



university of
groningen

campus fryslân



Internship Report

Minor

BSc Data Science and Society

Academic Year 2025 - 2026

Semester 1A & 1B / Year 3

S5673364

Britt Sleeuwenhoek

From Raw Data to Insights: A Data Science Internship at DataGrow

“Building Data Pipelines, Predictive Models, and Dashboards”

Student

S5673364

Britt Sleuwenhoek

Rijksuniversiteit Groningen

Campus Fryslan

Internship Organisation

DataGrow / Koningslaan 60, 3583 GN Utrecht The Netherlands

Internal Supervisor

Taís Fernanda Blauth / t.f.blauth@rug.nl

External Supervisor

Maarten San Giorgi / maarten@datagrow.nl

15-12-2025

Preface

During my bachelor's degree in Data Science and Society, I acquired foundational knowledge in various machine learning techniques, programming in Python, as well as contemporary ethical and legal considerations such as human rights in the digital age and GDPR compliance. Much of this education, particularly the technical training, was primarily theoretical in nature. The datasets we worked with were often pre-cleaned or carefully chosen to align with the specific topic of each course module. While this approach effectively conveyed core principles, it did not fully represent the complexities and realities of working with data in a professional environment.

Recognizing this gap, I wanted to ensure I also got hands-on experience. In particular, I was seeking this experience in the end-to-end processes of a real-world data organisation. My goal was to better understand how data pipelines function, from raw data to deliverables. To achieve this, I reached out to various companies to secure an internship opportunity.

Following multiple discussions with potential organisations, I secured an internship at DataGrow. They provided me with a project encompassing the complete data pipeline, allowing me to participate in every stage of data processing, integration, and analysis. This experience has given me practical insight into how theoretical skills are applied in practice and the challenges that arise in managing real-world data.

I would like to express my gratitude to Maarten San Giorgi at DataGrow for his guidance, encouragement and support throughout the internship period. I also thank my academic supervisor, Dr. Taís Fernanda Blauth, for her advice and assistance, and my colleagues at DataGrow for creating a welcoming environment that enriched my learning experience.

Britt Sleuwenhoek

Utrecht, 2025

Table of Contents

Preface	2
Table of Contents	3
1. Introduction	4
2. Description of Internship	5
2.1 Internship Organisation	5
2.2 Internship Assignments	6
2.2.1 Pipedrive Analysis	6
2.2.1 “2.0” Analysis	6
2.3 Internship Tasks	7
2.4 Internship Results and Output	10
2.4.1 Pipedrive Sales Analytics Dashboard	10
2.4.2 “2.0” Employee Engagement Dashboard	10
3. Evaluation	11
3.1 Learning Outcomes	11
3.2 Contributions to the Internship Organisation	11
3.3 Value to and of the Programme	12
3.4 Reflection and Future Development	13
4. References	14
5. Appendices	15
Appendix A. Internship Plan	15
Appendix B. Internship Logbook	16
Appendix C. Learning Outcome Table	19
Appendix D. Workflows	21
Pipedrive Sales Analytics Dashboard	21
Appendix E. Product	21
Pipedrive Sales Analytics Dashboard	22
2.0 Employee Engagement Dashboard	27

1. Introduction

The field of data science is rapidly evolving, with increasing reliance on data-driven decision-making across industries. As organisations accumulate vast amounts of data, effective data processing and analysis become critical to generating actionable insights and achieving competitive advantage. In this context, gaining practical experience in handling real-world data is essential to complement the theoretical knowledge acquired during academic training.

This internship report documents my internship at DataGrow, a company specializing in data solutions and business intelligence, which was carried out from September until December 2025. The purpose of this internship was to provide comprehensive exposure to the end-to-end data pipeline process, from data collection and cleaning to analysis and reporting. Through this project, I aimed to develop practical skills and better understand the operational challenges and best practices within a professional data environment.

The report first describes the internship organisation and context, followed by the internship assignment, tasks, results, and outputs. It then presents an evaluation of learning outcomes and contributions, and concludes with references and appendices containing the internship plan, logbook, learning outcome table, delivered products and evaluation forms.

