Carbon AIQ
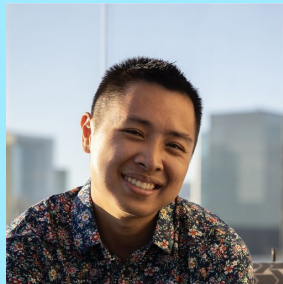
A Data Product for the Exploration and
Analysis of Carbon Offset Projects

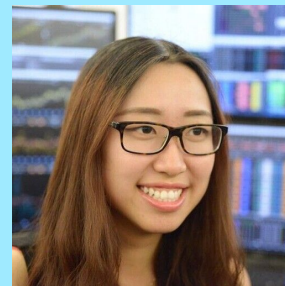# OUR TEAM

Kevin Hoang

• • •

Maggie Corry

Roshni Tajnekar

• • •

Roxy Rong

OUR BUSINESS PARTNERS

Barbara Haya, PhD

• • •

Director, Berkeley
Carbon Trading Project

Aline Abayo

• • •

Master's student at
Energy Resource Group
at UC Berkeley.

Berkeley Public Policy
The Goldman School
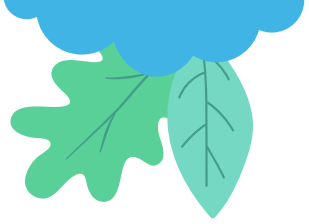
**Center for Environmental
Public Policy (CEPP)**

01

Overview

What is Carbon Offset?

Market Opportunity

Market size 2021

Compliance
$899B
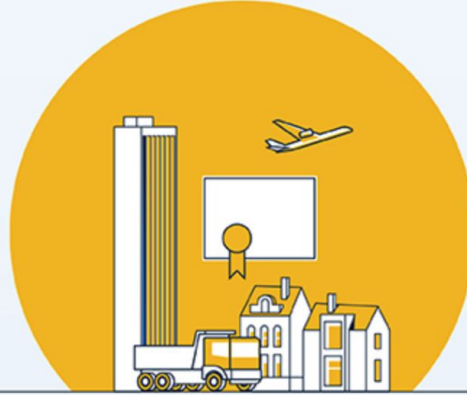
Voluntary
$2B

Sources: Refinitiv, Ecosystem Marketplace

As demand to cut emissions intensifies, voluntary carbon market volume has grown **five-fold in three years.**

Year	Value
2017	**$146M**
2018	**$296M**
2019	**$320M**
2020	**$520M**
2021	**$1,985M**

3-4X

The Status Quo

Fragmented

Inaccessible

Inconsistent

Standardized

Transparent

Reliable

Areas For Improvement
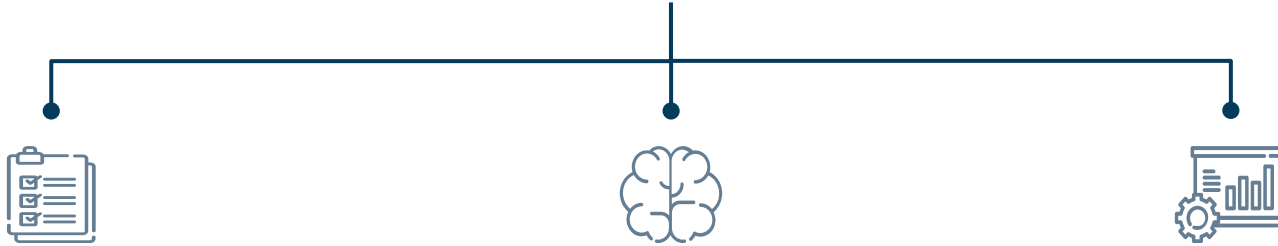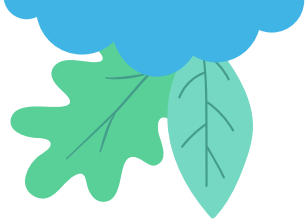
Automated Registry Data Integration

