CIKLUM

Predicting Unpredictable: How to bring value to the client who doesn’t know what he wants?

Oleksandr Makarevych, Data Scientist at Ciklum

Agenda.

1. Business problem
2. Modelling and Technical Solution
3. Business Value

Few words about me.

B. Sc. in Computer Science
Kyiv, Ukraine
2016 - 2020

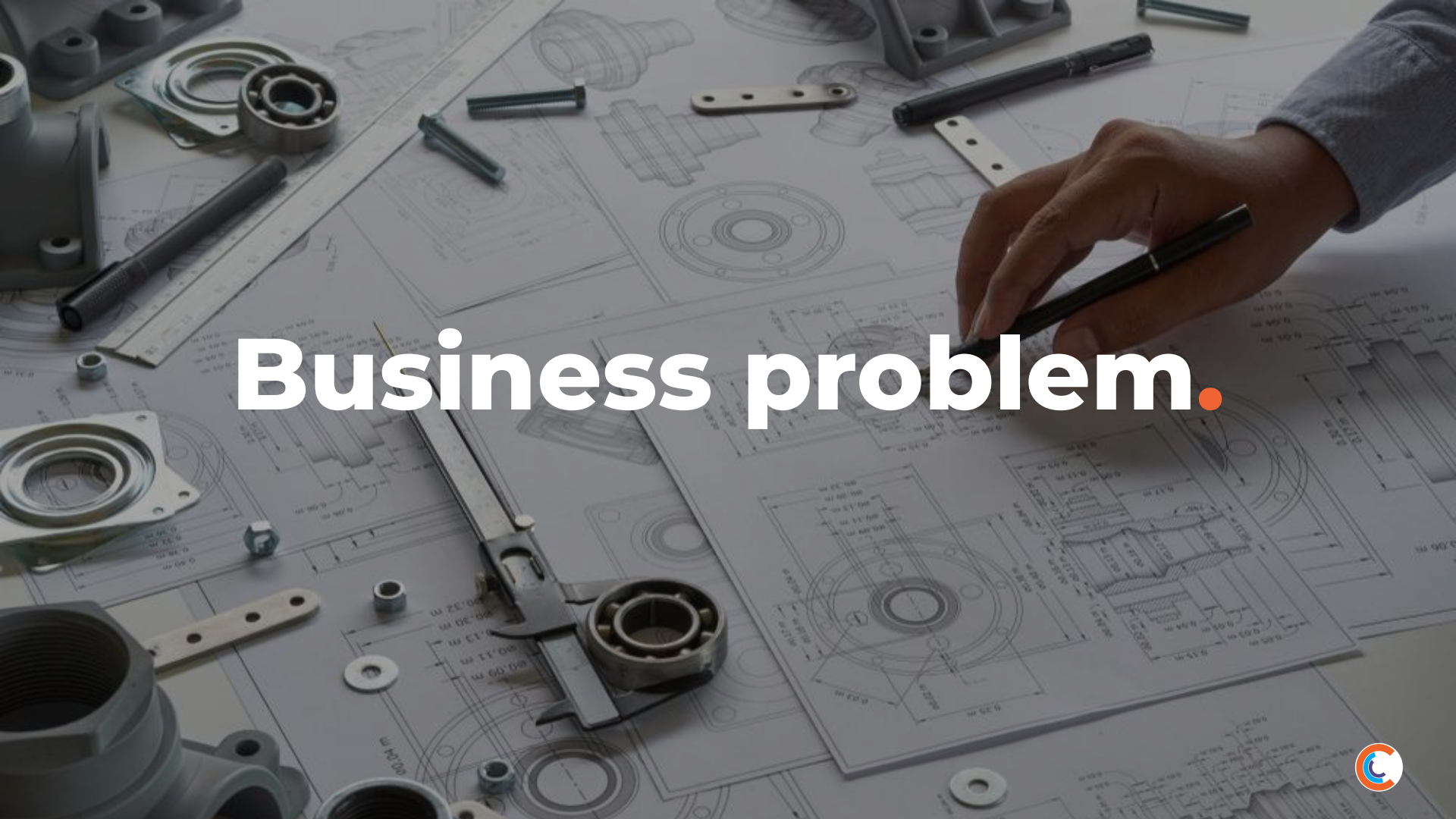
Elite M. Sc. In Data Science
Munich, Germany
2020 - Present

