# Amplemarket Challenge

Filipe Marques

# Predicting Company Type using Random Forest and Embeddings

**Objective:** To classify companies into B2B, B2C, or Hybrid categories based on various features.

**Features**: Utilizing embeddings of 'Technologies', 'Specialties', 'Company Hubs', 'Industry', and 'Categories'

**Model**: Random Forest

# Simplifying Prediction Model Inputs

**Rationale**

- Focused Approach: Prioritizing key columns central to prediction objectives.
- Simplicity: Minimize complexity for ease of understanding and model interpretability.
- Relevance: Chosen columns directly contribute to crucial insights for accurate predictions.

**Methodology**

- Embeddings Creation: Transform categorical data into numerical representations.
- Training Data: Utilize the selected columns to train the model.
- Prediction: Leverage embeddings for predictions based on the simplified dataset.

# Why Random Forest for this classification?

- After rigorous testing, Random Forest demonstrated superior performance, outperforming other models in terms of accuracy and efficiency
- **Ensemble Learning**: Random Forest is an ensemble learning method that combines multiple decision trees for robust predictions
- **Advantages:** High accuracy, robust to overfitting, inherent feature importance
- **Flexibility in Tuning:** Allows easy hyperparameter tuning for optimization

# Optimizing Random Forest for Classification

**Data Splitting:** Split data into training and evaluation sets (80/20)

**Standardization**: Utilized StandardScaler to standardize features for consistent model training

**Balanced weights**: Initialized Random Forest model with balanced class weights

**Hyperparameter Tuning:** Employed RandomizedSearchCV to explore optimal hyperparameters

# End-to-End Pipeline with Dockerized Model

**Training Process**: Jupyter Notebook for model training

**Docker Containerization**: Dockerized the entire workflow for portability and reproducibility

**Task-Based Architecture:** Divided the pipeline into tasks:

Preprocess Task: Handles data preprocessing before feeding it to the model.

Predict Task: Executes model predictions on preprocessed data.

**API Integration:**

Developed an API for easy integration into various applications

API tasks include preprocessing and model prediction, encapsulated for simplicity

# Future Work

- Add Features: Considering incorporating more features
- Add unit tests
- Create train pipeline, using the already implemented entrypoint
- Keeping Random Forest as algorithm embeddings don’t need to be created manually

# Questions