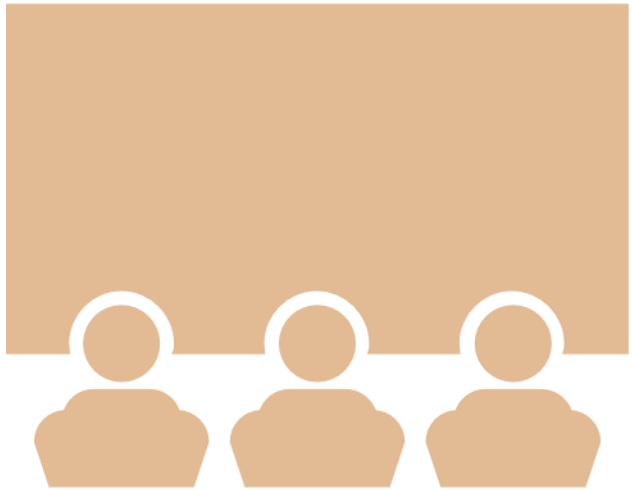
Data Science Capstone Project

Robson Leite da Cruz

https://github.com/yraion

01/11/2022

Outline



Sumario Executivo

- Dados coletados da API pública da SpaceX e da página da Wikipedia da SpaceX. Foi criado a coluna de rótulos 'class' que classifica os desembarques bem-sucedidos. Os dados explorados usando SQL, visualização, mapas de fólio e painéis. Colunas relevantes reunidas para serem usadas como recursos. Alterou todas as variáveis categóricas para binárias usando uma codificação quente. Dados padronizados e GridSearchCV usados para encontrar os melhores parâmetros para modelos de aprendizado de máquina. Visualize a pontuação de precisão de todos os modelos.
- Quatro modelos de aprendizado de máquina foram produzidos: Regressão Logística, Máquina de Vetor de Suporte, Classificador de Árvore de Decisão e K Vizinhos Mais Próximos. Todos produziram resultados semelhantes com taxa de precisão de cerca de 83,33%. Todos os modelos previram pousos bem-sucedidos. Mais dados são necessários para melhor determinação e precisão do modelo.

Introduction

SpaceX Falcon 9 Rocket – The Verge

Background:

- A Era Espacial Comercial Chegou
- Space X tem o melhor preço (US$ 62 milhões contra US$ 165 milhões)
- Em grande parte devido à capacidade de recuperar parte do foguete (Estágio 1)
- Space Y quer competir com Space X

Problema:

- O Espaço Y nos encarrega de treinar um modelo de aprendizado de máquina para prever uma recuperação bem-sucedida do Estágio 1

Métodologia

- Metodologia de coleta de dados:
- Dados combinados da API pública da SpaceX e da página da Wikipedia da SpaceX
- Execute a disputa de dados
- Classificar pousos verdadeiros como bem-sucedidos e malsucedidos de outra forma
- Realizar análise exploratória de dados (EDA) usando visualização e SQL
- Realize análises visuais interativas usando Folium e Plotly Dash
- Realize análises preditivas usando modelos de classificação
- Modelos ajustados usando GridSearchCV

Métodologia

VISÃO GERAL DA COLETA DE DADOS, DISCUSSÃO, VISUALIZAÇÃO, PAINEL E MÉTODOS DE MODELO

Data Collection Overview

O processo de coleta de dados envolveu uma combinação de solicitações de API da API pública do Space X e dados da web de uma tabela na entrada da Wikipedia do Space X.

O próximo slide mostrará o fluxograma de coleta de dados da API e o seguinte mostrará o fluxograma de coleta de dados do webscraping.

Colunas de dados da API Space X:

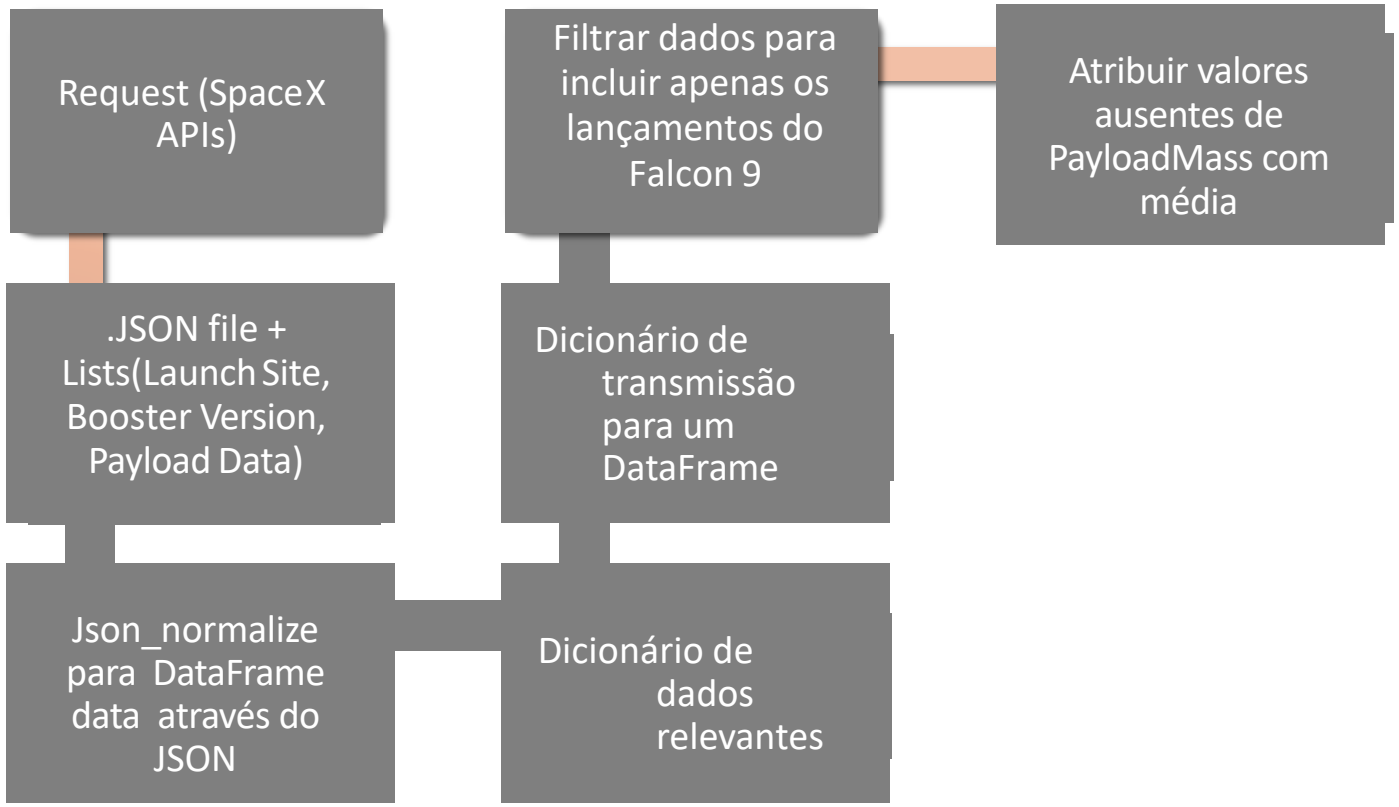
FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins,

Reutilizado, Pernas, LandingPad, Bloco, ReusedCount, Serial, Longitude, LatitudeWikipedia

Colunas de dados do Webscraping :

Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch result, Version Booster, Booster landing, Date, Time

