STEEILY
A cloud-based solution for implementing scalable fault classification in production

Ekaterina Rossi (150286)
Konrad Schulte (149872)
Marin Mes (149899)

Copenhagen Business School
MSc Business Administration & Data Science
Applied Machine Learning and Data Engineering in Business Context

Agenda

1. Introduction
2. Our Understanding of the Situation
3. Problem Statement
4. Our Approach
    a. Four phased model
    b. End-to-end architecture
5. Project Requirements, Benefits and Risks
6. Recommendations
7. Appendix
    a. Data Maturity Assessment
    b. Visualization and Reporting
    c. Classification model: Methodology
    d. Classification model: Model description
    e. Income Statement & Cost breakdown

Steeily is a traditional steel plate manufacturer from England and the global market leader with a current growth rate of 14.2%[1].

30 Countries
40,000 Employees
$ 25,000,000[2] Revenue
12.3%[3] Average Growth[4]
53% Global Market Share

General Information

- Steeily is a manufacturing company founded in 1850 in Sheffield, England.
- Steeily is specialised in producing plates from A300 and A400 steel.
- A key factor in Steeily's success is the continuous optimization of its production processes.

[1,2,3] Appendix, S.23
[4] over past four years

Steeily has a manual approach for detecting, differentiating, and reporting faults, and no consistent data management across its different entities.

Productional Level

- Steel plates can have various defects in production process, classified into 7 different types: Dirtiness, Stains, Pastry, Bumps, Z-Scratch, K-Scratch, and Others.
- Right now, these faults are detected manually at the end of the production process where employees take plates out, evaluate their severeness and report the fault.
- Depending on the severeness Steeily differentiates between two major fault groups:

Value Decreasing Faults
(Dirtiness, Stains, Pastry)

- Steel plates can still be sold but for a lower price

Dysfunctional Faults
(Bumps, Z-Scratch, K-Scratch)

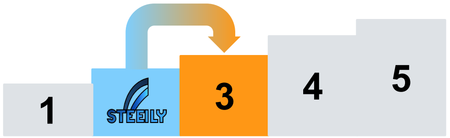
- Steel plates have to be recycled separately
- This recycling process needs a lot of energy

Steeily aims to automate and optimise this process

Data Management

- Currently the steel plate fault data is residing in on-premises data sources SAP and Oracle.
- Different Steeily countries engage in different data preprocessing and data management steps.
- There is no company-wide standard for collecting other data.
- Currently, there is no source to target mapping.
- In general, Steeily is not extracting value from the data that they generate.

- Therefore, according to our data maturity assessment[1], Steeily can be identified in Level 2.



Steeily aims to reach level 3 in two years

[1] Appendix, S.17

We identified two major problems which we propose to solve with a Machine Learning model and an integrated cloud solution.

Problems

With energy costs accounting for a quarter of all cost of revenue[1], current energy price increases will further reduce profits.

Steel plate fault data resides in silos and is not handled consistently across the company. Therefore, different units cannot benefit or learn from each other's data.

Approaches

We propose to implement a Machine Learning classification model to decrease energy consumption per item and labor time (and therefore total energy costs by 10%[2]).

We provide a scalable cloud solution to centralise and harmonise the company-wide steel plate fault data for all entities.

[1,2] Appendix, S.25

“

How can we leverage a Machine Learning classification model in combination with a cloud solution to increase Steeily’s net income by 15%?

Our approach is a four phased strategy which focuses on the migration of data storages and data processing tools to the cloud, and on applying a classification model to fault steel plates data

Orientation Phase

- ➔ Understand the current data structure and architecture.
- ➔ Source to target mapping with SME's[1].

Build Phase

- ➔ Set-up Azure DevOps services for project management.
- ➔ Set-up a Data Lakehouse in Azure.
- ➔ Develop a fault classification model.

Implementation Phase

- ➔ Migrate steel plate fault data to the Data Lakehouse.
- ➔ Migrate the data processing tools to Databricks.
- ➔ Implement solution in pilot country.

Scaling Phase

- ➔ Streamline data flows and output flows.
- ➔ After concept has been proofed, scale up to all 30 countries.

STEEILY

[1] Subject Matter Experts

First, we will investigate the current data structures and work together with SME's to perform source-to-target mapping in excel.

Data Sources

Things to consider:

- What are the different data sources?
- What is stored in each data source?