Consolidates project information and offset credit activities data

Project Type Classifier

Integrates machine learning with a rule-based system to classify projects by type

Interactive Visualization WebApp

Provides visual dashboards and tools to explore, analyze, and track information

02

CarbonAIQ Demo

Demo

03

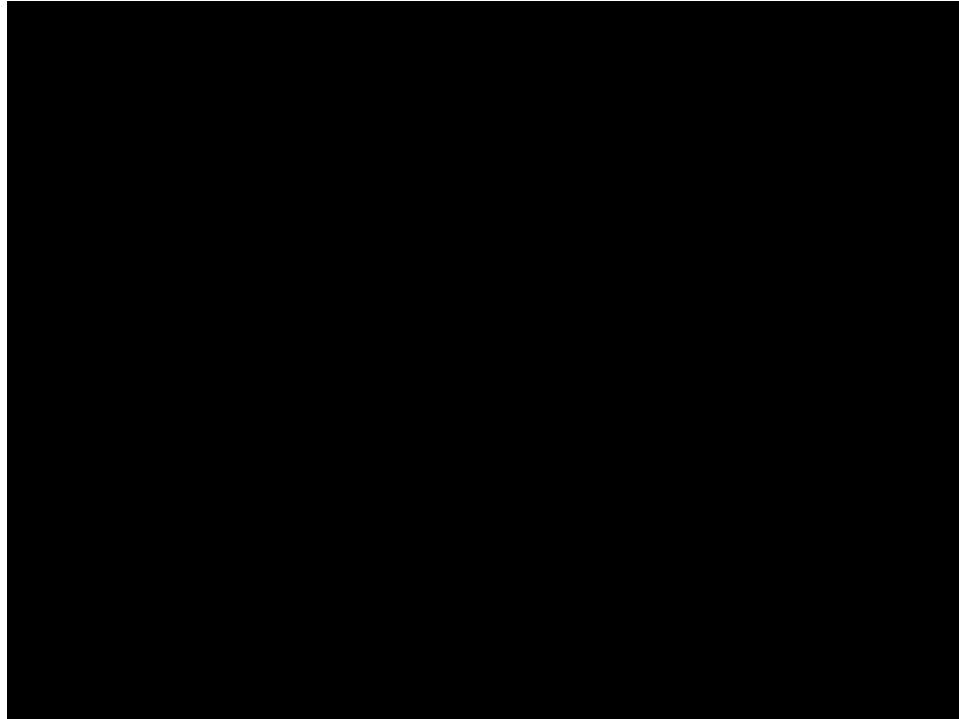## Data Product

Exploratory Data Analysis

AT A GLANCE

Multilingual project data across 6 registries, at 9,087 projects. Transaction data across all registries at 100K+ total records.

FEATURES

Continuous, categorical, text-based.

IMBALANCE

78 project types labeled by Berkeley Carbon Trading Project. Not evenly distributed - 55 types at <50 projects.

Frequency of Project Types
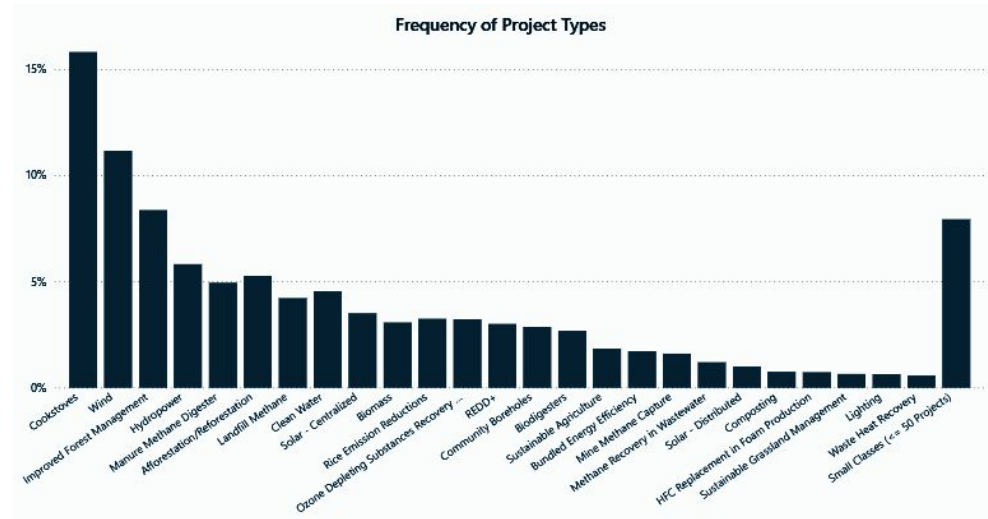
End-to-End Product Architecture