Intern
National Bank of Ukraine
2018

Data Engineer
Ciklum
2019

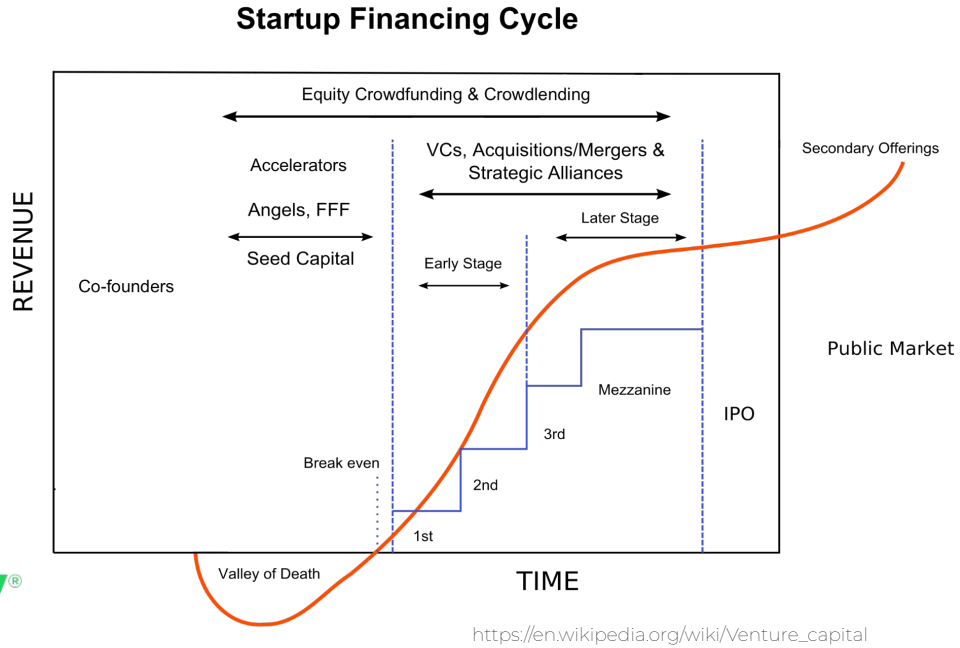
Data Scientist
SAP
2020

Data Scientist
Ciklum
2021


Business problem.

Venture Capital.

Venture capital is a form of private equity financing that is provided by venture capital firms or funds to **startups**, **early-stage**, and **emerging companies** that have been deemed to have **high growth potential** or which **have demonstrated high growth** (in terms of number of employees, annual revenue, scale of operations, etc).
Venture capital funds invest in these early-stage companies in exchange for equity, or an ownership stake. They take on the risk of financing risky start-ups in the hopes that some of the firms they support will become successful.



https://en.wikipedia.org/wiki/Venture_capital

First, the new firm seeks out "seed capital" and funding from **"angel investors"** and accelerators.
Then, if the firm can survive through the "valley of death"– the period where the firm is trying to develop on a "shoestring" budget – the firm can seek **venture capital** financing.

