DAREDATA ENGINEERING

Our Projects and Expertise.

We've had the luck to work in wonderful projects with great teams. **Either building our own solutions, helping customers raise the bar on their data iniatives or kickstarting proof-of-concepts, we're proud of everything we've built.**

Our Solutions.

DareData Engineering is a startup company providing consulting and outsourcing offers in Data Science & Engineering. We were born in 2019 and have built more than 75 Data Science and Machine Learning systems for companies of all sizes.

+75

projects delivered in 3 years

Our team is carefully curated to be able to assist with your data challenges.

Meet our solutions

DAREDATA ENGINEERING

Index

Telecommunications

- Cross Sell Recommendation
- Call Center Relapses

Retail

- On and Off Trade Forecasting
- Product Adoption
- Routing Optimization
- Promo Efficiency
- Assortment Optimization

Pharma

- Auto-Labelling Invoices
- Microscopy Consensus Map

Public Sector

- Data Lake Infrastructure Setup and Maintenance
- Company Growth Forecast

Utilities

- Roof Detection

Startups/Scale Ups

- Language Model
- Reporting Infrastructure
- Data Science and Engineering Advisory

DAREDATA ENGINEERING

Cross Sell Recommendation Engine

Summary:

We've built a recommendation engine for one of the major Telecommunication companies in Iberia. Our client had a major issue identifying customers they should contact with new offers for premium TV channels and outbound calls ended up turning into complain calls by the clients.

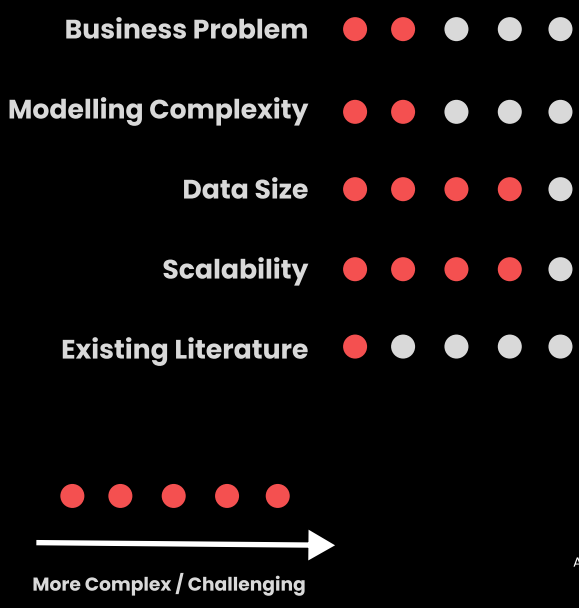
Our job was to predict which customers were more likely to buy the new products. Additionally, we needed to consume a huge amount of past data from the customer to set up business rules on top of our recommendations.

The Model

We've used a gradient boosting model that incorporates client characteristics, current customer packs and television usage, among others.

Ensuring Validity

One of the hardest jobs in high dimensionality problems is to ensure causality. We made sure that it was clear (for our Customer) that we needed to setup a controled experiment to make sure we could scale the solution.



What we've learned with this project:

- It's important to understand what is the metric we should beat so that business users are comfortable with our recommendations. For this to work, we needed to improve cold calls success rate.

Call Center Relapses

Summary:

We've built a technical call relapses prediction system for one of the major telecommunication companies in Iberia. Currently, 30% of their calls generated a relapse (a new call by their customer) and hit rate was low. These customers were at a higher risk of churning as they were already having a lot of issues with their service. The company's proactivity was the only measure that could improve the customer satisfaction.