Storage
- Raw Project Data
- Raw Transaction Data
- Supplemental Dictionaries
- Log Files

Code
- Project Info Processor
- Compare with Old Versions
- ML Type Classification
- Rule-Based Type Classification
- Project Info CSV
- Credit Transaction Processor
- Credit Activity CSV

Database
- PostgreSQL
- Project Info Table
- Credit Activity Table
- Manual Change Log
- Archive

Frontend
- Amazon EC2
- Hosting
- Streamlit
- Charts and Visuals
- Web App
- Monitoring and Alerting

Data Processing Challenge, Visualized

Raw Data

Verified Carbon Standard Projects

ID	Name	Methodology	Country/Area
1716	Matebe Hydroelectric Plant	AMS-I.L.	Congo, The Democratic Republic of The

- **7** add'l tables/dictionaries to join and map fields
- Standardization formulas for ID and registry fields

Verified Carbon Standard Credits

Issuance Date	ID	Name	Total Vintage Quantity	Retirement / Cancellation Date	Quantity Issued
20/03/2023	1716	Matebe Hydroelectric Plant	43147	15/05/2023	1273

- Standardization of ID, date, transaction type, and subtype fields for all registries
- Complex operations to calculate amount - joins, lookups, group bys, encoding fixes, etc. for **22 subtypes**

Processed Data

Projects

project_id	project_name	voluntary _registry	methodology	country	arb_wa_ project
VCS1716	Matebe Hydroelectric Plant	**VCS**	**AMS-I.L. Electrification of rural communities using renewable energy**	**DRC**	**No**

Credits

project_id	transaction_type	sub_transaction_type	amount	date
VCS1716	**issuance**	**issuance**	43147	**3/20/2023**
VCS1716	**retirement_ cancellation**	**retirement**	1273	**5/15/2023**

04

Project Type Classification

Rule-Based System

How it works?
Check if fields equal or contain certain strings from rule table, get the type

Achievements
- **Hours** of manual work automated in a **5 minute** job
- **5,199** / 9,087 **projects** mapped
- **57%** coverage - how to improve?

Example: Project Data

Project ID	Project Name	Registry	Methodology	ARB/WA Project?
GS7795	Bundled **Wind** Power Project in the State of Gujarat	Gold Standard	**AMS-I.D. Grid connected renewable electricity generation**	**No**

Applicable Rule

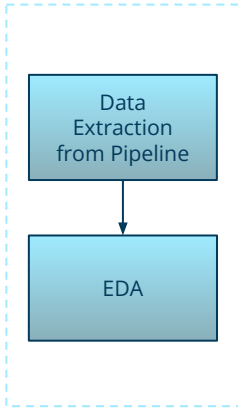| Project Name Logic	Project Name	Methodology Logic	Methodology	ARB Compliance Logic	ARB Compliance
Contains	Wind	Contains	AMS-I.D. Grid connected renewable electricity generation	Not	Compliance