Data Collection – SpaceX API



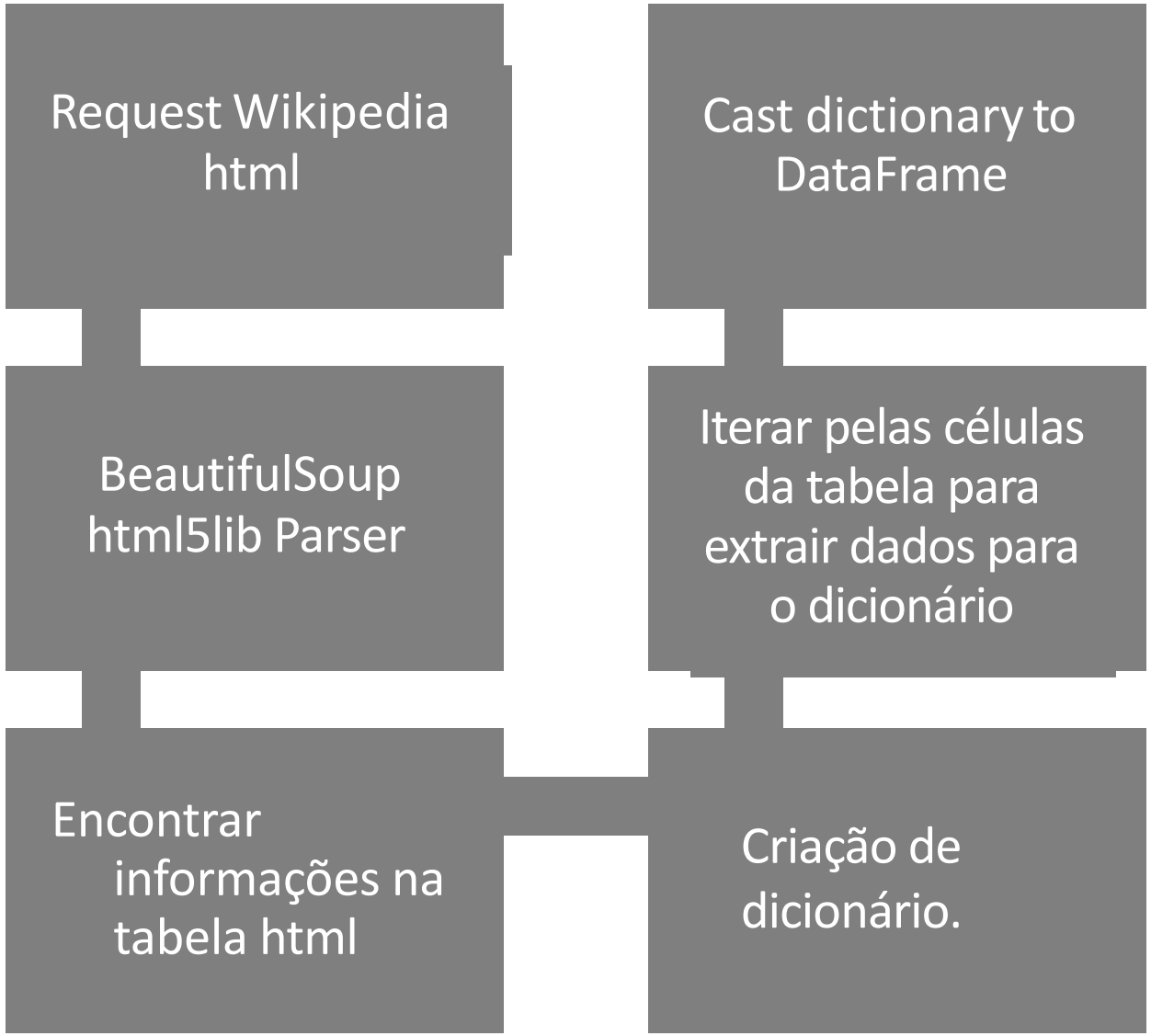
GitHub url:

https://github.com/yraion/IBM-Data-Science/blob/master/01-jupyter-labs-spacex-data-collection-api.ipynb

Data Collection – Web Scraping

GitHub url:

https://github.com/yraion/IBM-Data-Science/blob/master/02-jupyter-labs-webscraping.ipynb



Disputa de dados

Crie um rótulo de treinamento com resultados de pouso em que sucesso = 1 e falha = 0.

A coluna de resultado tem dois componentes: 'Resultado da missão' 'Local de aterrissagem'

Nova coluna de rótulo de treinamento 'class' com valor 1 se 'Resultado da missão' for True e 0 caso contrário. Mapeamento de valor:

True ASDS, True RTLS e True Ocean - definido como -> 1

Nenhum Nenhum, Falso ASDS, Nenhum ASDS, Falso Oceano, Falso RTLS – definido como -> 0

GitHub url:

https://github.com/yraion/IBM-Data-Science/blob/master/03-labs-jupyter-spacex-Data%20wrangling.ipynb

EDA com visualização de dados

Análise Exploratória de Dados realizada nas variáveis Flight Number, Payload Mass, Launch Site, Orbit, Class e Year.

Plots Used:

Número do voo versus massa da carga, número do voo versus local de lançamento, massa da carga versus local de lançamento, órbita versus taxa de sucesso, número do voo versus órbita, carga útil versus órbita e tendência anual de sucesso

Gráficos de dispersão, gráficos de linhas e gráficos de barras foram usados para comparar as relações entre as variáveis para

decidir se existe um relacionamento para que eles possam ser usados no treinamento do modelo de aprendizado de máquina

GitHub url: https://github.com/yraion/IBM-Data-Science/blob/master/06-jupyter-labs-eda-dataviz.ipynb

EDA with SQL

Conjunto de dados carregado no banco de dados IBM DB2.

Consultado usando a integração SQL Python.

As consultas foram feitas para obter uma melhor compreensão do conjunto de dados.

Informações consultadas sobre nomes de sites de lançamento, resultados de missões, vários tamanhos de carga útil de clientes e versões de reforço e resultados de pouso

GitHub url: https://github.com/yraion/IBM-Data-Science/blob/master/04-jupyter-labs-eda-sql-coursera.ipynb

Build an interactive map with Folium

Os mapas Folium marcam os Locais de Lançamento, pousos bem-sucedidos e malsucedidos e um exemplo de proximidade a locais-chave: Ferrovia, Rodovia, Costa e Cidade.

Isso nos permite entender por que os sites de lançamento podem estar localizados onde estão. Também visualiza pousos bem-sucedidos em relação à localização.

GitHub url: https://github.com/yraion/IBM-Data-Science/blob/master/07-lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

O painel inclui um gráfico de pizza e um gráfico de dispersão.

O gráfico de pizza pode ser selecionado para mostrar a distribuição de pousos bem-sucedidos em todos os locais de lançamento e pode ser selecionado para mostrar as taxas de sucesso do local de lançamento individual.

O gráfico de dispersão recebe duas entradas: Todos os sites ou site individual e massa de carga útil em um controle deslizante entre 0 e 10.000 kg.

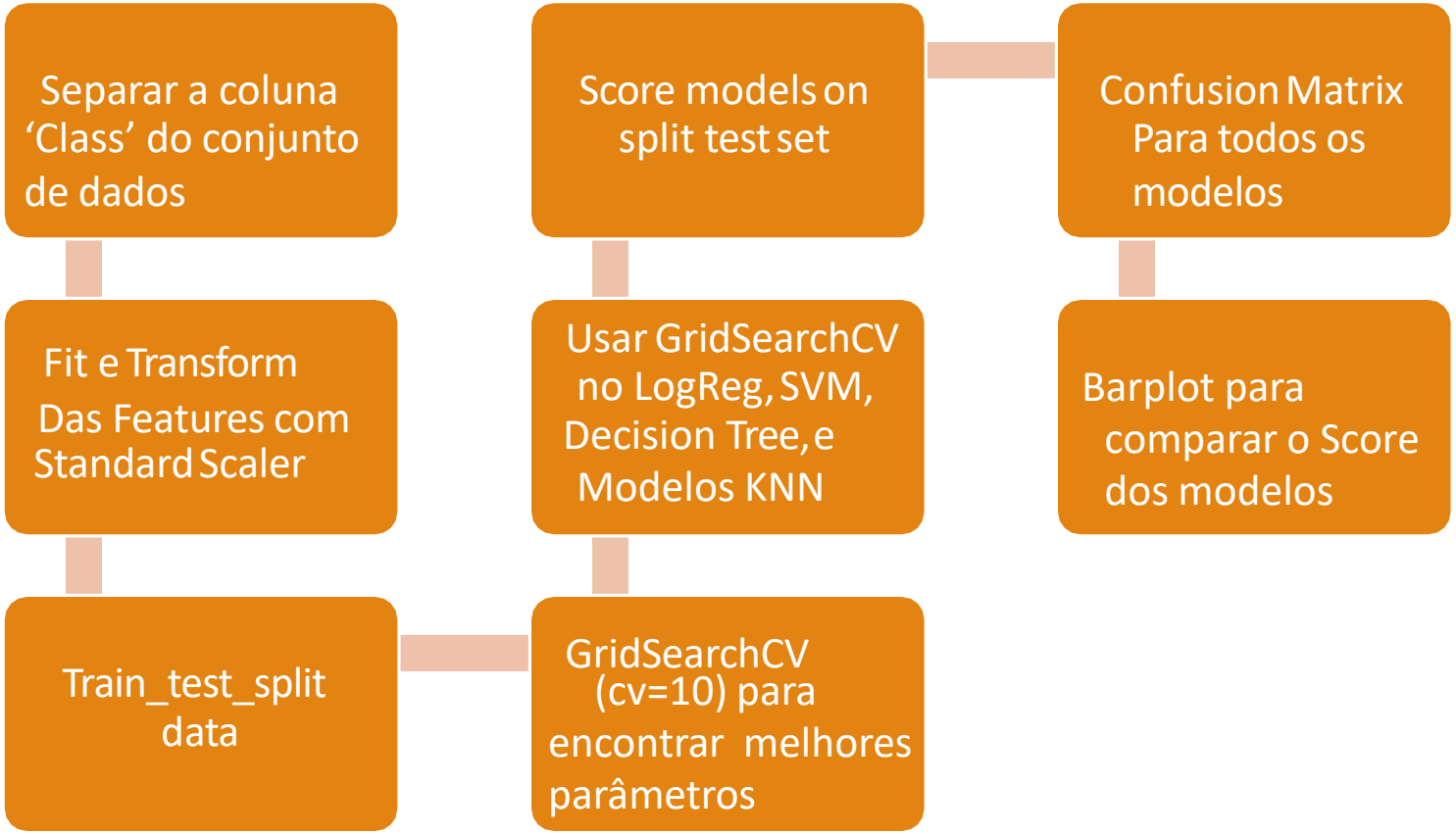
O gráfico de pizza é usado para visualizar a taxa de sucesso do site de lançamento.