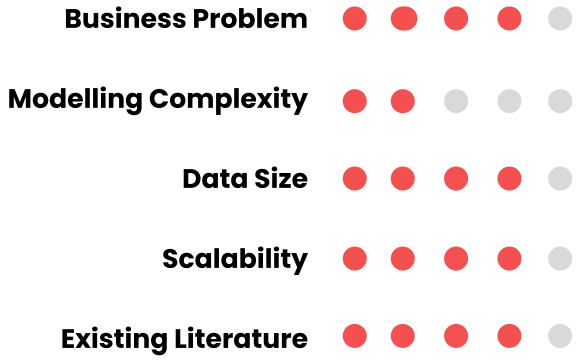
As all customers in the training base had a technical issue with the service in the past, we wanted to maximize for precision.

The Model

We've used a gradient boosting model that incorporates service level and customer interactions to predict relapses.

Defining the Metric

Using Precision for this problem was the right decision. The customer had a limited capacity of pro-actively calling customers so we needed to make sure that for the 1.000 customers we will select based on our model, most of them are indeed going to relapse in the next few days.



What we've learned with this project:

- Most Data Science problems are technically simple. The hard part is to find the right metric to optimize. This should ALWAYS be done with the business users input.

Sales Forecasting (On and Off Trade)

Summary:

Our sales forecasting machine learning algorithms are currently being deployed in two markets of one of the major beverage companies in the world.

We've done production-level forecasts, deployed using MLOps best practices, both for On-Trade (Restaurants, Cafes, etc.) and Off-Trade (Retail). We've achieved significant improvement in current MAPE (Mean Average Percentage Error), particularly in the On-Trade business.

The Model

We've tested both SARIMAX and Prophet models. The models are retrained to avoid overfitting and to capture significant changes in consumers demands.

We've used sales, weather, soccer games, holidays and other factors to capture consumers demand.

Over 500 Models

We have deployed models for more than 250 SKUs for both businesses. It's impossible to tackle and manage these models without proper MLOps techniques and working CI/CD pipelines.

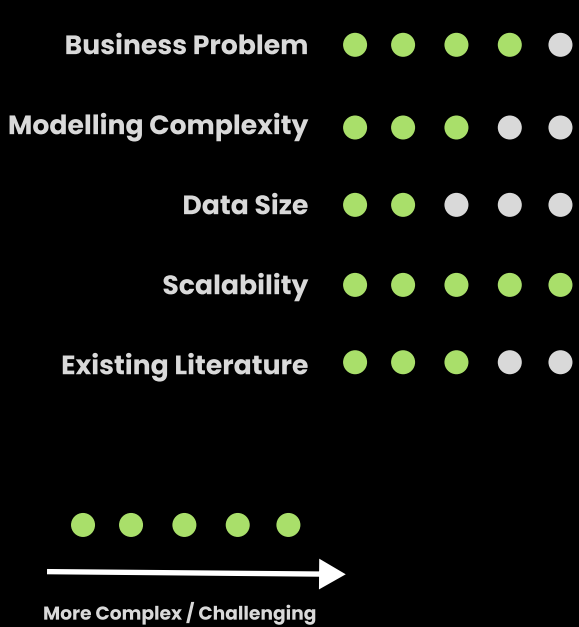


Figure 1 - Example of Features (Endo and Exogenous) used in the forecast.

Product Adoption

Summary:

Using a nearest neighbours approach, we've helped launch and push new products, influencing the consumer behavior on the mid-long term, achieving significant margin growth for our client while improving marketing efforts.

As a new product is launched and pushed for certain Points-of-Sale (POS), we leverage the available data about those POS to target similar ones and focus the marketing and sales efforts for a specific product.

The Model

We've used a two stage model approach:

- First, we classify existing POS based on their affinity with the brand.
- We then understand how many POS are "surrounded" in a multi-dimensional space by the POS with more affinity.

From Model to the Real World

Our dimensional space consists of explainable variables related with the POS relationship with our customer but also contextual information (street information, average income in the nearby neighbourhoods, etc.). It was easy to explain why a specific POS was being recommended, something that was very well received by the sales team.