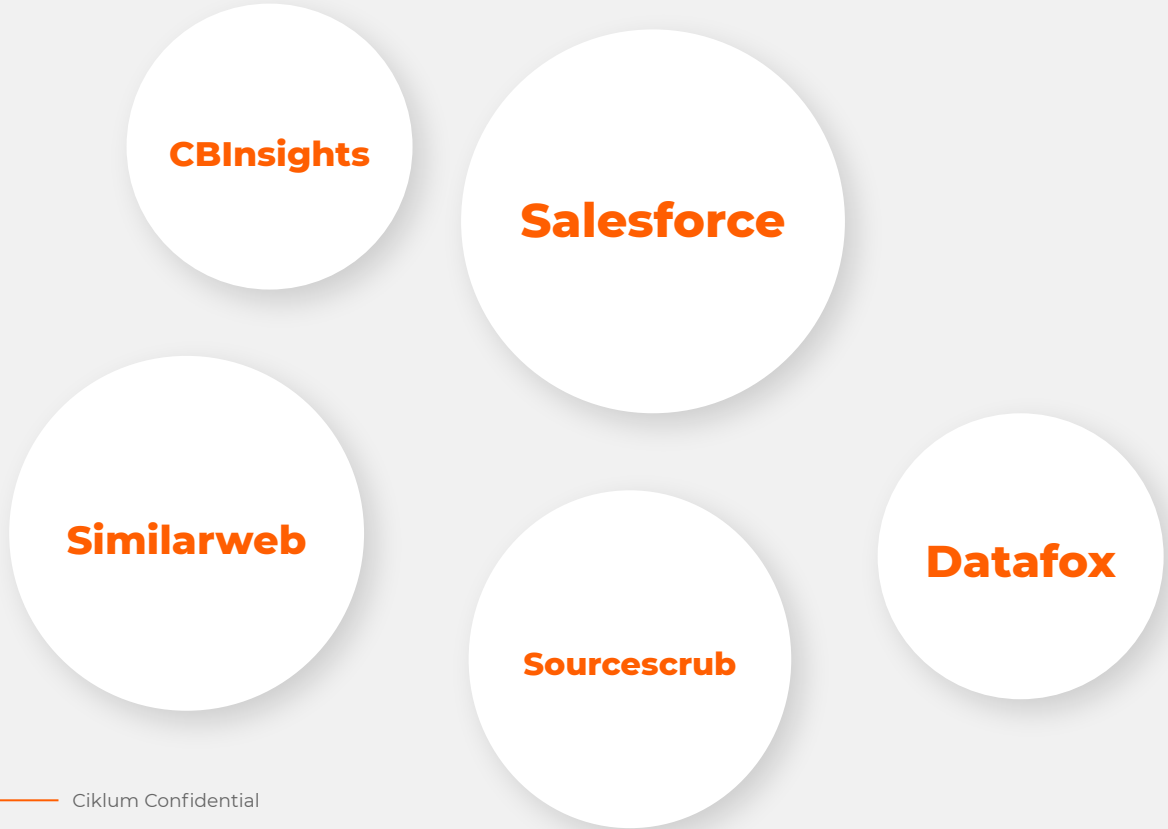
Situation.

The client is a large western venture capital investment fund, operating billions of dollars and investing into startups.

The client has large volume of company data in their data warehouse, ranging from financial, key business performance to scoring.

Financial analysts are no longer able to efficiently analyse such large quantity of company data to estimate the investment attractiveness of companies.

Data Sources in DWH.



Salesforce: primary source of company information including information on investment rounds, number of employees, etc

Similarweb: provides information on traffic through the websites like number of clicks, user growth, etc

Datafox: provides information on company appearance in news, conferences, key investors, etc

Sourcescrub: provides financial info describing company, including contact details

CbInsight: provides information on investments and social media appearance of the company

Challenges.

1

Datasets in the data warehouse are of varying quality and is not ready/suitable for analysis:

- Datasets are raw and unprepared for modeling
- Data comes from 8 uncorrelated sources that includes financial, third party insights, web data of various types etc
- Data is very sparse, noisy and sometimes have missing values (around 50%)

2

Existing analysis techniques performed by analyst are yet to be **translated/formulated as mathematical problems** that can be modelled and solved **algorithmically**

Actions.

1

Work on data exploration and data management to prepare dataset for modelling:

- Conduct missing value analysis and evaluate a range of techniques to impute missing values from ‘clean data’
- Feature engineer to selecting best features and focus on improving their data quality

2

Translate business problem into mathematical problem and choose target variables to represent data and features.

- Develop **Company Score Model**, which measures how attractive a particular startup is for investment. Higher scores correspond to higher investment attractiveness.
- Iteratively train, validate and deploy model to continuously deliver business value to client

Modelling and Technical Solution.

Data Management: Exploratory Analysis.