Data Warehouse

Things to consider:

- How are the different data sources connected?
- What does the raw data look like?
- How are the summary and metadata stored?

Reporting & Analysis

Things to consider:

- Which transformations are done?
- What data processing tools are used?
- Which data is used for reports / visualization?

Source-to-target Mapping

➔ Source-to-target mapping explains where data comes from. Without a source-to-target sheet no one in the company will know how to retrieve the steel plate fault data.

➔ Source-to-target mapping guides employees on how to work with data.

➔ Source-to-target mapping also serves as detailed documentation.

➔ Source-to-target mapping is done in collaboration with SME's and it will serve as a standard set of rules for all countries.

Next, we will build a new cloud infrastructure and build the classification models on real data.

Azure DevOps Services: Set of integrated services to manage projects, specifically version control using Azure Repos

Data Lakehouse: In the cloud the steel plate fault data will be stored in a Data Lakehouse, which must be set-up with the correct configurations.

Databricks: The environment running on top of the Lakehouse where the ML models will be running, must be installed.

Classification Model: The machine learning model will be trained and optimized on real data (the current POC is based on representative subset data).

Then, the migration strategy and the classification model will be implemented for the selected pilot country.

Pilot Country

Steel plate faults data

Data processing tools

New Data Structure

Storage: Azure Data Lakehouse

Data Processing: Databricks

Classification Model Set-Up

1. A picture will be taken of each steel plate that is picked out manually and that is considered to be faulty.
2. The picture goes into the classification model.
3. The picture is pre-processed such that it fits the input requirements of the model.
4. The pre-processed data is classified into a fault class.
5. The steel plate is directed in a corresponding direction to reheat or to the scrap pile.

Lastly, the data flows towards and from the Data Lakehouse will be streamlined, and the solution will be scaled to all factories.

Scaling

Pilot Study

Proposed solution is being tested in the pilot country

Scaled Solution

Start scaling the proposed solution to more countries, until all factories in all countries have migrated to the cloud and have incorporated the classification model as part of the production process.

The classification model will be continuously adapted to any changes or additions in the type of faults of steel plates, making it a scalable and dynamic model.

Since the Data Lakehouse is massively scalable there are many opportunities to migrate all other Steeily data to the cloud and scale up in that way.

With more and more data in the cloud, scaling up in the number of Azure data services used will allow for deeper insights and analyses of Steeily.

The following end-to-end architecture will achieve the aforementioned solutions which in turn allow for scaling and optimization.

	1. Data Sources	2. Ingestion	3. Storage	4. Model & Transformation	5. Deployment	6. Insights
	Where is the data now? Currently, the steel plate fault data is residing in on-premises data sources SAP and Oracle.	**Where will the data be ingested in?** Azure Data Factory: Data will be ingested in the Azure Data Factory (ADF) which will be the software responsible for loading the data, scheduling the loading of the data, and doing any primary transformations to the data before it is being stored in Azure.	**Where will the data be stored?** Azure Data Lakehouse: This is where the different data loads from the ADF will be stored in a secure way. Confluence: Confluence will be used to secure consistent documentation throughout the project.	**Where will the data be processed?** Databricks: The platform which will be used on top of the Data Lakehouse to do preprocessing. It is also the platform where the classification model is developed. ML Flow: ML flow will be used to manage the Machine Learning lifecycle of the classification model.	**Where will the model run?** The model will run on image data that comes in directly from the factories into Azure. The model will run on Databricks and the results will be sent back to the machine, instructing it what to do.	**Where will insights be collected?** Confluence: Just like the documentation of the project, the insights of both the data migration and the ML classification models will be documented in Confluence. Reporting and visualization is out of scope, but initial ideas are proposed in Appendix (S.18).
Azure Services		Azure Data Factory	Data Lakehouse	Databricks + ML Flow + Azure DevOps	Databricks + ML Flow + Azure DevOps	
DevOps	Azure DevOps (Project Planning) + Confluence (Documentation)			Version Control (CI)	CD	Monitor
Environment	Test Environment				Production Environment	

The implementation of the migration of data storage and data processing tools to the cloud will take 12 months.