2. Description of Internship

2.1 Internship Organisation

I completed my internship at DataGrow, a data and business intelligence specialist based in Utrecht. DataGrow helps small to medium size enterprises and government organisations with growth ambitions and an innovative mindset. Their services include business intelligence, data visualisation and analytics, data-driven strategy development, AI and machine learning (ML) applications, cloud solutions, and data integration through platforms such as Azure.

DataGrow's business model consists of two main components. The first part is consultancy services providing specialized expertise and advisory services to assist clients. The consultancy work carried out by the organisation encompasses companies such as: Daiwa House, Tennen and PLTRM. This includes activities such as conducting market research, optimising processes, and offering guidance. The second part of their business comes from secondment, whereby qualified professionals are temporarily assigned to client organisations to support specific projects, fill short-term skill gaps or strengthen existing teams. Some clients out of their portfolio are: Beweging 3.0, KPN, Nationale Nederlanden and SMT.

A notable characteristic of DataGrow is its transparency and approachable culture as a relatively small organisation. The company promotes knowledge sharing, maintains an open and collaborative atmosphere, and leverages its size to tailor solutions closely to client needs. This combination of openness, innovation, and supportiveness creates a constructive environment for professional learning.

2.2 Internship Assignments

The assignment was designed collaboratively with my internship supervisor, allowing flexibility to explore areas of interest while aligning with the company's needs. While the scope was defined by focusing on a specific dataset related to deal performance, meaning how many deals are won or lost and how they move through the sales pipelines, I was given freedom to choose which data science methods and techniques to apply. The learning outcomes indicated a clear objective to work on the full data pipeline, from data acquisition to dashboard development. This semi-structured approach allowed me to tailor my work to address real business needs while experimenting with different analytical approaches. This approach created a meaningful balance between guidance and independent initiative.

2.2.1 Pipedrive Analysis

Pipedrive is the software that DataGrow uses to keep track of their secondment data. In this programme you can easily define in what stage of the hiring process people are, what organisation to address, and how long the deal has been on hold (Pipedrive Inc / Pipedrive, 2025b). Pipedrive is a CRM platform widely used by sales and recruitment teams to manage contacts, deals and sales activities. It also provides tools to visualize, track, and organize pipeline stages and deal progress. My assignment builds on these capabilities by enhancing insights through customized visualisations and predictive analytics. The project aimed to demonstrate the potential of ML on existing data and therefore create value for the business developer.

2.2.1 "2.0" Analysis

In addition to the primary data pipeline project, I also worked on a secondary initiative that kept the internship varied and engaging. This involved developing a system to monitor and evaluate internal company social activities also known in the company as 2.0. This analysis captures participation, feedback, and outcomes through Google Forms, and integrates the data into Databricks. This project was particularly engaging and provided the company with valuable insights into activities already conducted, overall employee satisfaction, and activities that could be reintroduced. By working on this project, I was able to apply the data processing and analytical skills learned during my internship in a different, more dynamic context.

2.3 Internship Tasks

At the start of my internship, I was introduced to the company, the team, and the various online work environments used, including tools like Databricks, a cloud-based platform for data engineering and analytics, and Pipedrive (*Databricks, 2023*). This onboarding phase was crucial for familiarizing myself with the data infrastructures and organisational processes. Early in the internship, I also gained access to the shared Google Drive, laying the groundwork for my data exploration and analysis. These first weeks were focused on understanding the available datasets, clarifying project objectives, and developing a detailed plan to guide my work.

To achieve the objectives of the internship and address the defined research question, I engaged in a variety of tasks spanning the full data science lifecycle, combining technical execution with business-focused problem solving. A significant part of my responsibilities involved the extraction, transformation, and structuring of deal data from Pipedrive. This required developing a robust and scalable data pipeline capable of supporting near real-time analytics, which included integrating APIs, cleaning and transforming raw data, and organising it in a format suitable for both reporting and ML applications (Pipedrive Inc / Pipedrive, 2025a). The workflow of this pipeline is visualised in Appendix D.

While working on the Pipedrive API, I used exported datasets from Pipedrive to initiate the ML models. These data were first cleaned and prepared focusing specifically on the Detachering Kandidaten pipeline (secondment data). Key preprocessing steps included renaming columns for clarity, exploding the "Grower" (the person attached to the vacancy) column to ensure accurate counting of individual deals, and converting date columns to the proper date format.