Business Problem ● ● ● ● ○

Modelling Complexity ● ● ○ ○ ○

Data Size ● ● ○ ○ ○

Scalability ● ○ ○ ○ ○

Existing Literature ● ● ● ● ●

What we've learned with this project:

- Having back and forths with the business was crucial to understand misclassified examples. Although these examples were not misclassified in terms of a supervised "apprach", they helped us understand data quality issues and which POS we were recommending were a bit "off", making it easier to tweak the solution.

● ● ● ● ●

More Complex / Challenging

Routing Optimization

Summary:

Building a Routing plan is always a tricky project for most companies. The need to balance between Resources and ensuring each Customer's need is always hard to tackle.

We've optimized commercial routes for several of our retail customers, helping them manage the workload between the team and ensuring the maximum coverage possible.

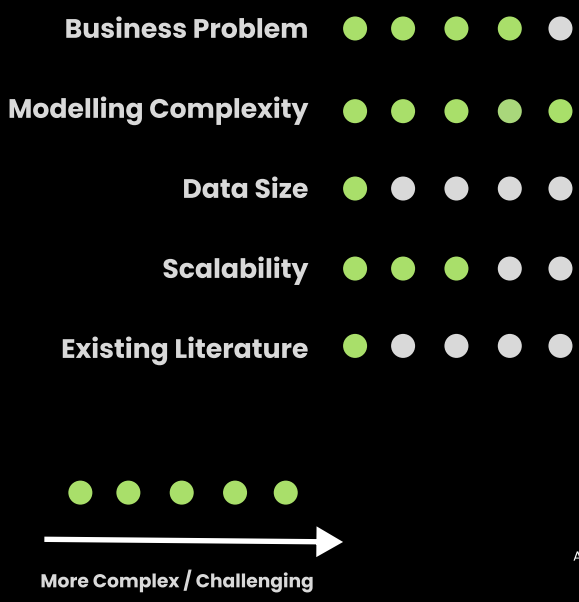
The Model

Using OSRM (Open Source Routing Machine) library we were able to obtain the shortest path between our geo-referenced points of sale (POS).

From here, we could build a routing engine that took into account several constraints such as number of times a POS should be visited and POS closing days.

Ensuring Coverage

Before our solution, our customer had a 50% to 60% coverage of all the POS they needed to visit. Our solution is able to maximize this coverage to around 90% while ensuring that top customers have a priority in the visit plan. Our optimization algorithm tried to balance coverage and the time for each visit.



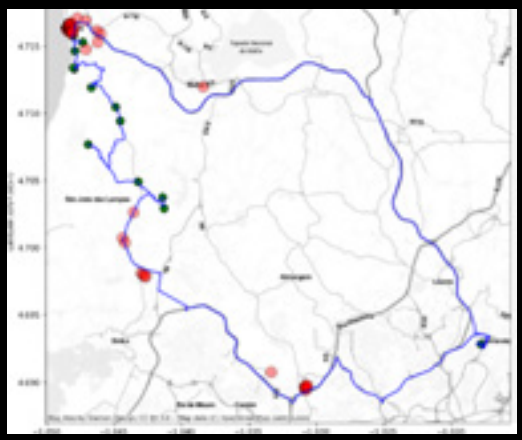


Figure 2 - Example of an Optimized Route produced by our algorithm.

Promo Efficiency

Summary:

Leveraging data to boost promo efficiency is one of the quickest measures to improve retailers margins. We've built a model that incorporates seasonality data and product features to recommend the optimum price according to two axis:

- Improving Margin;
- Retaining Market Share;

For retailers, the optimum point between these two variables is their bread & butter. Due to the high dimensionality of the business, achieving that sweet spot is impossible without leveraging machine learning or statistics.