- Steeily
- Our team
- Both

	Quarter 1			Quarter 2			Quarter 3			Quarter 4		
	1	2	3	4	5	6	7	8	9	10	11	12+
Admin tasks	Project startup for one factory	Provision of documentation about current processes									Scaling up for all factories	
IT tasks	Setup Azure AD[1]									Maintenance of solutions		
Infrastructure Development		Migration data to Lakehouse										
			Migration data processing tools to Databricks		Implementation of CI/CD pipelines							
Infrastructure handover						Documentation	Training			Handover		
ML Implementation				Data Processing	Modelling		Integration / Training					
ML Deployment								Deployment		Hypercare		
ML Handover								Documentation	Training	Handover		

[1] Active Directory

If all requirements are satisfied and risks are resolved, this project can result in an increase of $408,000,000[1] in Steeily's net income.

ML Model related Requirements

- Installation of a new Camera System
- Databricks Subscription
- Confluence Subscription
- Azure DevOps Services Subscription
- Maintenance
- Steel Plate Faults Data

Costs: $68,000,000[2]

Cloud Solution related Requirements

- IT Department Support
- Contact to SMEs
- Backup of Data
- Azure Data Lake Subscription
- Azure Data Factory Subscription
- Hevo Subscription

Costs: $88,000,000[3]

Project Benefits

- Reduction of electricity consumption and costs by 10% ($690,000,000)[4]
- Automatic documentation provides further analysis and improvement of production processes
- Synergy Effects: future projects will also benefit from the setup cloud environment
- Steeily reaching level 3 in data maturity assessment

Net Income Increase

+$408,000,000[5]

Project Risks

- No data documentation
- Unreliable SMEs
- Lack of clear cloud migration strategy
- Locked in with Microsoft products/services

[1,4,5] Appendix, S.26
[2,3] Appendix, S.25

Improved data labeling and implemented Faulty Steel Plates Classifier Model will help to reduce variable production costs by $534,000,000[1].

Project related recommendations

- Investigate the causes of k scratches and bumps faults and ways to reduce their occurrence to reduce production costs.
- Sort out “other” category into defined categories to improve accuracy of the model by 5%.
- Implement the Faulty Steel Plates Classifier Model in the production to reduce energy costs by $534,000,000.[2]

Big scale recommendations

- Add the Faulty Steel Plates Detection Model into pipeline as part of CI / CD practices.
- Add reporting into pipeline to monitor changes in models or visualise insights using Power BI tools.
- Test the accuracy of the Faulty Steel Plates Classifier model using raw image data as an input.

[1,2] Appendix, S.25

Appendices

In this Data Maturity Framework, Steeily can today be identified as an Opportunistic Archetype.

STEEILY

LEVEL 1 - Apprentice / ad-hoc

Characteristics:
- Little or no governance
- Limited tool set / slowly adoption
- Driven by individuals
- No roles defined
- Controls applied inconsistently
- Data quality issues not addressed
- Reactive in nature
- Data a mean to solve operational issues and loose to no methods

LEVEL 2 - Opportunistic

Characteristics:
- No to slow emerge of governance
- Consistent tool set / low-tech
- Driven by individuals in silos
- Some roles and process defined
- Growing awareness of impact of data quality issues
- Business begins to understands the value of data but reactive
- Description modelling but repeatable
- Driven by teams or single BA's
- Performing below peer/market

LEVEL 3 - Defined

Characteristics:
- Data viewed as business driver and tech as a key enabler
- Digitizing to pursue operational excellence
- Process outcomes, including data quality is more predictable
- Business understands the value of data start being proactive asset
- Driven by experts or single BA's
- Data is categorized and defined but no clear ownership
- Loose to no methods and standards present and not in a coherent company framework
- Effort vs outcome not balanced
- Performing at peer/market level

LEVEL 4 - Developed

Characteristics:
- Central planning and governance
- Managing risk related to data
- Data management is embedded in process and function
- Data quality metrics in place
- Data used proactively and a crucial asset for decision making
- Driven experts / CoE's
- Detailed methods and standards in place along the data life cycle.
- Data is categorized, defined and owned
- High effort and resources invested in building capabilities
- Measurable and targeted business outcomes with clear value creation
- Exceeding peers/market performance