O gráfico de dispersão pode nos ajudar a ver como o sucesso varia entre os locais de lançamento, massa de carga útil e categoria de versão de reforço.

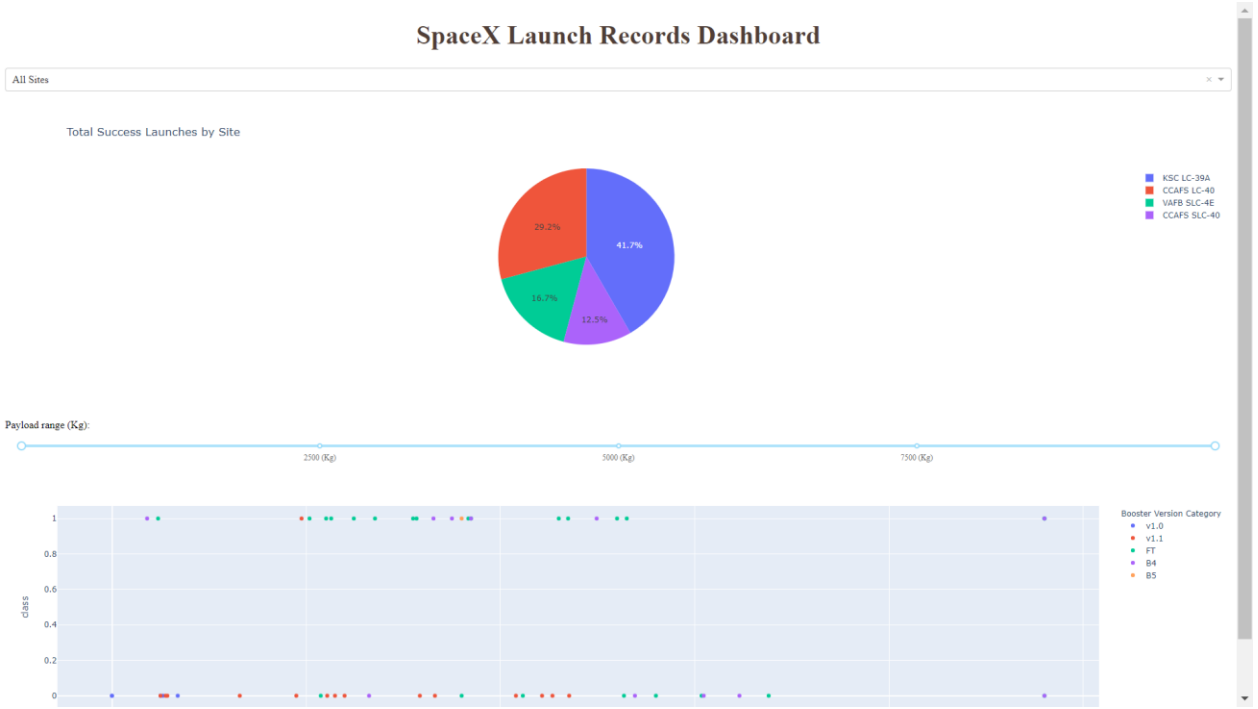
GitHub url:https://github.com/yraion/IBM-Data-Science/blob/master/06-jupyter-labs-eda-dataviz.ipynb

Predictive analysis (Classification)

GitHub url:https://github.com/yraion/IBM-Data-Science/blob/master/08-SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb



Resultados

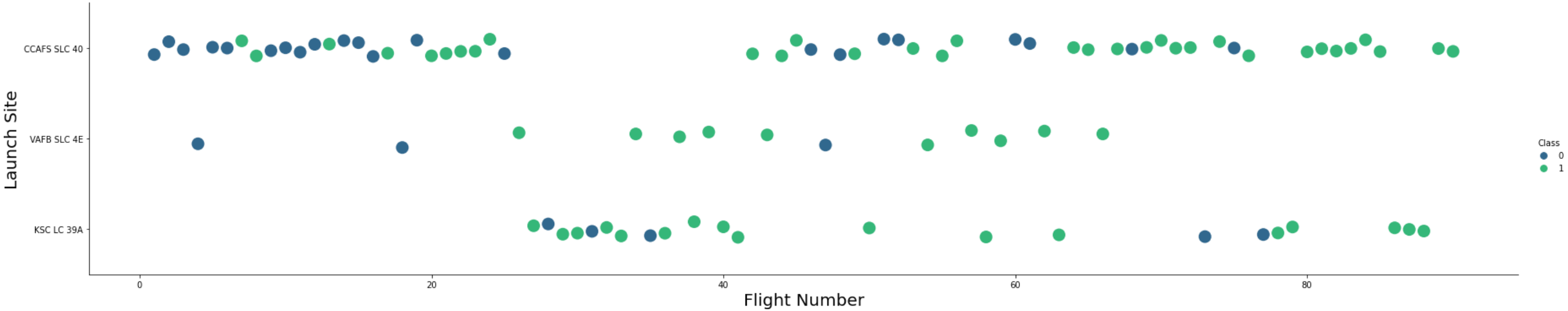


Esta é uma visualização do painel Plotly. Os lados a seguir mostrarão os resultados do EDA com visualização, EDA com SQL, Mapa Interativo com Folium e, finalmente, os resultados do nosso modelo com cerca de 83% de precisão.

EDA com Visualização

EXPLORATORY DATA ANALYSIS WITH SEABORN PLOTS

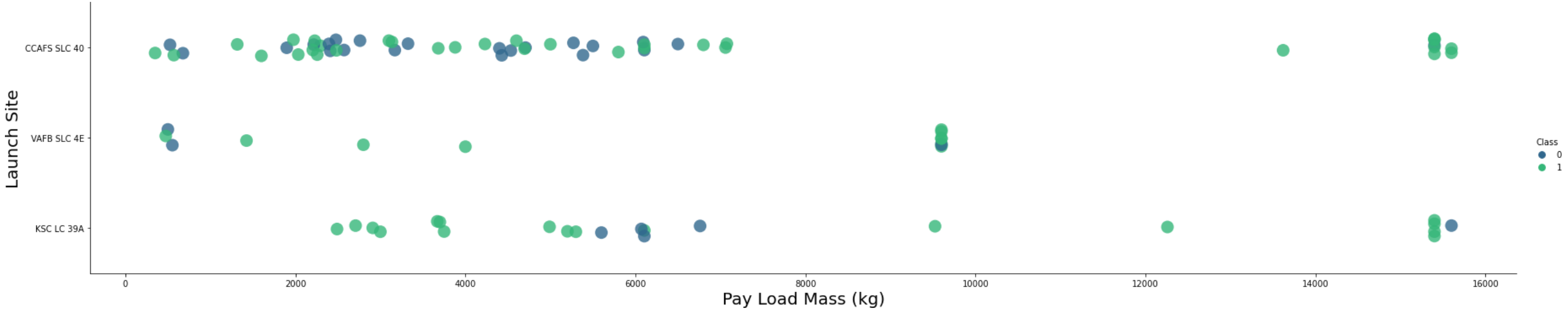
Flight Number vs. Launch Site



Verde indica lançamento bem sucedido; Roxo indica inicialização malsucedida.