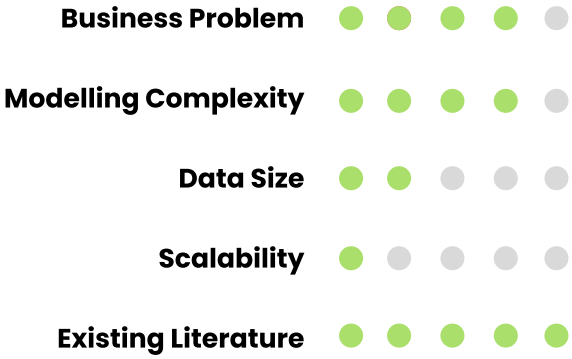
The Model

We've tried a forecasting approach but due to the inability of our customer to change prices so quickly, we could only recommend monthly promos. Tree-Based models were more efficient to do that.

Listening to the Business

Although we know that a forecasting approach would work better, our customer couldn't act on it.

This is the perfect use case where there is a conflict between model development and business perspective - in the end, we always prioritize the business view.



What we've learned with this project:

- Retail's High Dimensionality is a really complex issue. We really had to protect for a high variance model here. This was really important because our customer didn't have the ability to retrain the model often (they acquired data every quarter).

Assortment Optimization

Summary:

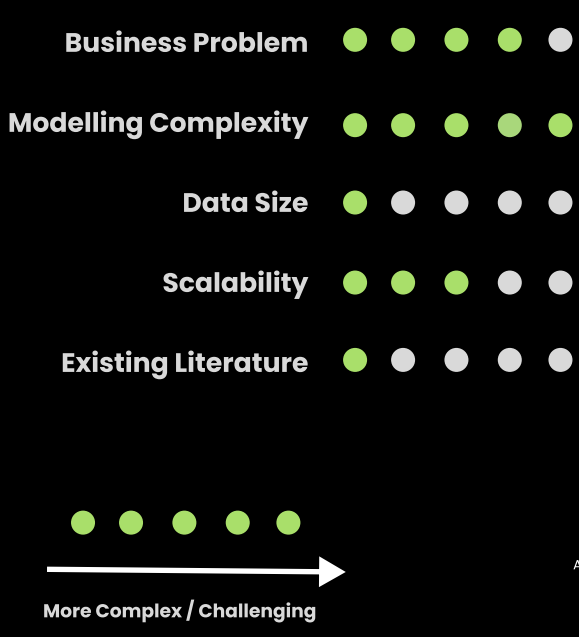
When a company sells a lot of products physically via other businesses, understanding if the product you are selling is the right one for that market is always challenging.

Our assortment optimization method helped our customer map which POS's needed different segments of products (premium, economy, etc.) to achieve target margin and market share while nailing the product that the consumer was searching for in each specific POS.

The Model

First, we used a clustering algorithm to group similar POS.

After clustering, we ran an optimization network to maximize the target sales as a function of the SKU's. This will give us a mapping of the % we should alocate to each SKU type.



Known Unknowns

As our customer does not have 100% market share in all POS, we don't want to include POS where we only know a small % of the market (for instance, they can buy a really big chunk of SKU's from a competitor).

This is where the clustering algorithm was useful! We could optimize for the POS where we are sure that we know most of the "demand" and apply the assortment rules for the cluster as a whole (even for lower share POS).

What we've learned with this project:

- Literature in the field is pretty opaque. We had to tackle the product with some custom solutions and ended up choosing the best hypothesis.

Auto-Labelling Invoices

Summary:

Operational efficiency is key for most industries, specially for pharma. With a lot of sales rep on the road, it's essential that their day-to-day job is largely spent on selling.
Our customer had a manual process that consumed a lot of sales rep time, providing almost zero added value. At the end of each week, they spent around 3 to 4 hours just labelling invoices from their work week. We've developed a model that took into account the invoice info to suggest categories such as "Hotels", "Restaurants", "Transportation", "etc." with only the sales rep approving the suggestion.

The Model

An XGBoost was built on top of info from the invoice. We've achieved more than 90% accuracy on labelling the correct category of the invoice.