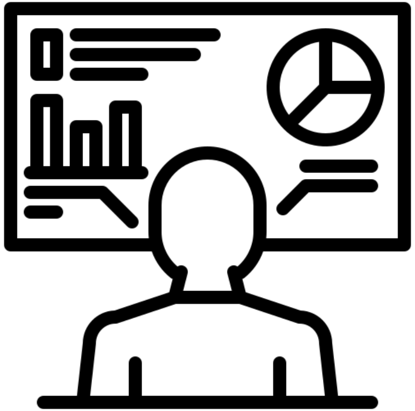
LEVEL 5 - Institutionalized

Characteristics:
- Highly predictable processes
- Able to reduce data risk
- Well understood metrics to manage data and process quality
- Data as an crucial assets
- Data has become a competitive edge for the company
- Truly data driven - it's culture and defines how they operate
- Frontrunners - setting new standards
- Act proactively as an organization
- Operational efficiency in how to build and deploy data use cases, models and systems
- Technology stack is constantly evolving
- Deliver best in class outcomes that drives strategic direction setting

Visualization and reporting: suggestions and recommendations

- Suggested platform: PowerBI.
- PowerBI creates dynamic reports displaying trends and patterns in the data.
- Regarding fault data, think of:
  - Most common faults
  - Fault development over time
  - Location on the plates where most faults are detected
  - Time patterns of the faults in the steel plates
- Analysing these trends may lead to insights further improving the production process and increasing the bottom line.

Power BI



Methodology

Preprocessing

Data visualisation
Outlier detection
Correlation analysis
Normalisation
Data split

- The outliers were detected visually using a boxplot. The Z-score, a measure of the relative spread of values (Illowsky & Dean, 2013), was calculated for each feature. Higher z score, more standard deviations there are from the mean (Ibid.). Observations with a z-score above 4 will be removed from the dataset.
- Correlation analysis, strength of linear relationship between features (Kim, 2019), is conducted to discover possible multicollinearity.
- Data was scaled to a range between 0 and 1.
- The dataset was splitted into training and test dataset, 70% and 30%, respectively.

Feature engineering

PCA | RFE

- Principal Component Analysis (PCA) is a data extraction method that project datapoints into lower dimension hyperplane to preserve as much variance as possible (Géron, 2019). N components, that explains more than 99% of the variance, are passed to the classifier.
- Recursive Feature Elimination (RFE) is a data selection method, which removes features with the smallest weights recursively (sklearn.feature_selection.RFE, n.d.). Step is 5.

Modeling

Random Forest | Gradient Boosting

- Two ensemble techniques were utilised, Random Forest and Gradient Boosting. Random forest is a collection of decision trees trained in parallel (Géron, 2019) and decision is made by voting, while Gradient Boosting trained sequentially, where each classifier learns using errors of the previous (Ibid.).
- These algorithms were chosen since it was proven to handle multicollinearity (Belgiu, & Drăguţ, 2016; Chen, Benesty & He, 2018).