O gráfico sugere um aumento na taxa de sucesso ao longo do tempo (indicado no número do voo). Provavelmente um grande avanço em torno do vôo 20, que aumentou significativamente a taxa de sucesso. O CCAFS parece ser o principal local de lançamento, pois possui o maior volume.

Payload vs. Launch Site

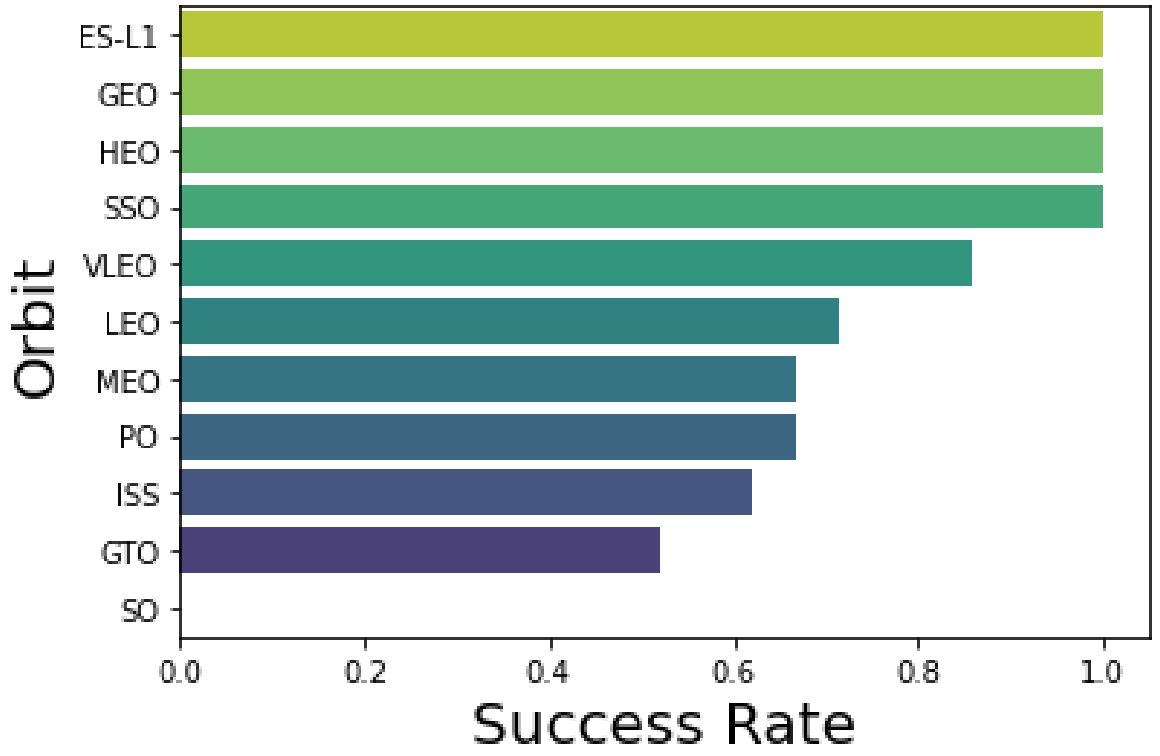


Verde indica lançamento bem sucedido; Roxo indica inicialização malsucedida.

.

A massa da carga útil parece cair principalmente entre 0-6000 kg. Diferentes sites de lançamento também parecem usar diferentes massas de carga útil.

Success rate vs. Orbit type



Escala de Taxa de Sucesso
0 as 0%
0.6 as 60%
1 as 100%

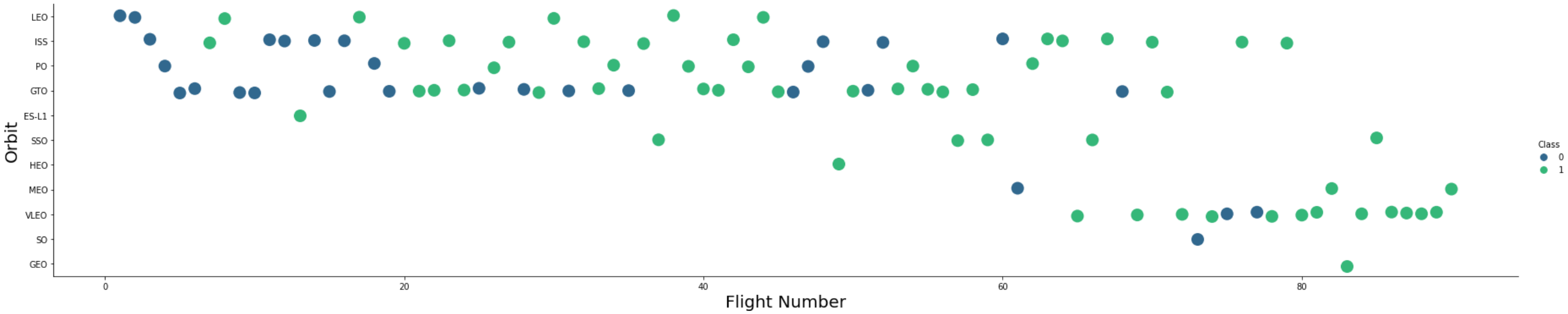
ES-L1 (1), GEO (1), HEO (1) têm 100% de taxa de sucesso (tamanhos de amostra entre parênteses) SSO (5) tem 100% de taxa de sucesso

VLEO (14) tem uma taxa de sucesso decente e tenta

SO (1) tem 0% de taxa de sucesso

GTO (27) tem a taxa de sucesso de cerca de 50%, mas também a maior amostra

Flight Number vs. Orbittype



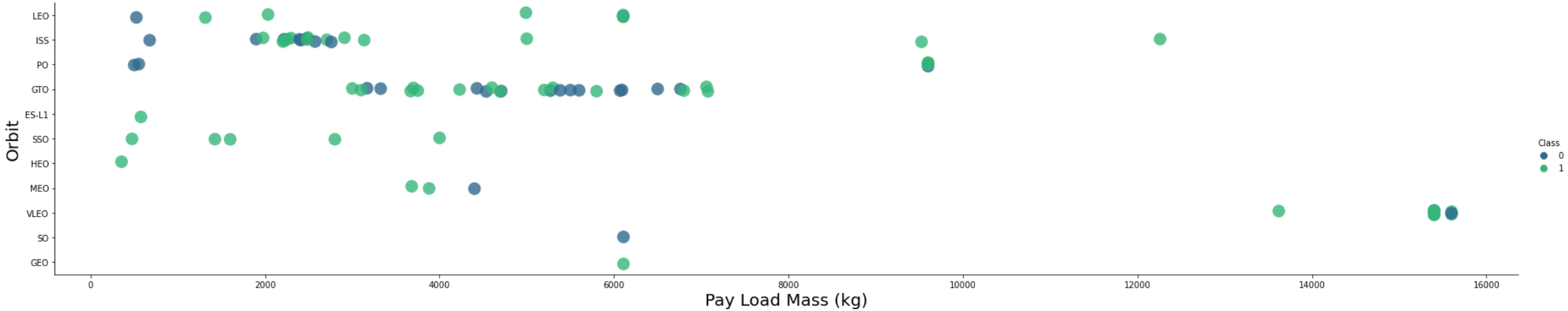
Verde indica lançamento bem sucedido; Roxo indica inicialização malsucedida.

.

As preferências de lançamento do Orbit foram alteradas em relação ao número do voo. O resultado de lançamento parece se correlacionar com essa preferência.