I engineered several features to help the model learn patterns in the data, including temporal attributes (such as days since deal creation and expected deal duration), win rates per individual grower and organisation, and a progression score representing the stage of the deal in the pipeline (Guyon & Elisseeff, 2003). To improve model performance, I added interaction features that combined these core metrics, capturing how deal progression was influenced by organisation or grower-specific win rates, as well as temporal effects.

The initial RandomForestClassifier results produced relatively low maximum win probabilities, indicating room for improvement (Breiman, 2001). Through iterative feature additions and testing multiple models, I encountered issues of overfitting, evidenced by high

performance scores on training data but limited generalization. I attempted techniques like feature selection and stratified cross-validation to mitigate overfitting, but these yielded limited success. An additional challenge was the imbalanced nature of the dataset (8 won vs. 284 lost deals), with far more lost deals than won (He & Garcia, 2009). This complicated model training. This was mitigated by applying an oversampling method called SMOTENC (Omari et al., 2025). This method helped to lessen the imbalance and yielded better results.

The Classification model development concluded with an AUC score of approximately 0.97 after implementing SMOTENC oversampling, feature selection based on importance, and model selection across multiple algorithms. While the high AUC indicated strong discriminatory power, the extreme class imbalance limited maximum win probability predictions. Future enhancements include collecting additional data and further investigating feature selection methods.

Following discussions with a colleague about the Classification model, a colleague suggested exploring anomaly detection as an alternative approach, given the highly imbalanced nature of won vs. lost deals. This led me to develop an Anomaly Detection model using an Isolation Forest, treating won deals as rare anomalies within closed deal datasets (Liu et al., 2012). Trained on one-hot encoded categorical features like pipeline names, grower names and stage names with 5% contamination, the model achieved up to 0.98 accuracy after restricting training to historical closed deals only.

All the work on the data pipeline and model came together in a PowerBI dashboard built using the cleaned, structured data from the Data Warehouse. The dashboard spans nine pages, each providing focused insights to support management decisions.

- **The Deal Conversion page** presents key metrics such as deal counts, conversion rates between pipeline stages, average deal duration, and overall win ratios.
- **The Value Conversion page** offers similar metrics with an additional bar chart of realized deal value.
- **The Grower Deals section** visualizes the number of vacancies per grower and their progression through pipeline stages, with filters for grower, month, and stage.
- **The Organisation and Heritage page** highlights responsive organisations, their origins by platform, and trends over time, while the **Organisation Details and Organisation**

Conversion pages provide detailed and comparative conversion metrics per organisation and stage.

- **The Reason for Loss page** summarises causes for lost deals with an accompanying timeline chart to reveal recurring patterns.
- Finally, the dashboard integrates outputs from the **Anomaly Detection and Classification models** for open deals, offering predictive signals on potentially successful or atypical cases.

In addition to the main assignment, I contributed to a secondary internal project as stated before. For this I developed a system using Google Forms, which is connected to Google Sheets, to capture attendance, participation feedback, and outcomes of events. The overall dataflow for this process is also included in the workflow overview in Appendix D. I chose Google Forms because it is easy to set up, maintain and expand. However, this approach also introduces issues such as typos and people describing the same activity in different ways. To reduce these problems and create more uniform data, I implemented two separate forms: one for participants and one for organizers. Both have predefined fields that are all necessary to fill in before the form can be submitted. This does not fully eliminate human error, but it significantly reduces the impact on the dataset.

The matching of these two sheets was done by code in Databricks, which was then loaded into PowerBI so the information could be displayed in a dashboard. The matching logic was also designed with imperfections in mind. Both forms require a date field, and matches are made on this column, but with a ten-day matching window. This means that even when participants enter a date that is slightly different from the organizers' date, the records can still be linked correctly, improving robustness against small input mistakes.

Automation of the data pipeline was supported by scheduled nightly jobs on Databricks, maintaining up-to-date data with minimal manual intervention. Job setup and workflow management benefitted from close collaboration with my external supervisor, leveraging his expertise. Documentation was also part of my responsibilities, allowing the system to be maintained and extended in the future.

Overall, these tasks allowed me to engage in end-to-end project development, combining technical expertise in data engineering, ML, and dashboard design with business skills in

requirements gathering, project management, and stakeholder communication. The experience not only enabled me to deliver concrete results for the company but also fostered significant professional and academic growth.