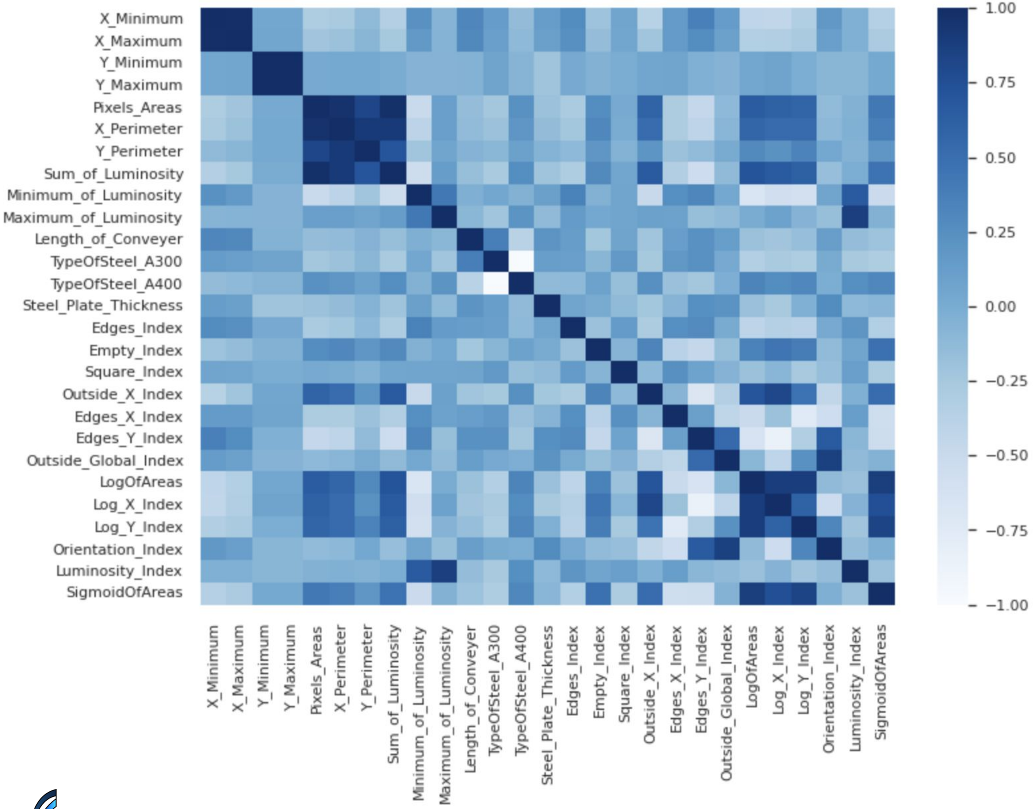
Evaluation

F1-score
Precision
Recall

- F1-score - harmonic mean of precision and recall (Ibid.), is used to train models.
- F1 score, precision and recall is used to evaluate models

Multicollinearity problem was detected using correlation analysis.

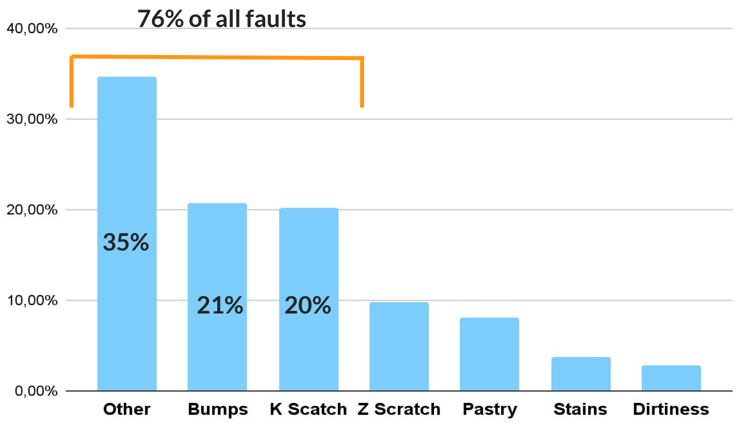
Correlation analysis of features



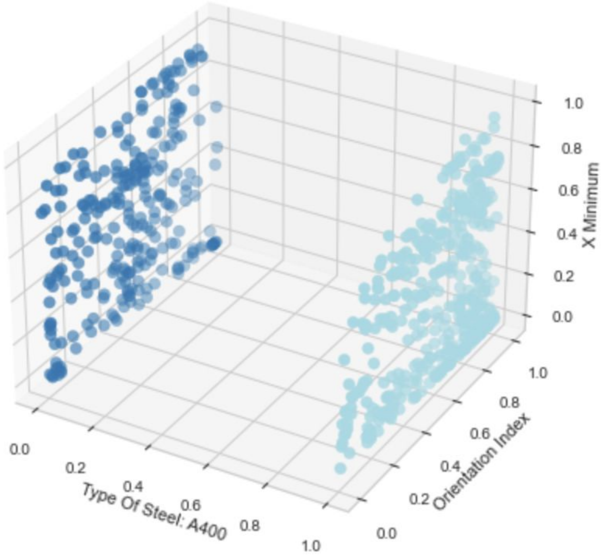
- The heatmap of correlation matrix shows a strong correlation between some features.
- Variance Inflation Factor (VIF) was calculated for each feature. VIF higher than 10 demonstrates significant multicollinearity problem (Illowsky & Dean, 2013). Current dataset contained 18 features that had VIF score higher than 10.
- Perfect multicollinearity was revealed between the A300 and A400 steel types, X maximum and minimum, and Y maximum and minimum. Since Gradient boosting cannot handle perfect collinearity (Chen, Benesty & He, 2018), features A400 steel types, X maximum, Y maximum were dropped.

Analysis of faulty steel plates reveals occasional misuse of 'other' category

Analysis of faulty steel plates



Clustering of the category "others"



- 76% of all defects account for 3 types of defects: k scratches, bumps, and others. At the same time, steel plates with k scratches and bumps must be recycled before sale.
- The "other" category is 35% of all defects

- Using K means clustering, 2 clearly separated clusters were identified. This may indicate the possibility to identify new categories of defects and divide the category others into these categories.