- EDA yielded only 40% of the data suitable for the analysis
- More than 300+ columns were drop due to lack of information
- Feature engineering was performed to improve the data quality
- Prepared Dataset contained no missing values

NY	Prospect	Applications	Jonathan Kerstein	False	4	Evaluating	10.0	Social	...	Outbound
OR	Prospect	Applications	Unassigned	False	2	Evaluating	10.0	Business Management	...	Outbound
FL	Prospect	Applications	Unassigned	False	2	Evaluating	10.0	Healthcare	...	Outbound
CA	Prospect	Applications	Unassigned	False	1	Evaluating	10.0	Business Management	...	Outbound
CA	Prospect	Computer Software	Michael Shephard	False	2	Evaluating	10.0	Business Management	...	Outbound

Missing Values: Pre-Processing.

To prepare a dataset for any type of modeling missing value in Data Warehouse had to be analyzed

MFV

- DWH contained 100+ categorical columns with missing value
- Missing values in categorical columns were imputed using most frequent value

Mean Mode Median

- DWH contained 150+ numerical values
- Depending on the distributions, missing values were imputed

Log

- Numerical features distributions were analyzed
- Depending on the distribution, log transform was applied for normalization

Scaling

- Numerical features showed a wide range value
- To normalize the dataset scaling were applied

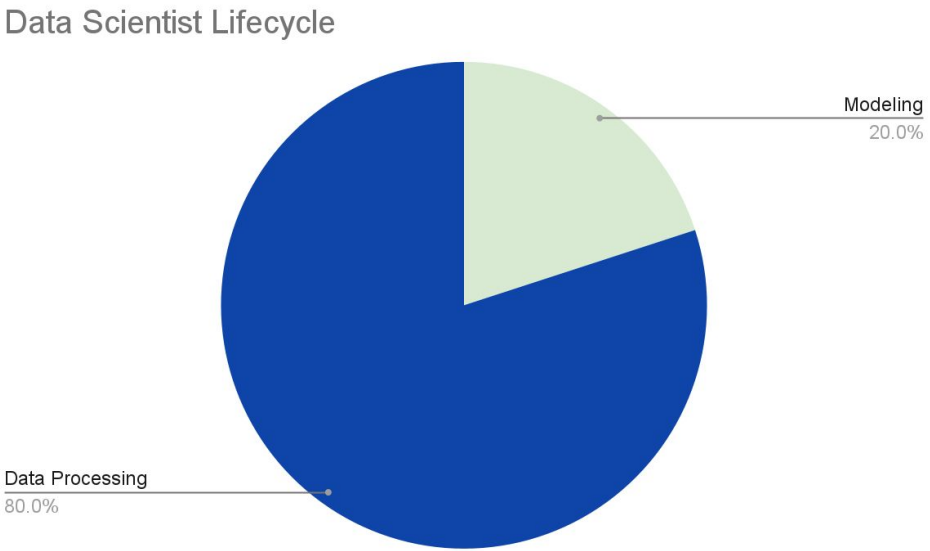
Date Split

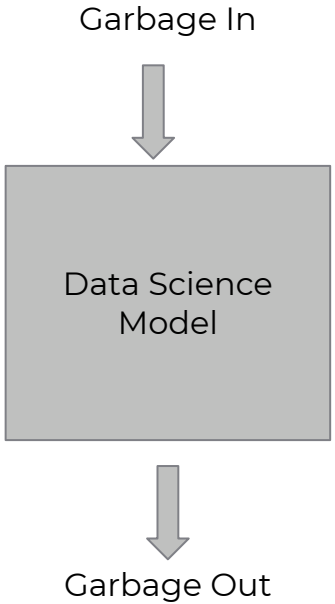
- DWH contained 30+ data columns
- Columns were analyzed and split into respective features if needed

Type transform

- Many of the columns contained the wrong data type
- Type conversion was performed for the data preparation and cleaning

Data Management: The Golden Rule.





Data Scientist work lifecycle almost always consists of spending 80% of the time on data processing and only 20% on actual data modeling. The key rule of Data Science says: “Garbage In = Garbage Out”. Our client was no exception, we put in a lot of efforts to be able to model the data with the dataset we prepared.

Technical Solution: Our Approach.

1

Define mathematical problem to translate business problem into technical solution

2

Choose an appropriate target variable to to represent the data and features included

3

Build a ML model to find a connection between the input data and the observed value

4

Validate the model to make sure that the Company Score is stable

5

Deploy to production to bring business value to the client

Technical Solution: How it is done.