2.4 Internship Results and Output

The internship produced two main deliverables that expanded DataGrow's analytical capabilities: a comprehensive Pipedrive Sales Analytics Dashboard with integrated ML models and the 2.0 Employee Engagement Dashboard. Both dashboards were designed in the company's house style, ensuring a consistent look and feel with existing reports. The requirements for these dashboards came directly from the intended users through discussion and feedback sessions, increasing the likelihood that they will be adopted in practice.

2.4.1 Pipedrive Sales Analytics Dashboard

The primary output is a 9-page interactive PowerBI dashboard built from the cleaned data in Databricks, providing unprecedented visibility into sales performance. Key pages include:

- **Deal Conversion:** Pipeline stage counts, conversion rates, average deal duration, and win ratios.
- **Value Conversion:** Revenue tracking with realized value bar charts alongside conversion metrics.
- **Grower Deals:** Vacancy involvement and stage progression per grower, filterable by name, month, and stage.
- **Organisation & Heritage:** Responsiveness by organisation, platform origins, and engagement trends.
- **Organisation & Heritage Details:** Conversion rates of two organisations for easy comparisons.
- **Organisation Conversion:** Conversion rates per phase and per organisation.
- **Reason for Loss:** Timeline analysis of loss reasons to identify recurring failure patterns.
- **ML Model Outputs:** Near real-time predictions from both the Classification model and Anomaly Detection model, flagging high-potential and outlier open deals.

This dashboard overcomes Pipedrive's native reporting limitations by enabling custom metrics, advanced filtering, and predictive insights previously unavailable.

2.4.2 “2.0” Employee Engagement Dashboard

The secondary project delivered a PowerBI dashboard visualizing team activity data collected via automated Google Forms and matched in Databricks. Key features include:

- **KPIs:** Attendance rates, organizers, activity grades, and winners.
- **Yearly overview:** Best activities, top participants, most/least present employees.
- **Interactive filtering:** By activity type and year.
- **Timelines:** Activity ratings over time and cumulative win rankings.

This dashboard provides management with insights into employee happiness and engagement trends for inclusion in monthly reporting alongside financial metrics.

3. Evaluation

3.1 Learning Outcomes

During my internship, I have achieved the learning outcomes outlined in the Data Science and Society (DSS) syllabus. A more detailed overview is provided in Table 1 in Appendix E. I worked effectively within the organisation by remaining adaptable, cooperative, and clear in my communication. This was evident through regular meetings with my internship supervisor to discuss progress and any problems that arose. It was also clear through constructive feedback that was consistently integrated into iterative improvements of models and dashboards. I successfully made quality deliverables, including two interactive dashboards and two predictive models, which align with the expectations described in my internship plan.

In addition to the DSS outcomes, I also met the three personal learning outcomes defined in my internship plan: developing proficiency in end-to-end data handling, enhancing analytical and machine learning capabilities, and improving data storytelling and communication skills. End-to-end data handling was practiced through building and automating the Pipedrive data pipeline in Databricks. My analytical and ML skills were strengthened by developing the Classification and Anomaly Detection models, including dealing with imbalanced data and model evaluation. Data storytelling and communication were developed through designing and presenting both dashboards in close collaboration with their intended users.

My internship report comprehensively documents these contributions and my learning journey. In addition, I critically reflected on my competencies and learning, recognizing strengths such as practical data engineering and analytical skills, and identifying areas of improvement especially in handling data imbalances and deploying machine learning models in production. This experience expanded my competencies in ML techniques, feature engineering, and unsupervised learning within a real business context. The reflective process further strengthened my analytical and communication skills, bridging theory and practice from the DSS curriculum.

3.2 Contributions to the Internship Organisation

I contributed to DataGrow by delivering practical solutions that enhanced business insights and reporting capabilities. The Classification model and the Anomaly Detection model offer predictive analytics previously unavailable, aiding decision-making and signaling meaningful patterns. Importantly, by extracting data from Pipedrive and developing custom analytics and dashboards, I enabled far more comprehensive visibility into deal performance. Previously, the organisation was limited to the basic reports and visuals provided directly in Pipedrive; now, advanced analysis, tailored metrics, and deeper reporting are possible, supporting more informed strategy and planning.