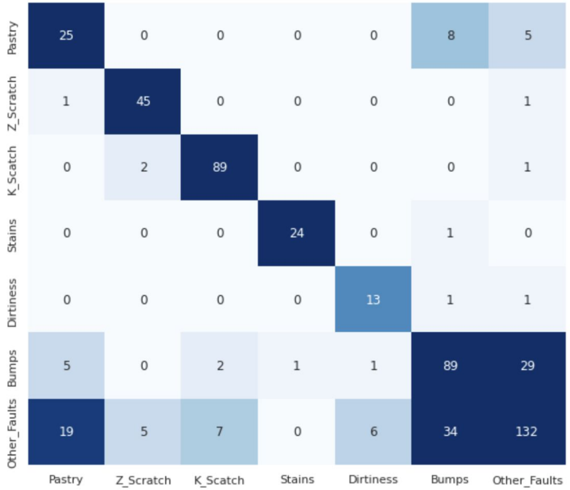
Accuracy of the best faulty steel plates classifier model is 78%.

Result of built models

	Random Forest		Gradient Boosting	
	PCA	RFE	PCA	RFE
F1-score	0.71	0.78	0.58	0.71

- On this dataset, models using feature selection techniques showed better results than models using dimensionality reduction technique.
- On this dataset, random forest had higher accuracy than gradient boosting.
- The accuracy of the Random Forest model that used RFE is the highest, 78%.

Confusion matrix of Random Forest with RFE



- The most ambiguous class is the "others", since it was the most misclassified, while other classes were mostly mislabeled as an "other" class.
- Precision of 5 classes out of 7 is higher than recall.

Income Statement of Steeily (I)

Items	TTM	2021	2020	2019	2018
Total Revenue	28,378,724,817	25,083,036,330	21,964,129,886	19,698,771,198	17,494,468,204
Cost of Revenue	23,952,604,531	21,170,931,870	18,538,469,238	16,626,429,809	14,765,923,453
Gross Profit	4,426,120,285	3,912,104,460	3,425,660,647	3,072,341,387	2,728,544,748
Operating Expense	962,551,052	850,767,720	744,980,490	668,143,937	593,378,274
Operating Income	3,463,569,233	3,061,336,740	2,680,680,157	2,404,197,450	2,135,166,474
Net Non Operating Interest Income Expense	59,396,574	52,498,710	45,970,849	41,229,460	36,615,861
Other Income Expense	101,597,724	89,798,940	78,633,047	70,522,912	62,631,360
Pretax Income	3,505,770,382	3,098,636,970	2,713,342,355	2,433,490,901	2,161,181,972
Tax Provision	826,552,396	730,562,910	639,722,338	573,742,007	509,539,970
Net Income	2,679,217,986	2,368,074,060	2,073,620,017	1,859,748,894	1,651,642,002
Other under Preferred Stock Dividend	2,490,305	2,201,100	1,927,408	1,728,617	1,535,183
Average Dilution Earnings	23,583	20,845	18,253	16,370	14,538
Diluted NI Available to Com Stockholders	2,464,167,951	2,177,998,295	1,907,178,892	1,710,474,342	1,519,071,351
Basic EPS	8,150	7,204	6,308	5,657	5,023
Diluted EPS	7,967	7,042	6,166	5,530	4,911
Basic Average Shares	113,047	99,919	87,494	78,469	69,688

STEEILY

Income Statement of Steeily (II)