Define mathematical problem	Classification can help us separate ‘good’ and ‘bad’ companies for investments. Company score is a probability of a company being a potentially good investment
Choose an target variable	The target choised as the variable in the Salesforce dataset which represent the last date of modification for the companies being analyzed by the analysts. We predict the probability that the company will be interesting for investors for 1 year
Build a ML model	Decision tree based algorithms are a particularly suitable choice for Company Score as it allowed to build decision rules based on a large number of features.
Validate the model	We performed K-Fold cross validation with 5 folds to ensure model stability. We use Recall and AUC to estimate model predictive performance
Deploy to production	Company Score model had to be integrated into the existing Data Warehouse. Amazon Web Services was a chosen cloud provider for the project

Model Comparison.

01

Logistic Regression

- **AUC Score:** 0.9799
- **Train Recall:** 0.8847
- **Test Recall:** 0.8358

02

Gradient Boosting Classifier

- **AUC Score:** 0.9842
- **Train Recall:** 0.9559
- **Test Recall:** 0.8507

03

LGBMClassifier

- **AUC Score:** 0.9859
- **Train Recall:** 0.9966
- **Test Recall:** 0.9104

* Comparison performed on 600 000 companies

Model selection and validation.

LGBM Classifier.
Library that utilizes decision trees and gradient boosting algorithm for classification problems. Decision trees use the data to split it according to the rules inferred from the training dataset. It then classifies new data point according to the inferred rules.

KFold.
Cross validation with 5 folds were used to estimate the mode performance. Cross validation is the method used to test predictions on multiple test sets and estimate the stability of the model.

Parameter Search.
Randomized search were performed to choose between 25+ sets of parameters. Randomized search is a hyper parameter optimization technique when the model is trained on all available set of parameters to find the combination that yields the best results.

100+ Features for training

95% Recall on a training set

91% Recall on a test set

98% AUC on a test set

Code Snippets.

```
objects_to_download = s3.ls(readPath)
for file in objects_to_download:
    if file.endswith(".parquet"):
        lst = file.split("/")
        filename = lst[-1]
        print("Downloading: ", filename)
        # path = "./tmp/" + filename
        s3.download(file, path + filename)
print("Downloaded files: ", len(objects_to_download))
companies_info_ex = pq.ParquetDataset(path).read_pandas()
companies_info_ex = companies_info_ex.to_pandas()
```

Data Input

```
yPred = clf.predict(X_main.values[trainIdx][selected])
results.loc[mdlIdx, 'train_recall'] = metrics.recall_score(Y.values[trainIdx][selected],
yPred = clf.predict(X_main.values[testIdx])
auc_score = roc_auc_score(y_test, clf.predict_proba(X_main.values[testIdx])[:, 1])
results.loc[mdlIdx, 'test_recall'] = metrics.recall_score(Y.values[testIdx], yPred)
```

Model Test

```
cv2 = model_selection.KFold(5, shuffle=True, random_state=0)
clf = model_selection.RandomizedSearchCV(clf2, searchParams, cv=cv2, n_iter=25,
                                         random_state=0, scoring=metrics.make_scorer(
        lambda x, y: FSpecial(x, y)[0]), verbose=2)
clf.fit(X_main.values[trainIdx][selected], Y.values[trainIdx][selected],
        sample_weight=sampleWeights.iloc[trainIdx].values[selected])
```

Model Train with KFold and parameter search

```
clf = lgb.LGBMClassifier(
    num_leaves=6, learning_rate=0.025,
    n_estimators=800,class_weight='balanced',
    random_state=0, n_jobs=n_jobs)
clf.fit(X_main.values[trainIdx][selected],
        Y.values[trainIdx][selected],
        sample_weight=sampleWeights.iloc[trainIdx].values[selected],
        categorical_feature=catFeatIdxs if catFeatIdxs else 'auto')
```

Model Train with tuned parameters

Model Feature Importances.