The 2.0 Dashboard brings a new perspective on employee engagement, supporting management's understanding of social activity participation and overall team sentiment. This initiative was valuable: activity data were systematically collected and organized for the first time, providing new insights into participation trends and employee satisfaction.

My work helped automate aspects of data processing and visualisation, increasing efficiency and consistency in regular reporting, and building a foundation for ongoing improvement in both areas. Because all pipelines and models are implemented within a robust Databricks production environment and scheduled as recurring jobs, the workflows now run continuously without manual intervention. This means the organisation can keep using the dashboards and model outputs with minimal extra effort, which was an important success factor for this project.

3.3 Value to and of the Programme

The internship reinforced the DSS programme's pillars by providing a practical testing ground for the specific learning outcomes I defined in my internship plan. In particular, I:

- **Extracted and accessed real-world data** by building the Pipedrive API integration and loading structured deal data into Databricks.
- **Applied data cleaning and transformation techniques** such as exploding categorical columns, handling date formats, and implementing the medallion architecture (bronze, silver, gold).

- **Implemented ML models** through the Classification model and the Anomaly Detection model.
- **Interpreted analytical results and translated them into actionable insights** for sales performance and deal management.
- **Developed visual presentations of data insights** by creating both dashboards in PowerBI.
- **Strengthened end-to-end data workflow skills** by orchestrating automated pipelines from API ingestion to dashboard deployment.

Whereas coursework often focuses on individual techniques, the internship required integrating these competencies simultaneously under real business constraints: dealing with imbalanced datasets, iterative stakeholder feedback, and production deployment considerations. This holistic application revealed nuances absent in academic settings, such as the impact of data quality on model performance and the importance of aligning technical solutions with organisational priorities. Core skills from courses like Machine Learning and Visualisation were directly leveraged but significantly enriched through this end-to-end project experience. The practical context also illuminated areas for further growth, particularly in advanced unsupervised learning techniques, and real-world model deployment strategies, gaps that coursework alone cannot fully address.

3.4 Reflection and Future Development

This internship clarified a career direction towards data science roles that emphasise ML and data visualisation. The hands-on experience with anomaly detection and predictive modelling expanded my technical toolkit and boosted my confidence in tackling complex, ambiguous problems. In doing so, it strongly contributed to my personal learning outcomes of enhancing analytical and ML capabilities and improving data storytelling and communication skills. The design and implementation of the dashboards, together with regular discussions with stakeholders, also supported the DSS learning lines related to effective communication with stakeholders. I aim to deepen knowledge in model interpretability, deployment, and ethical AI, and to pursue a master's degree focused on Data Science and AI. The supportive team

environment was pivotal to my learning, underscoring the importance of collaboration and feedback in professional growth.

In addition to technical growth, I experienced clear personal development at DataGrow. Over time, I became more confident in collaborating with colleagues, gathering and refining requirements from end users, communicating progress and setbacks, and sharing planning and priorities in a structured way. Regular update meetings, feedback sessions on interim dashboard versions, and troubleshooting of pipeline issues helped me move from mainly executing tasks to actively engaging in the process.

Looking ahead, I plan to pursue targeted improvements for each deliverable. For the Classification model, future development should focus on collecting additional data and implementing feature importance analyses. The Anomaly Detection model could be enhanced by testing additional algorithms (e.g., Local Outlier Factor, One-Class SVM), adding numerical features like deal value and duration. The 2.0 Dashboard would benefit from expanding the Google Forms to also accept photo uploads.

I now have a clearer appreciation for the importance of model interpretability to build trust with stakeholders. Additionally, deployment considerations such as automating pipelines, maintaining scalable workflows, and monitoring models in production have emerged as critical competencies I aim to develop further. A concrete step in this direction during the internship was implementing all pipelines and models in a robust Databricks production environment with scheduled jobs, so the system runs continuously without manual work and can be used by the organisation on an ongoing basis.


Equally important, the supportive and collaborative team environment was foundational to my learning. The experience reinforced the idea that professional growth is maximized in environments that encourage collaboration, openness, and continuous learning. These principles will also guide my future career choices.