Items	TTM	2021	2020	2019	2018
Diluted Average Shares	114,529	101,228	88,641	79,498	70,602
Total Operating Income as Reported	3,276,339,739	2,895,850,650	2,535,771,147	2,274,234,212	2,019,746,191
Rent Expense Supplemental	10,708,684	9,465,060	8,288,143	7,433,312	6,601,520
Total Expenses	24,915,155,583	22,021,699,590	19,283,449,728	17,294,573,747	15,359,301,729
Net Income from Continuing & Discontinued Operation	2,466,681,840	2,180,220,240	1,909,124,553	1,712,219,330	1,520,621,074
Normalized Income	2,599,529,582	2,297,640,059	2,011,944,009	1,804,434,088	1,602,516,952
Interest Income	109,111,575	96,440,190	84,448,502	75,738,566	67,263,380
Interest Expense	164,213,027	145,142,580	127,095,078	113,986,617	101,231,453
Net Interest Income	59,396,574	52,498,710	45,970,849	41,229,460	36,615,861
EBIT	3,669,983,410	3,243,779,550	2,840,437,434	2,547,477,519	2,262,413,427
Reconciled Cost of Revenue	22,702,532,644	20,066,033,790	17,570,957,784	15,758,706,532	13,995,298,873
Reconciled Depreciation	1,336,418,266	1,181,216,850	1,034,340,499	927,659,640	823,854,031
Net Income from Continuing Operation Net Minority Interest	2,466,681,840	2,180,220,240	1,909,124,553	1,712,219,330	1,520,621,074
Total Unusual Items Excluding Goodwill	173,838,971	153,650,640	134,545,218	120,668,356	107,165,502
Total Unusual Items	173,838,971	153,650,640	134,545,218	120,668,356	107,165,502
Normalized EBITDA	5,180,240,648	4,578,647,040	4,009,323,152	3,595,805,517	3,193,432,963

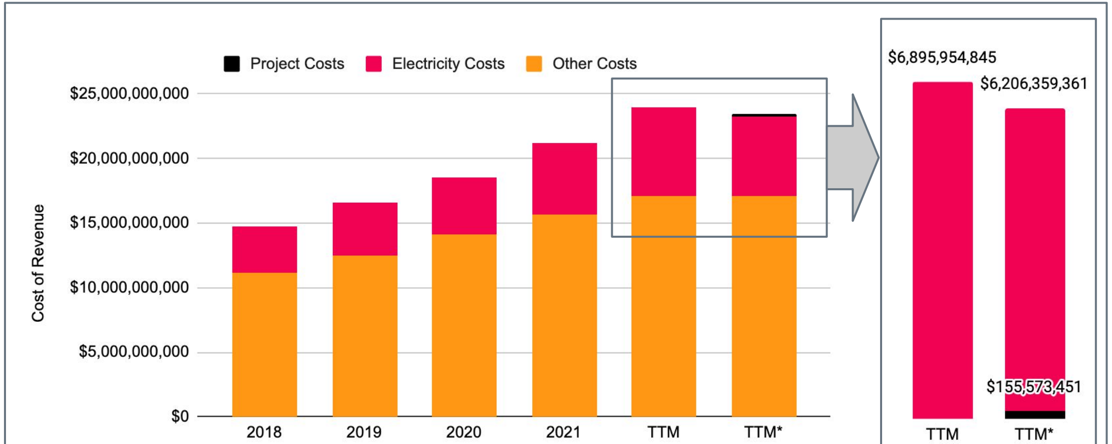
STEEILY

The cost breakdown reveals: With this project, Steeily can reduce its Cost of Revenue by $534,022,034

- We calculated that Steeily can globally **reduce electricity costs by 10%** after successfully implementing the proposed ML model in an integrated Cloud environment
- The project-related costs are as follows:

Model Implementation	
Installation of new Camera System	$11,869,951
Databricks Subscription	$14,000,000
Azure DevOps Services Subscription	$14,000,000
Confluence Subscription	$254,000
Cloud Implementation	
Azure Data Lake Subscription	$28,000,000
Azure Data Factory Subscription	$32,000,000
Hevo Subscription	$449,500
Other Costs	
Maintenance	$45,000,000
Consulting	$10,000,000
Total	**$155,573,451**

Cost of Revenue (with and without implementation)