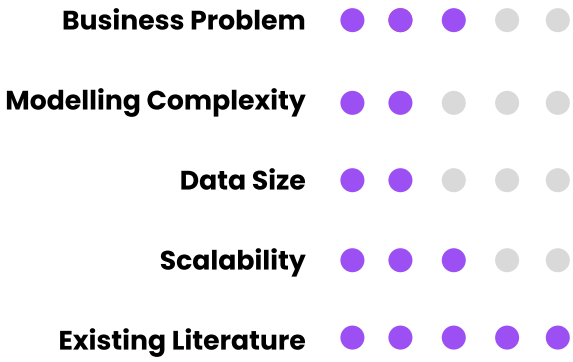
Big Impact in Small Processes

The most successful machine learning and data science projects tend to change small processes, incrementaly adding value to the organization.

Who knew that just by adding a small suggestion on the invoice labelling would bring coutless of more hours for sales representatives do to their value-added job?

What we've learned with this project:

- There are certain products that are able to do some of the custom solutions we've developed. Although they have a plug-and-play approach they have some shortcomings such as vendor lock-in or lack of integration with your tech stack.



Microscopy Consensus Map

Summary:

Physicians have a hard time to target cancer cells based on microscopic images. Currently, our customer wanted to build a consensus map across several images hand-labeled by physicians targeting suspicious cells.

We've developed a clustering algorithm able to build a consensus map for the same image - these images can now be used for a better diagnostic but also as training data for other ML models.

The Model

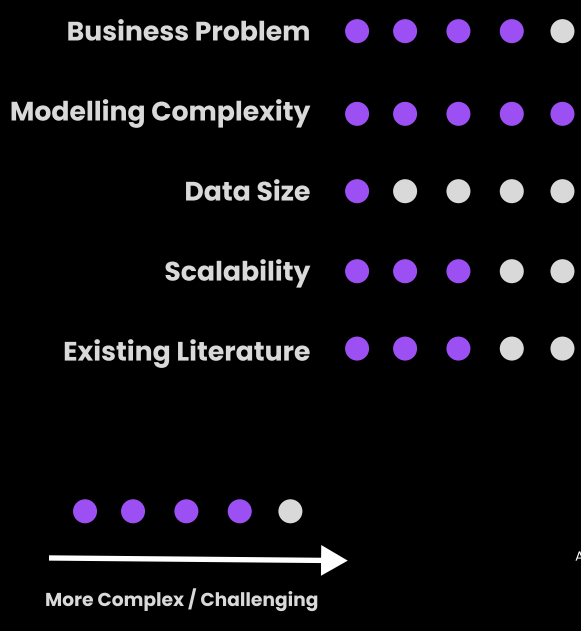
We've used an image clustering algorithm, targeting areas that were considered by several physicians as suspicious.

Above our clustering algorithm, we had a couple of heuristic rules running to lower the error rate.

POC Basis

This project was a good example of a simple POC that was given to the customer for further improvements.

The customer was pretty happy with our first iteration and is now able to use our model to train more advanced ML supervised models on their end with more accurate data.



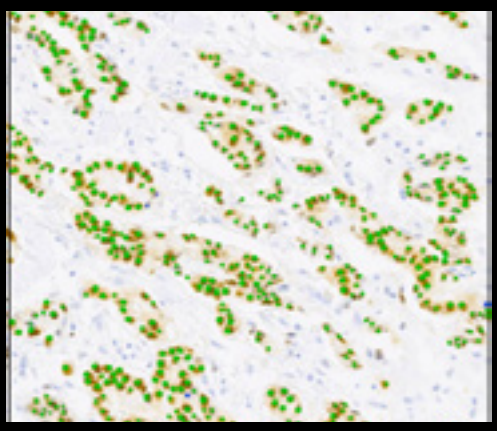


Figure 3 - Example of Application of Clustering Method to Target Likely Cancer Cells

Data Lake Setup and Maintenance

Summary:

One of our most challenging projects was to completely transform the database infrastructure from a big government agency.

We completely redesigned the whole architecture, enabling our client to be able to use data in a variety of formats - such as text and even video data. Our data lake is currently being used by hundreds of people inside the organization, feeding multiple applications that are able to process huge chunks of data in the cloud.