4. References

- Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32.
<https://doi.org/10.1023/A:1010933404324>
- Databricks: Leading Data and AI Solutions for Enterprises*. (2023, October 13). Databricks.
<https://www.databricks.com/>
- Guyon, I., & Elisseeff, A. (2003). An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research*, 3, 1157–1182.
- He, H., & Garcia, E. A. (2009). Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9), 1263–1284.
<https://doi.org/10.1109/TKDE.2008.239>
- Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2012). Isolation-Based Anomaly Detection. *ACM Trans. Knowl. Discov. Data*, 6(1), 3:1-3:39. <https://doi.org/10.1145/2133360.2133363>
- Omari, K., Taoussi, C., & Oukhatar, A. (2025). Comparative Analysis of Undersampling, Oversampling, and SMOTE Techniques for Addressing Class Imbalance in Phishing Website Detection. *International Journal of Advanced Computer Science and Applications*, 16(2). <https://doi.org/10.14569/IJACSA.2025.0160276>
- Pipedrive Inc / Pipedrive. (2025a). *Pipedrive API Reference and Documentation*.
<https://developers.pipedrive.com/docs/api/v1>
- Pipedrive Inc / Pipedrive. (2025b). *Sales Pipeline Management Software*. Pipedrive.
<https://www.pipedrive.com/en/features/pipeline-management>

5. Appendices

Appendix A. Internship Plan

**university of
 groningen**

Form for Approval of Internship
- To be filled in digitally by the student in accordance with supervisor* of the internship.
- Digital copy (pdf) with signatures of the student and supervisor to be sent to the Exam Board (cf examboard@rug.nl) asap, but preferably no later than April 1st 2024.
- Note that the remaining 15 ECTS of the minor spots need to be approved as well, approval should be requested through this [link](#).

Student name	Britt Steuwerhoek
Student number	55673364
Name of internship	Stage by DataGrow
Amount of ECTS	15 ECTS
CF supervisor	Tais Fernanda Blauth
Internship organisation (and location)	DataGrow located in Utrecht (Koningstaan 60)
Supervisor at internship company	Maarten San Giorgi
Supervisor contact details	maarten@datagrow.nl

Justification

List the main topics of the internship.	For my internship at DataGrow I will go over the following topics: <ul style="list-style-type: none">- Unlocking data from an application- Extracting data to make it accessible for analysis- Preparing data in a Python environment- Cleaning and transforming the data using Python and accompanying libraries to ensure it is structured properly and ready for analysis- Analyzing/ predicting based on this data<ul style="list-style-type: none">- Applying analytical and/or machine learning techniques to uncover patterns and/or make predictions- Presenting the results (e.g. dashboard or report)- Communicating insights through visualizations or reports
Mention the learning outcomes that will be achieved after successful completion of the internship.	Based on the above mentioned topics I plan on achieving the following learning outcomes: <ul style="list-style-type: none">- Demonstrate the ability to extract and access data from real-world applications- Apply data cleaning and transformation techniques- Implement statistical and machine learning models- Interpret analytical results and translate findings into actionable insights- Develop visual or written presentations of data insights- Strengthen practical skills in end-to-end data workflows

**university of
 groningen**

Specify why the internship adds to your DSS programme.	I am pursuing this internship to gain hands-on experience in the full data lifecycle (extracting, preparing, analyzing, and presenting data) in a real-world setting. This opportunity allows me to apply the academic material of the DSS programme to practical tasks, deepening my understanding of how data can be used to draw insights and decision-making. So far, much of the data we have worked with in the DSS programme has come from platforms like Kaggle, where datasets are often pre-cleaned and simplified. While useful for learning concepts, they don't fully reflect the complexity and messiness of real-world data. This internship gives me the opportunity to work with unstructured or less curated data, developing practical skills in data cleaning, transformation, and integration. These skills are essential but harder to practice in a classroom setting, it helps bridge the gap between academic exercises and the real-world challenges faced by data professionals. In addition, the experience of turning raw data into meaningful insights and communicating those insights through dashboards or reports will strengthen my ability to translate complex findings into clear, actionable outcomes, something that is crucial for roles outside of academia. Ultimately, this internship adds a layer of depth to my DSS education by developing my practical, technical and communication skills in a way that classroom learning alone cannot fully achieve.
Specify the ECTS and workload by giving an estimated time schedule of the internship where you describe the frequency and planned period of meetings with your CF supervisor and your internship supervisor. (Remember: 1 ECTS equals 25 hours of workload.)	The workload of this internship will be 560 hours, which will equal to 20 ECTS. This will be based on a 40 hour workweek, for 14 weeks. The internship will start the first of September. Meetings with the internship supervisor will take place on a regular basis, preferably every week. This will be to ensure I stay on track and can get feedback. Meetings with the CF supervisor will be held if either the supervisor or I request one. Halfway through the internship there will also be an interim evaluation, where both the CF supervisor and internship supervisor will be present. Preferably I will also be present for this evaluation, to be able to get feedback and also share my thoughts.
Describe the method of assessment and the assessment criteria.**	For the method of assessment and the assessment criteria I refer to the assessment forms appended to the 24-25 Internship Manual DSS.