A SpaceX começou com órbitas LEO que tiveram sucesso moderado LEO e retornou ao VLEO em lançamentos recentes A SpaceX parece ter um desempenho melhor em órbitas mais baixas ou órbitas síncronas com o Sol

Payload vs. Orbit type



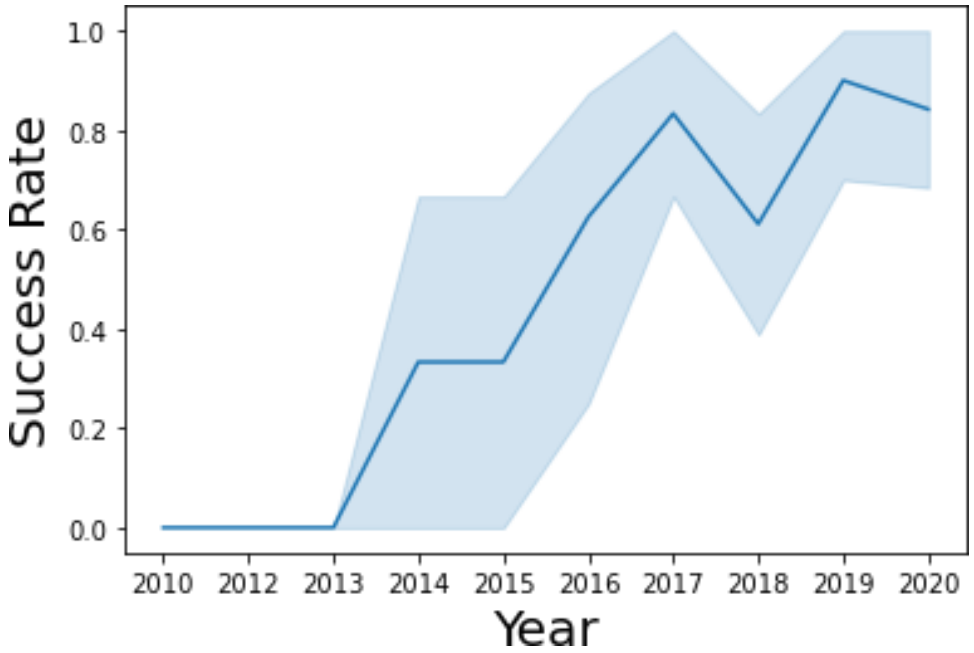
Verde indica lançamento bem sucedido; Roxo indica inicialização malsucedida.

A massa da carga parece se correlacionar com a órbita

LEO e SSO parecem ter uma massa de carga útil relativamente baixa

O outro VLEO orbital de maior sucesso tem apenas valores de massa de carga útil na extremidade superior da faixa

Launch Success Yearly Trend



Intervalo de confiança de 95% (sombreamento azul claro)

O sucesso geralmente aumenta ao longo do tempo desde 2013, com uma ligeira queda em 2018

Sucesso nos últimos anos em cerca de 80%

EDA COM SQL

EXPLORATORY DATA ANALYSIS WITH SQL DB2

INTEGRATED IN PYTHON WITH SQLALCHEMY

All Launch Site Names

```
In [4]: %%sql
SELECT UNIQUE LAUNCH_SITE
FROM SPACEXDATASET;
```

Done.

Out[4]:

launch_site
CCAFS LC-40
CCAFS SLC-40
CCAFSSLC-40
KSC LC-39A
VAFB SLC-4E

Consultar nomes de sites de lançamento exclusivos do banco de dados.

CCAFS SLC-40 e CCAFSSLC-40 provavelmente representam o mesmo

site de lançamento com erros de entrada de dados.

CCAFS LC-40 era o nome anterior. Provavelmente apenas 3 valores exclusivos do site de lançamento: CCAFS SLC-40, KSC LC-39A, VAFB SLC-4E

Launch Site Names Beginning with `CCA`

In [5]:
```
%%sql
SELECT *
FROM SPACEXDATASET
WHERE LAUNCH_SITE LIKE 'CCA%'
LIMIT 5;
```

Done.

Out[5]:

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Primeiras cinco entradas no banco de dados com o nome do Site de Lançamento começando com CCA.

Total Payload Mass from NASA

```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_) AS SUM_PAYLOAD_MASS_KG
FROM SPACEXDATASET
WHERE CUSTOMER = 'NASA (CRS)';
```

Done.

sum_payload_mass_kg
45596

Esta consulta soma a massa total da carga útil em kg onde a NASA era o cliente.

CRS significa Commercial Resupply Services, que indica que essas cargas foram enviadas para a Estação Espacial Internacional (ISS).

Average Payload Mass by F9 v1.1

```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_) AS AVG_PAYLOAD_MASS_KG
FROM SPACEXDATASET
WHERE booster_version = 'F9 v1.1'
```

Done.

avg_payload_mass_kg
2928

Esta consulta calcula a massa média de carga útil ou inicializações que usaram a versão de reforço F9 v1.1

A massa média de carga útil de F9 1.1 está na extremidade inferior de nossa faixa de massa de carga útil

First Successful Ground Pad Landing Date

```
%%sql
SELECT MIN(DATE) AS FIRST_SUCCESS
FROM SPACEXDATASET
WHERE landing__outcome = 'Success (ground pad)';
```

Done.

first_success
2015-12-22

Essa consulta retorna a primeira data de aterrissagem do bloco de solo bem-sucedida.

A primeira aterrissagem no solo não foi até o final de 2015.

Desembarques bem sucedidos em geral
aparecem a partir de 2014.

Successful Drone Ship Landing with Payload Between 4000 and 6000

```
%%sql
SELECT booster_version
FROM SPACEXDATASET
WHERE landing__outcome = 'Success (drone ship)' AND payload_mass__kg_ BETWEEN 4001 AND 5999;
```

Done.

booster_version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Essa consulta retorna as quatro versões de reforço que tiveram pousos bem-sucedidos de drones e uma massa de carga útil entre 4.000 e 6.000 não-inclusivamente.

Número total de cada resultado da missão

```
%%sql
SELECT mission_outcome, COUNT(*) AS no_outcome
FROM SPACEXDATASET
GROUP BY mission_outcome;
```

Done.

mission_outcome	no_outcome
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Esta consulta retorna uma contagem de cada resultado da missão.
A SpaceX parece atingir o resultado da missão quase 99% das vezes.
Isso significa que a maior parte do desembarque falhas são pretendidas.
Curiosamente, um lançamento tem um status de carga útil pouco claro e, infelizmente, um falhou em voo.

Boosters que carregavam carga máxima

```
%%sql
SELECT booster_version, PAYLOAD_MASS__KG_
FROM SPACEXDATASET
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXDATASET);
```

Done.

booster_version	payload_mass__kg_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600

Esta consulta retorna as versões de reforço que carregavam a maior massa de carga útil de 15600 kg.

Estas versões de reforço são muito semelhantes e todas são da variedade F9 B5 B10xx.x.

Isso provavelmente indica que a massa da carga útil se correlaciona com a versão de reforço que é usada.

Registros de desembarque de drones falhados em 2015

```
%%sql
SELECT MONTHNAME(DATE) AS MONTH, landing__outcome, booster_version, PAYLOAD_MASS__KG_, launch_site
FROM SPACEXDATASET
WHERE landing__outcome = 'Failure (drone ship)' AND YEAR(DATE) = 2015;
```

Done.

MONTH	landing__outcome	booster_version	payload_mass__kg_	launch_site
January	Failure (drone ship)	F9 v1.1 B1012	2395	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	1898	CCAFS LC-40

Essa consulta retorna o mês, o resultado de pouso, a versão do booster, a massa de carga útil (kg) e o local de lançamento dos lançamentos de 2015 em que o estágio 1 falhou ao pousar em um navio drone.

Houve duas ocorrências desse tipo.