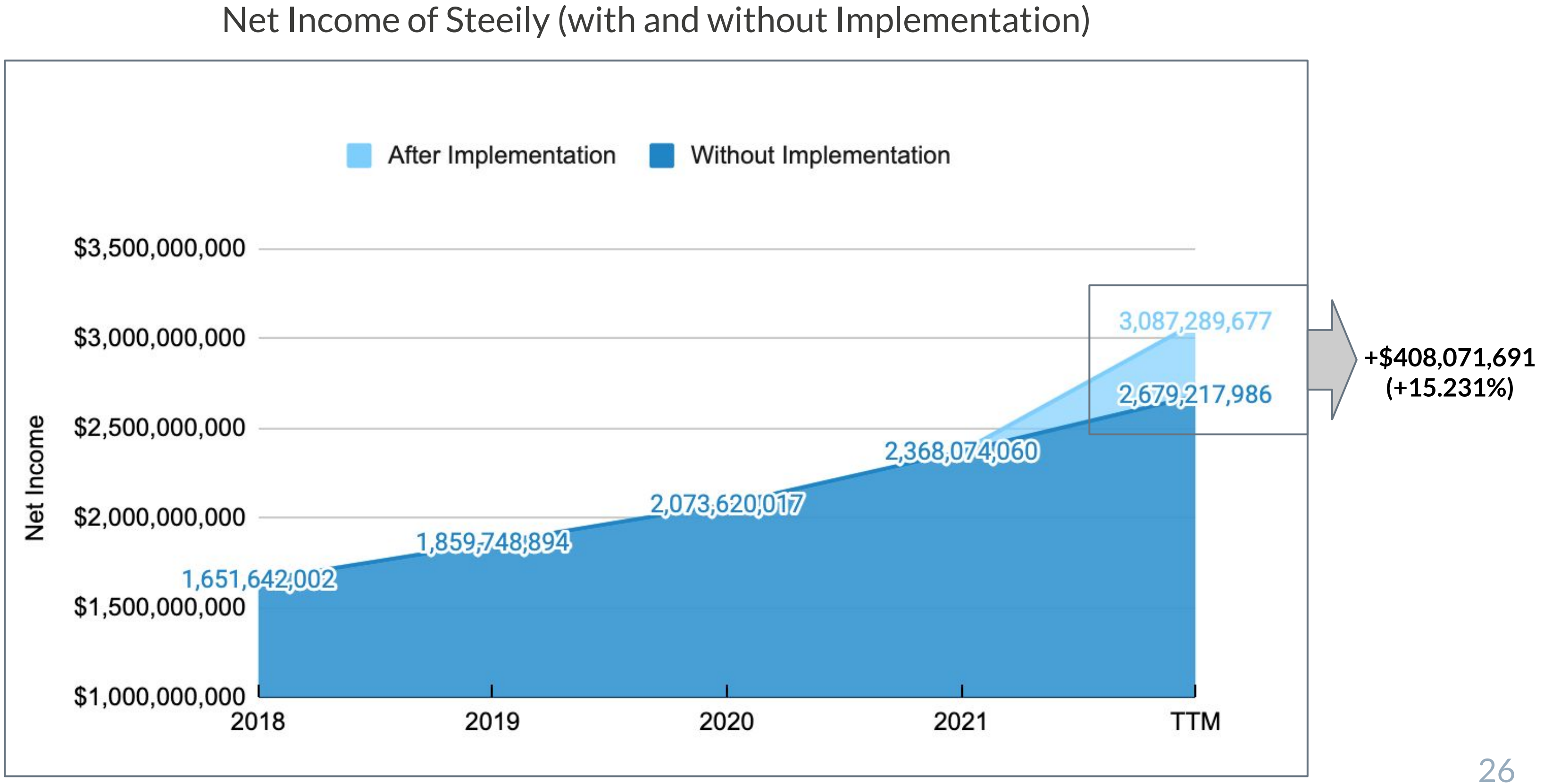
*with Implementation

With taking into account both electricity cost reduction and the project related cost, **Steeily would benefit from a total cost reduction of $534,022,034**. Since some of the costs are one-time payments (and possible synergy effects) the cost reduction in the following years might be even higher.

Electricity Cost Reduction	$689,595,485
Project Costs	-$155,573,451
Total Cost Reduction	**$534,022,034**

STEEILY

With Implementation of our proposed solution Steeily can increase its Net Income by 15%



References

References

- ▷ Belgiu, M., & Drăguţ, L. (2016). Random forest in remote sensing: A review of applications and future directions. ISPRS Journal of Photogrammetry and Remote Sensing, 114, 24–31. https://doi.org/10.1016/j.isprsjprs.2016.01.011
- ▷ Chen T., Benesty M., He T. (2018). Understand your dataset with XGBoost. https://cran.r-project.org/web/packages/xgboost/vignettes/discoverYourData.html
- ▷ Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems (3rd ed.). O'Reilly Media
- ▷ Illowsky, B., & Dean, S. L. (2013). Introductory Statistics. Amsterdam University Press.
- ▷ Kim, J. H. (2019). Multicollinearity and misleading statistical results. Korean Journal of Anesthesiology, 72(6), 558–569. https://doi.org/10.4097/kja.19087
- ▷ Sklearn.feature_selection.RFE. (n.d.). Scikit-learn. https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.RFE.html