Digital Transformation

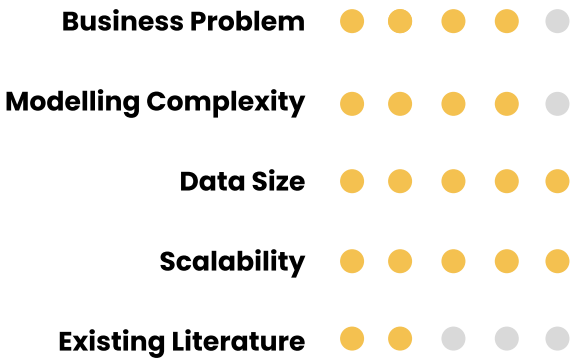
Helping traditional organizations becoming more digital is not an easy feat.

We were the main advocates towars using Data Lake and ML solutions at the customer with all the challenges that is brings.

Leveraging Open Source Tools

Our Airflow Orchestrator is currently being used to run more than 300 DAG processes that execute daily, enriching the decision making of our customer.

Our solution was one of the first major cloud implementations in Portugal, being used as a major case study throughout other governamental agencies as a successful case study.



What we've learned with this project:

- With such a challenging project, choosing the right battles to tackle first is key. We first developed some case studies around the data lake that showcased value so that users were confident of the massive improviments in their data collection capacbilities that our infrastructure would give them.

Company Growth Forecast

Summary:

Data Science can be applied to a lot of problems that may seem impossible to solve using statistics and machine learning.

One of our projects was trying to understand how to model a company's growth.

We wanted to target the Portuguese companies that are more likely to increase their exports in the next couple of years so we've used data about the company, sector and online presence combined with macroeconomic variables to understand which ones are more likely to do that.

F-Score vs. Complexity

Our target was a really rare one (< 2%). Although we wanted to improve the f-score of our model, we also needed to understand that a model that was too complex had the chance of being left on the sidelines.

Finding the Right Metric

After speaking with the business, we understood that precision @ k, was the best metric to optimize for. Currently, our customer had a work force that was able to deal with k companies at each time.

Every year, we want to make sure that the top k companies we recommend will likely improve their exports - and the ones that will not could potentially do it with a small nudge (our customers' job).

Roof Detection

Summary:

Computer vision techniques open up a ton of new possibilities that seemed farfetch a couple of decades ago.

One of our coolest project was focused on targeting roofs on satellite imaging for a utilities customer. We were not only targeting roofs but also looking at any obstruction that might prevent the setup of solar panels or other devices - we achieved an IOU of 0.7 using a U-NET architecture, producing extremely good results in less than 1 month.

The Models

Our model is based on a U-Net convolutional architecture. With just a couple of weeks we were able to produce a really decent baseline for roof detection.

During the project, we also developed two more models: obstruction detection and roof facing.

Few Data, Small Problems

While having few tagged data was an issue in the past, with most state-of-art computer vision models that is not the case.

Just with 60 tagged images, we were able to produce a really performant computer vision model with generalization power. Our POC ended up opening up a ton of new possibilities for our customer with few risk and investment.

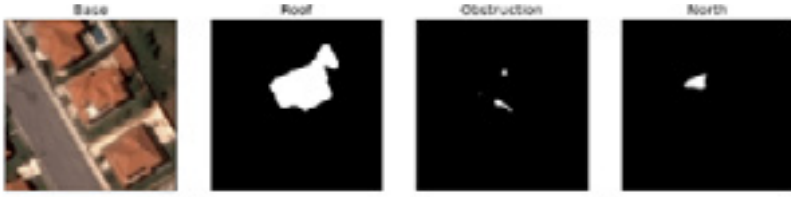
Business Problem ●●●●●

Modelling Complexity ●●●●●

Data Size ●●○○○

Scalability ●●●○○

Existing Literature ●●○○○



●●●●● More Complex / Challenging

Language Modelling

Summary:

We've used state of the art Natural Language Processing models (BERT, among others) to classify text based on precoded labels.