Classificação das contagens de desembarques bem-sucedidos entre 2010-06-04 e 2017-03-20

```
%%sql
SELECT landing__outcome, COUNT(*) AS no_outcome
FROM SPACEXDATASET
WHERE landing__outcome LIKE 'Succes%' AND DATE BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY landing__outcome
ORDER BY no_outcome DESC;
```

Done.

landing__outcome	no_outcome
Success (drone ship)	5
Success (ground pad)	3

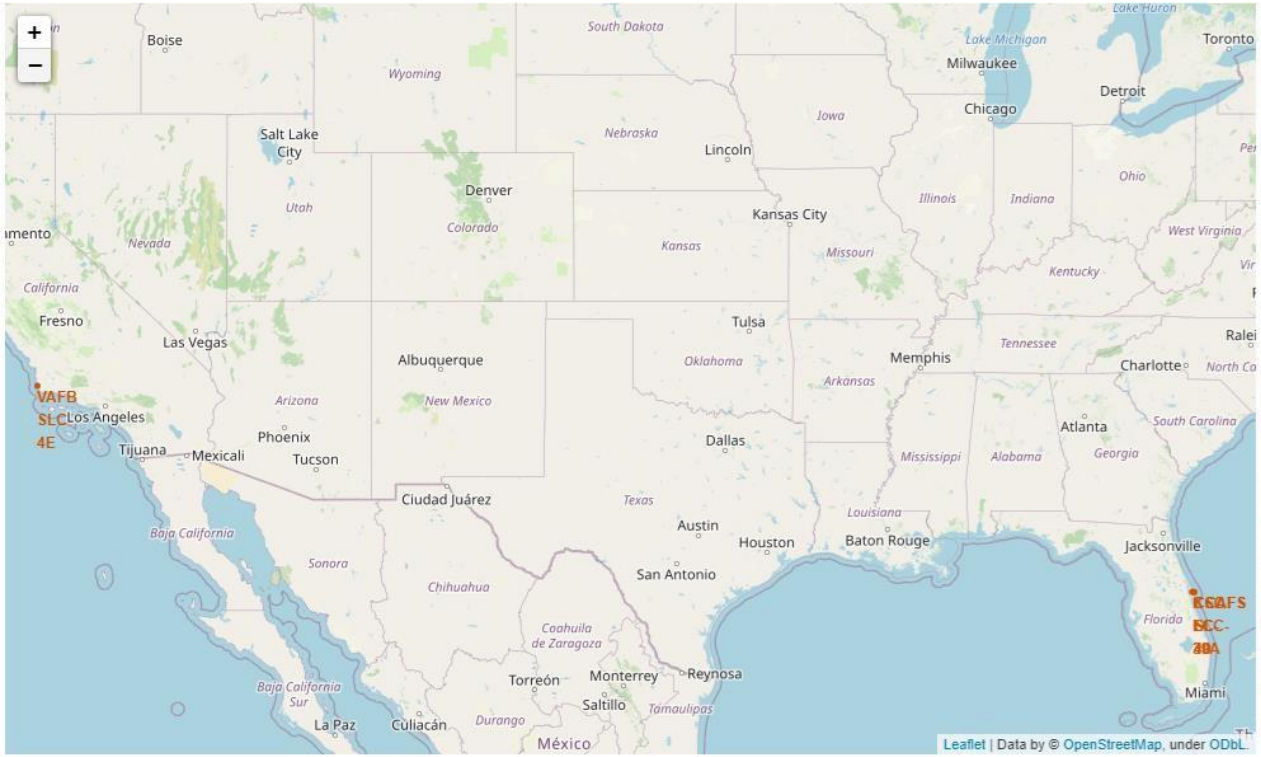
Esta consulta retorna uma lista de desembarques bem-sucedidos e entre 2010-06-04 e 2017-03-20 inclusive.

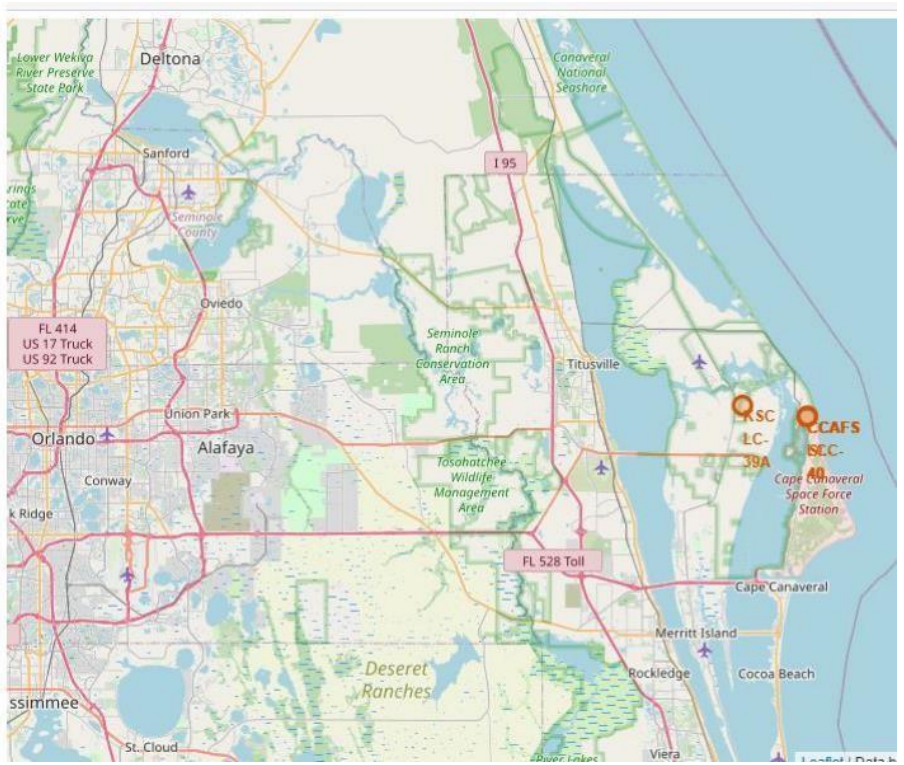
Existem dois tipos de resultados de pouso bem-sucedidos: naves drones e pousos no solo.

