesday	11	48340	Yes	35.6	5.34	8
7650	CAMP00007	Campaign 7	2024-10-12 ...	4	7558.44	7464.99	4	Newsletter	Thursday	9	41230	Yes	33.8	4.73	10
2120	CAMP00016	Campaign 16	2024-07-12 ...	3	2905.2	2071.32	4	Transactional	Monday	17	21560	No	19.8	2.18	7
14890	CAMP00042	Campaign 42	2024-08-14 ...	3	14353.7	13721.6	4	Promotional	Wednesday	8	52670	Yes	40.2	6.03	8
2570	CAMP00047	Campaign 47	2024-10-11 ...	4	2724.54	2467.92	4	Welcome	Wednesday	11	18230	Yes	29.7	3.86	10
5340	CAMP00048	Campaign 48	2024-07-23 ...	3	5203.79	5233.91	4	Reengageme...	Friday	8	32120	Yes	31.8	4.29	7
6340	CAMP00010	Campaign 10	2024-08-19 ...	3	5936.65	5534.61	5	Newsletter	Friday	8	36780	Yes	31.2	4.21	8
13890	CAMP00017	Campaign 17	2024-09-24 ...	3	13418.3	13353.3	5	Promotional	Thursday	9	51230	Yes	37.2	5.58	9
5340	CAMP00025	Campaign 25	2024-09-28 ...	3	4638.57	5244.54	5	Newsletter	Monday	16	34230	No	25.6	3.2	9
5980	CAMP00032	Campaign 32	2024-07-19 ...	3	5890.15	5547.78	5	Newsletter	Friday	15	36890	No	28.1	3.65	7
5890	CAMP00035	Campaign 35	2024-10-21 ...	4	6258.27	5297.72	5	Promotional	Monday	18	29450	No	22.6	2.94	10
11245	CAMP00001	Campaign 1	2024-09-15 ...	3	10336.2	7768.24	6	Promotional	Tuesday	9	45230	Yes	32.45	4.87	9
13420	CAMP00006	Campaign 6	2024-08-14 ...	3	13936.9	13651.6	6	Promotional	Tuesday	10	52340	Yes	38.6	5.79	8
14670	CAMP00028	Campaign 28	2024-07-28 ...	3	14446.8	13727.4	6	Promotional	Tuesday	8	53890	Yes	39.8	5.97	7
5120	CAMP00034	Campaign 34	2024-08-17 ...	3	5080.86	5347.03	6	Reengageme...	Wednesday	8	31780	Yes	32.4	4.37	8
13230	CAMP00045	Campaign 45	2024-09-16 ...	3	12873.2	13394.5	6	Promotional	Thursday	9	49230	Yes	36.9	5.54	9

Anexo C - Código Python (Referência)

```python
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import r2_score, mean_squared_error, mean_absolute_error

# Carregar dados
df = pd.read_csv('email_campaigns.csv')

# Preparar features e target
features = ['type', 'dayOfWeek', 'hour', 'listSize',
            'isSegmented', 'openRate', 'clickRate', 'month']
target = 'qtdEnvios'

# One-hot encoding para categóricas
df_encoded = pd.get_dummies(df, columns=['type', 'dayOfWeek', 'isSegmented'])

X = df_encoded.drop([target, 'campaignId', 'campaignName',
                     'createdAt', 'quarter'], axis=1)
y = df_encoded[target]

# Dividir dados
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.3, random_state=42
)

# Treinar modelo
model = LinearRegression()
model.fit(X_train, y_train)

# Predições
y_pred = model.predict(X_test)

# Métricas
r2 = r2_score(y_test, y_pred)
rmse = mean_squared_error(y_test, y_pred, squared=False)
mae = mean_absolute_error(y_test, y_pred)

print(f"R²: {r2:.4f}")
print(f"RMSE: {rmse:.2f}")
print(f"MAE: {mae:.2f}")

# Coeficientes
coef_df = pd.DataFrame({
    'feature': X.columns,
    'coefficient': model.coef_
}).sort_values('coefficient', ascending=False)

print("\nCoeficientes principais:")
print(coef_df.head(10))
```

Anexo D - Dataset Melhorado (Amostra)

campaignId,type,dayOfWeek,hour,listSize,isSegmented,openRate,clickRate,qtdEnvios
CAMP001,Promotional,Tuesday,9,45230,Yes,32.45,4.87,11245
CAMP002,Newsletter,Thursday,10,38560,Yes,35.20,5.28,6890
CAMP003,Transactional,Monday,14,22340,No,22.15,2.65,2450
...