Our customer wanted to generalize a model that was trained on the English Language to Spanish - we trained a new model and set it up on an API that was able to guarantee precision and accuracy. During the process we tested several BERT-like models such as BertIn and roBERTa.

Iterating Fast

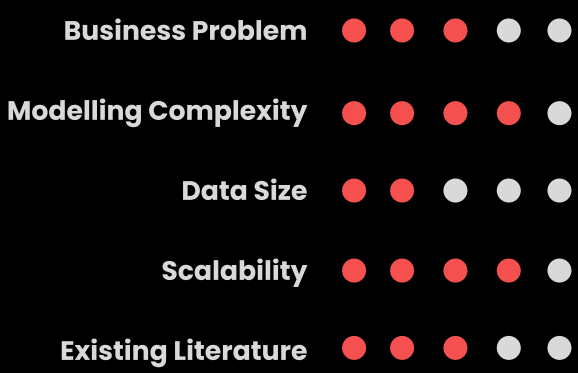
With current state-of-art NLP models, you can do a small budget NLP POC really fast with good performance.

We only took two weeks from gathering the data and training a good baseline on top of the spanish data.

Startup Budget

Most startups have a very tight budget to work with. If you're working with Data Science / Engineering projects, it is extremely important that you choose the right provider to work with you.

Only experienced providers can understand which paths you should be focusing on during development and what are the shortcuts to make sure you achieved a good model with less time.



Reporting Infrastructure

Summary:

We've built reporting infrastructures for start ups to help they make sense of their data and make crucial business decisions towards profitability and growth.

Our team managed the whole infrastructure, including:

- Provisiniong Resources
- Extracting data from data sources
- Manage KPI's and requirements

IoT Project

Our customer had lots of data coming from IoT systems that we needed to manage accordingly. This requires good coordination between ETL pipelines, infrastructure provisioning and reporting update.

Flexible Tech Stack

Within open source tools, we are mostly agnostic and work with different stacks.

We advocate for tools like Terraform or Ansible that even help you avoid cloud vendor lock in. Most of our solutions integrate with the customers tech stack, helping them manage the solutions at a lower cost.

Business Problem ● ● ○ ○ ○

Data Size ● ● ● ● ○

Scalability ● ● ● ● ○

● ● ● ● ●

More Complex / Challenging

DAREDATA ENGINEERING

Data Science & Engineering Advisory

Summary:

We've helped several startups implementing their data science & engineering infrastructure. Most startup companies have a hard time implementing proper data science & analytics initiatives due to lack of resources or strategy.

We've helped them kickstarting their data departments, guiding them on the conundrums of implementing scalable data infrastructures and helping them avoid the common pitfalls in the industry.

Time / Money Investment

Normally, startups are on a tight budget and have a really hard time finding a good financial plan to hire, train and maintain a team of data scientists and engineers.

Long-Term ROI

As data initiatives tend to have a long-term (and sometimes intagible) return-on-investment, startups and scale ups with a 1 to 3 year runway have a hard time allocating a huge portion of the budget to hire specialized data teams.

By hiring experts as advisors, companies can derisk themselves in the short term and prepare for growth in the long term.

Business Problem ● ● ● ● ●

Data Size ● ● ● ○ ○

● ● ● ● ●

More Complex / Challenging

WHY DAREDATA?

You are in great company

Our brand is trusted by some of the most prestigious companies and startups around the world. Join them and unlock the value of your data.

About DareData & How We Can Help

About DareData

DareData Engineering is an expert Data & AI technology consulting and talent agency company. We have contributed to data initiatives across diverse industries and sectors using the latest cloud and open-source technologies.

We are an asset in the upcoming challenges involving the digital transformation of organizations.

Get in touch to talk about your organization's plans for the technological future:

https://daredata.engineering

contact@daredata.engineering

(+351) 932 174 951