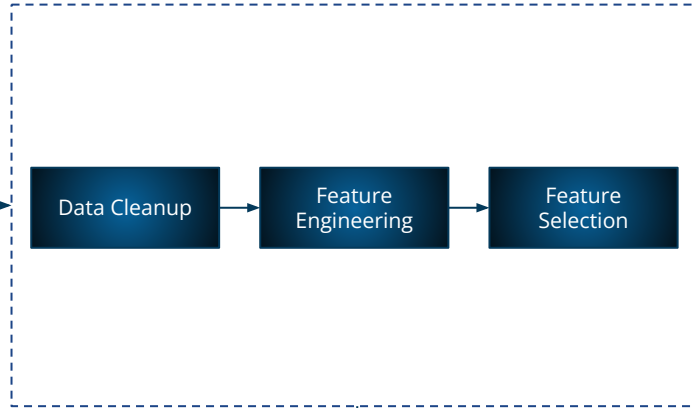Type Classification: **Wind**

05

ML/AL Models

ML/AI Pipeline

Data Extraction + Exploration

Data Preparation

Model Development

Data Extraction from Pipeline

EDA

Data Cleanup → Feature Engineering → Feature Selection

Train Set (70%)

Validation Set (10%)

ML/AI Models

Model Evaluation

Test Set (20%)
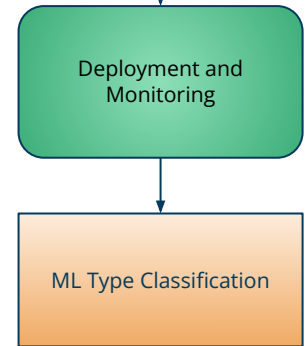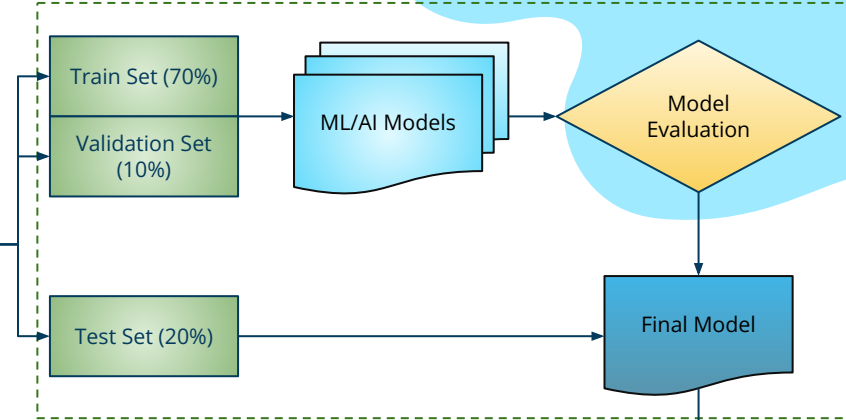
Final Model

Deployment and Monitoring

ML Type Classification

Initial data - 40 features, 9087 projects and 78 Project Types
Null data analysis - 24 features with >50% Null Values
EDA & Clustering - 7 features were unrelated to text classification
Removing minority classes - 19 classes with less than 6 records spread across 47 projects

Features Identified For Project Type Classification

TEXT FEATURES

• • •

- ❖ Project Name
- ❖ Project Developer
- ❖ Methodology
  - ➤ ACM0001 Flaring or use of landfill gas
- ❖ Project Type From Registry

CATEGORICAL

• • •

- ❖ Voluntary Registry (ACR, CAR..)
- ❖ ARB or WA project
- ❖ Region
- ❖ **Project Type (Target)**