SimilarWeb Top 5

- Total Pages Visited Rank (2.83)
- Total Pages Visited Growth (1.52)
- Total Unbound Visits Rank (1.03)
- Desktop Visits Per User (1.01)
- Total Visits (0.92)

Datafox Top 5

- Revenue Estimate (2.0)
- Number of Employees (1.51)
- Total New Mentions (1.11)
- Finance Score (1.025)
- Days Since Last Funding (0.95)

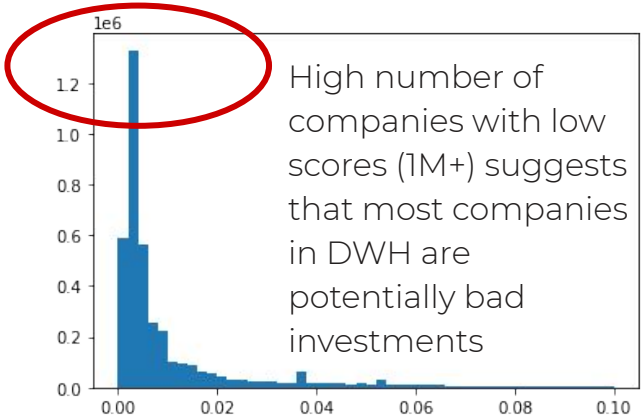
Cbinsight Top 5

- Sector (2.26)
- Industry (2.028)
- Mosaic Overall (1.61)
- Sub Industry (1.38)
- Overall Rank (0.91)

SourceScrub Top 5

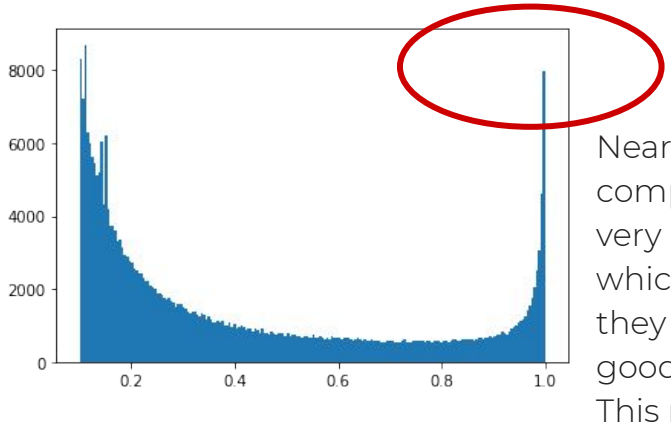
- Companies Source Count (3.055)
- Specialities (3.02)
- 6 month Employee Count Growth (2.92)
- Latest Employee Count (2.02)
- Specialities (1.66)

Scores distribution.



High number of companies with low scores (1M+) suggests that most companies in DWH are potentially bad investments

Scores distribution [0.0;0.1]



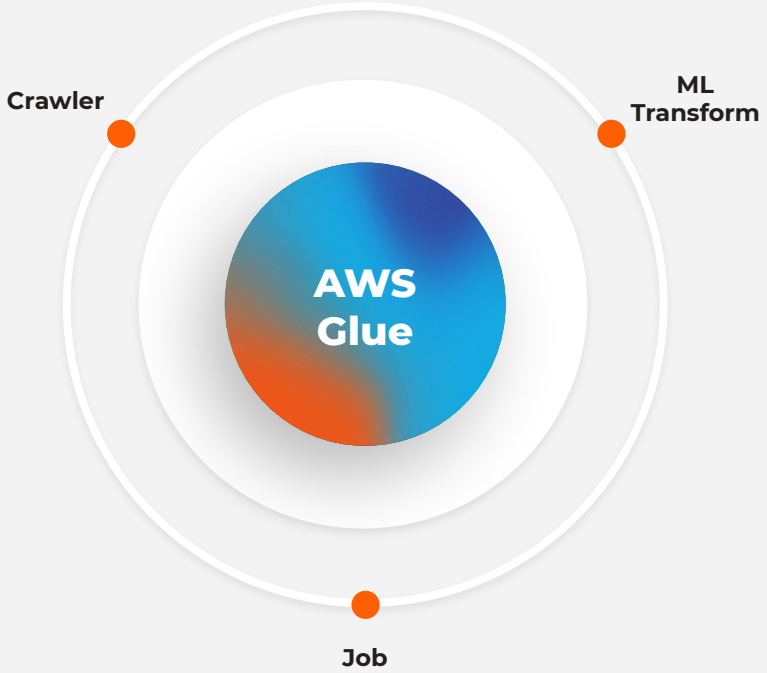
Nearly 8000 companies have a very high score which suggests they might be a good investment. This represents a 0.1% of all companies in DWH

Scores distribution [0.1;1.0]]

Further analysis suggest that most of the companies are indeed not worth investing in and only a handful of companies can be considered a worthy investment

Company Identity.