Houve 8 pousos bem sucedidos no total durante este período de tempo

Mapa interativo com Folium

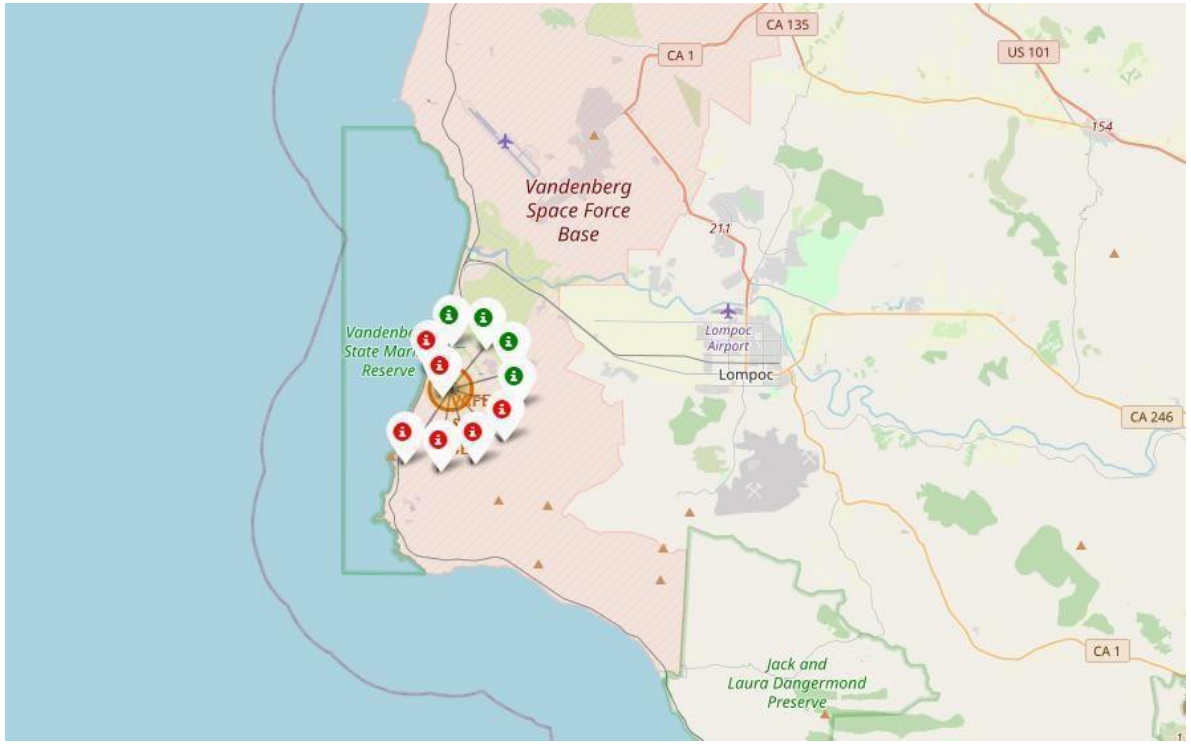
Launch Site Locations





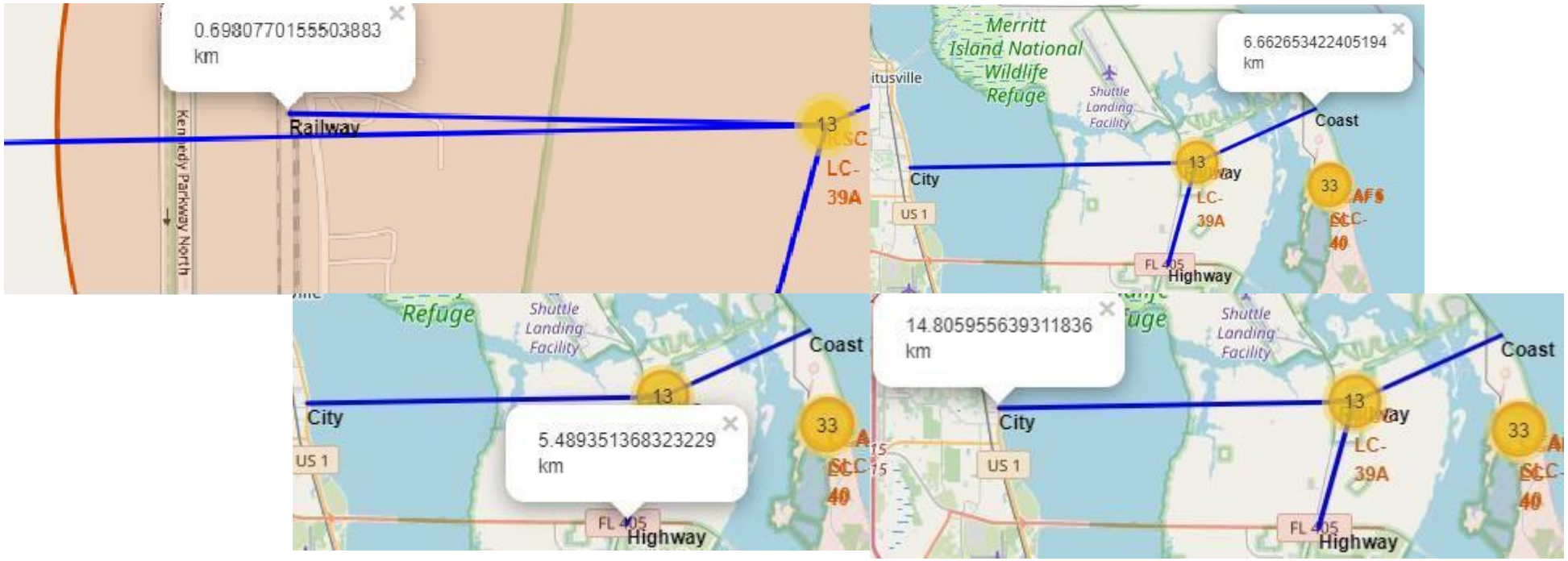
O mapa da esquerda mostra todos os locais de lançamento relativos ao mapa dos EUA. O mapa da direita mostra os dois locais de lançamento da Flórida, pois estão muito próximos um do outro. Todos os locais de lançamento estão perto do oceano.

Color-Coded Launch Markers



Clusters no mapa Folium podem ser clicados para exibir cada pouso bem-sucedido (ícone verde) e com falha pouso (ícone vermelho). Neste exemplo, o VAFB SLC-4E mostra 4 pousos bem-sucedidos e 6 pousos com falha.

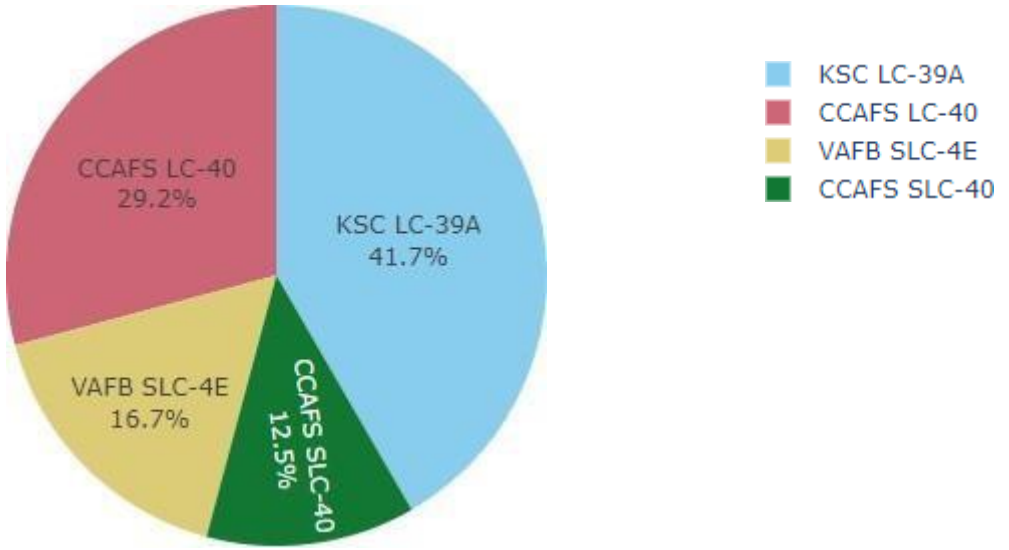
Principais Proximidades do Local



Usando o KSC LC-39A como exemplo, os locais de lançamento estão muito próximos das ferrovias para grande parte e transporte de suprimentos. Os locais de lançamento estão próximos a rodovias para transporte de pessoas e suprimentos. Os locais de lançamento também estão próximos às costas e relativamente longe das cidades, de modo que as falhas de lançamento podem pousar no mar para evitar que os foguetes caiam em áreas densamente povoadas.

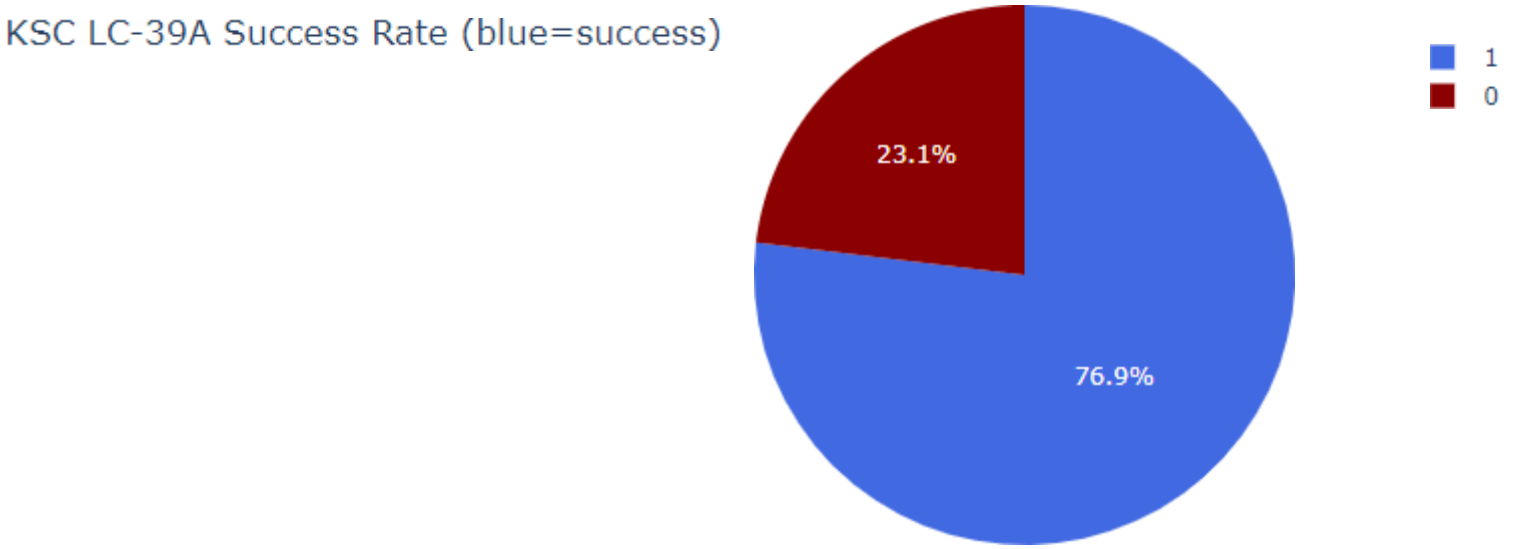
Construindo um Dashboard com o Plotly Dash

Lançamentos bem-sucedidos nos sites de lançamento



Esta é a distribuição de pousos bem-sucedidos em todos os locais de lançamento. CCAFS LC-40 é o nome antigo do CCAFS SLC-40, então CCAFS e KSC têm a mesma quantidade de pousos bem-sucedidos, mas a maioria dos pousos bem-sucedidos foi realizada antes da mudança de nome. VAFB tem a menor parcela de desembarques bem sucedidos. Isso pode ser devido à menor amostra e aumento da dificuldade de lançamento na costa oeste.