06

Model
Experiments &
Evaluation

	Project Types	No. of Features	Preprocessing Steps	Train Size	Accuracy	Weighted F1-score	Training Time	Remaining Project Types
Rule-Based Classification (No ML)	61	2 Text, 2 Categorical	Methodology Remapping	9,087	57.21%	N/A	N/A	18 (148 Projects)
BERT MultiClass Classification (Base Model)	**46**	**4 Text**	**None**	**6,433**	**86.15%**	**9.16%**	**86 min (CPU)**	**32 (148 Projects)**
Canvas Text Prediction (Standard Build)	29	4 Text, 3 Categorical	Default	5,475	96.232%	92.951%	50 min (GPU)	49 (491 Projects)
BERT MultiClass Classification	59	4 Text, 3 Categorical	Methodology Remapping, Data Balancing	47,200	92.70%	93%	31 min (GPU)	19 (47 Projects)
Logistic Regression with BERT + One Hot Encoding	59	4 Text, 3 Categorical	Methodology Remapping, Data Balancing	47,200	82.18%	83%	9 min (GPU)	19 (47 Projects)
Logistic Regression with TF-IDF + One Hot Encoding	**59**	**4 Text, 3 Categorical**	**Methodology Remapping, Data Balancing**	**47,200**	**95.16%**	**94.96%**	**1 min (CPU)**	**19 (47 Projects)**

TEST RESULTS - LOGISTIC REGRESSION (TF-IDF)
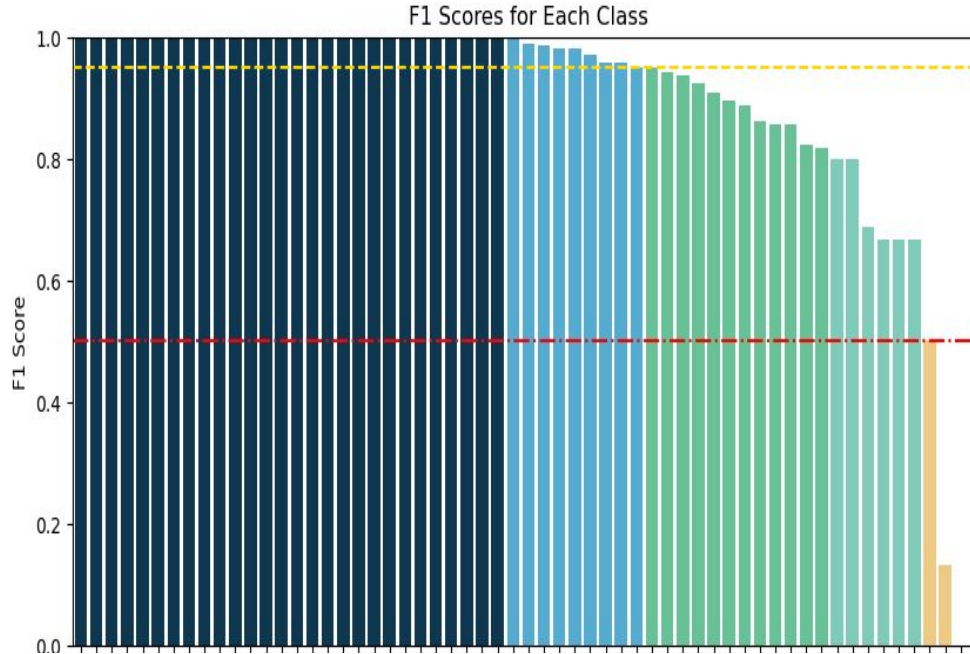
95%

F1 SCORE & ACCURACY

Classification report on the blind test data

92/1808

MIS-CLASSIFIED

Semantic Project Types, Different keywords used in Project Name & Methodologies

F1 Scores for Each Class

42/59

95-100% F1 SCORE

4/59

< 50% F1 Score

Low training data for minority classes

07

Closing

ACCOMPLISHMENTS

AUTOMATED DATA PIPELINE
· · ·
150+ hours per manual deployment to ~30 hours

EFFICIENT CLASSIFICATION
· · ·
Achieved extreme classification of 69 Project Types

ENHANCED USER INTERFACE
· · ·
20+ charts moved from Excel to web-based dashboards

READY-TO-DEPLOY PRODUCT
· · ·
Scalable to 100x projects and 10x users

Carbon AIQ, in partnership with Berkeley Goldman Public Policy