Q: HOW TO REMOVE DUPLICATES THAT COME FROM MULTIPLE SOURCES?
A: AWS FIND MATCHES



AWS Glue Crawler
Helps to crawl the data and infer the initial data schema

AWS Glue ML Transform
Used to create transformation and teach it to remove duplicates

AWS Glue Job
Used to apply taught ML transformation to remove data. There are two possibilities:

- Delete duplicates from dataset
- Append special ID to identify duplicates

Name	Contact Person	Title	Industry
ABC LLC.	Michael Doe	CEO	Crypto
ABC INC	John Applesed	CTO	Fin-tech
XYZ LIM	Maria Bohn	CFO	AGRO

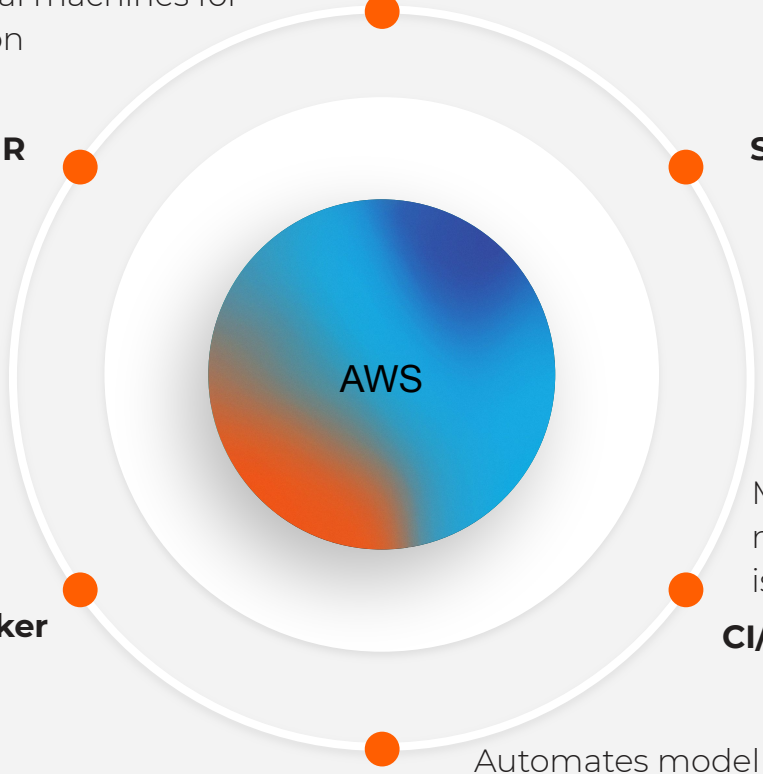
Technical Solution: Deployment.

EC2

Central AWS offering which provides virtual machines for code execution

EMR

Elastic MapReduce is a service used to run computation cluster of the pipeline and speed up the calculations.



S3

Service which acts as a file storage for compressed model and interim files

Makes sure that production state is not disrupted and each new feature is tested before deployment.

CI/CD

Docker

Encapsulates model into its own image so that the deployment is faster and more stable

Airflow

Automates model execution inside existing pipeline.

Results - Business Value.

Recommendation on prospective investments

By utilizing Company Score model, our client can receive reliable suggestions on which companies to invest in.

28% increase in ROI from recommendation on prospective investments

Easy to understand predictions

By utilizing standard classification we receive a score between 0 and 1 which serves as a measure of certainty.

Improved search speed for analysts

As of now, analysts don't need to go through the whole DWH, but rather focus on companies with highest scores.

35% reduction in time spent on investment sourcing by analysts

Dedicated Company Score model

The model has successfully been integrated into the existing pipeline and give daily predictions

Q&A

Let’s keep in touch!

O.Makarevych@campus.lmu.de
https://www.linkedin.com/in/oleksandrmakarevych/