Highest Success Rate Launch Site



KSC LC-39A has the highest success rate with 10 successful landings and 3 failed landings.

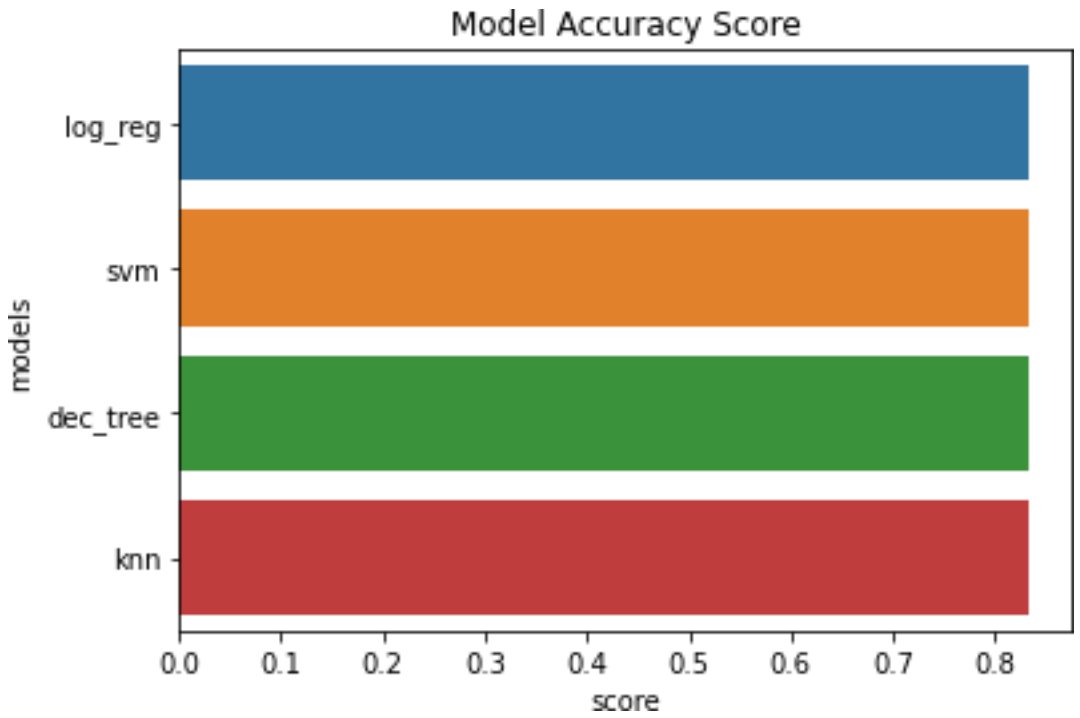
Payload Mass vs. Success vs. Booster Version Category

Plotly dashboard has a Payload range selector. However, this is set from 0-10000 instead of the max Payload of 15600. Class indicates 1 for successful landing and 0 for failure. Scatter plot also accounts for booster version category in color and number of launches in point size. In this particular range of 0-6000, interestingly there are two failed landings with payloads of zero kg.

Analise Preditiva (Classificação)

GRIDSEARCHCV(CV=10) na REGRESSÃO LOGISTICA, SVM, ÁRVORE DE DECISÃO, E KNN

Precisão da Classificação

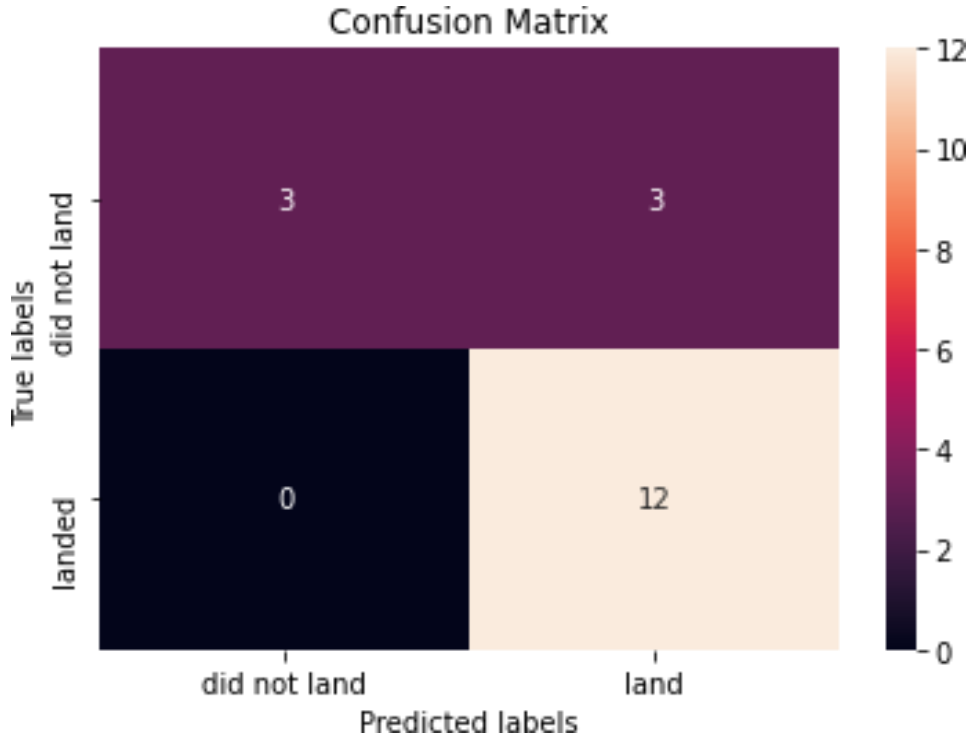


Todos os modelos tiveram praticamente a mesma precisão no conjunto de teste com 83,33% de precisão. Deve-se notar que o tamanho do teste é pequeno apenas com tamanho de amostra de 18.

Isso pode causar grande variação nos resultados de precisão, como os do modelo de classificador de árvore de decisão em execuções repetidas.

Provavelmente precisamos de mais dados para determinar o melhor modelo.

Matriz de Confusão



As previsões corretas estão em uma diagonal do canto superior esquerdo ao canto inferior direito.

Como todos os modelos tiveram o mesmo desempenho para o conjunto de teste, a matriz de confusão é a mesma em todos os modelos. Os modelos previam 12 pousos bem-sucedidos quando o rótulo verdadeiro era um pouso bem-sucedido.

Os modelos previram 3 pousos malsucedidos quando o rótulo verdadeiro era um pouso malsucedido.

Os modelos previram 3 pousos bem-sucedidos quando o rótulo verdadeiro era pousos malsucedidos

CONCLUSÃO

- Nossa tarefa: desenvolver um modelo de aprendizado de máquina para a Space Y que deseja concorrer à SpaceX
- O objetivo do modelo é prever quando o Estágio 1 chegará com sucesso para economizar ~ $ 100 milhões de dólares
- Dados usados de uma API pública da SpaceX e da página da Wikipedia da SpaceX na Web
- Rótulos de dados criados e dados armazenados em um banco de dados DB2 SQL
- Criei um dashboard para visualização
- Criamos um modelo de aprendizado de máquina com precisão de 83%
- Allon Mask da SpaceY pode usar esse modelo para prever com precisão relativamente alta se um lançamento terá um pouso bem-sucedido no Estágio 1 antes do lançamento para determinar se o lançamento deve ser feito ou não
- Se possível, mais dados devem ser coletados para determinar melhor o melhor modelo de aprendizado de máquina e melhorar a precisão

APENDIX

GitHub repository url: https://github.com/yraion/IBM-Data-Science

Instrutores:

Rav Ahuja, Alex Aklson, Aije Egwaikhide, Svetlana Levitan, Romeo Kienzler, Polong Lin, Joseph Santarcangelo, Azim Hirjani, Hima Vasudevan, Saishruthi Swaminathan, Saeed Aghabozorgi, Yan Luo

Agradecimento especial a todos os Instrutores e ao Corsera:

https://www.coursera.org/professional-certificates/ibm-data-science?#instructors