**university of
 groningen**

Signature of the student, who by signing additionally confirms to be aware of any further mandatory administrative steps to take after approval is received, as indicated on Brightspace

Name: Britt Steuwerhoek
Signature: 
Date: 03-07-2025

Approval of the CF internship supervisor

Name: Tais Fernanda Blauth
Signature: 
Date: 03-07-2025

Approval of the Exam Board

Name: Hilda Boskma, on behalf of ExB
Signature: 
Date: 04-07-2025

*Any CF teaching staff member can act as a supervisor.

** For archiving reasons, the internship supervisor has to send all relevant documentation concerning assessment (this form, the student's final report, the assessment form) to the Student Service Desk (cf see@rug.nl).

Appendix B. Internship Logbook

Week	Weekly Tasks	Notes
1	<ul style="list-style-type: none">- Introduced to company, team, and work environments.- Planned and gathered project info.- Explored Jira, and Pipedrive.- Studied Azure, cloud storage, medallion architecture, data modeling.- Reviewed Databricks data.- Set up contact with both supervisors- Created planning overview.	Introduction and setup phase.
2	<ul style="list-style-type: none">- Reviewed existing dashboards and environments.- Developed project ideas.- Researched ML methods including clustering and matching.	Research and foundational work

	<ul style="list-style-type: none"> - Created test Forms for 2.0 project. 	
3	<ul style="list-style-type: none"> - Set up meetings with colleagues to get a better picture of the company. - Connected Google Sheets with Databricks for 2.0 project. - Started on matching code for 2.0 project. 	Data integration setup
4	<ul style="list-style-type: none"> - Reviewed new ideas from colleagues. - Revised plans and goals. - Improved Google Forms questions for 2.0 project. - Explored Pipedrive data. - Exported Pipedrive data and restructured it. - Weekly meeting and GPTG explanation. 	Planning and data prep
5	<ul style="list-style-type: none"> - Fixed data inconsistencies in secondment data. 	Initial modeling and planning

	<ul style="list-style-type: none"> - Created Gantt Chart for better overview. - Created planning and tickets in Jira. - Started on the initial ML model. - Finished matching code for 2.0 project. - Weekly meeting and project go-ahead for revised plans. 	
6	<ul style="list-style-type: none"> - Developed a Classification model. - Integrated matching code for 2.0 project into Databricks - Sent out the Google Forms for 2.0 project, so data collection could begin. 	First predictive models development
7	<ul style="list-style-type: none"> - Added features to Classification model - Conducted feature testing - Enhanced matching code after testing on first data 	Model and code refinement
8	<ul style="list-style-type: none"> - Planned out and started on Pipedrive 	API integration and automation

	<p>API setup in Databricks</p> <ul style="list-style-type: none"> - Automated 2.0 workflow 	
9	<ul style="list-style-type: none"> - Developed 2.0 dashboard ideas. - Connected Team Outings data to PowerBI. - Continued Pipedrive API 	Dashboard development and review
10	<ul style="list-style-type: none"> - Continued working on Pipedrive API - Interim evaluation. - Expanded Classification model with sampling methods. - Integrated Pipedrive API and matching code into production. 	Evaluation and production preparation
11	<ul style="list-style-type: none"> - Continued dashboard work - Sought ML model feedback - Incorporated model feedback - Sought feedback on 2.0 dashboard 	Feedback

12	<ul style="list-style-type: none"> - Continued dashboard work - Started an anomaly detection model after feedback - Tested anomaly detection code - Moved model code to Databricks 	Anomaly detection and dashboard development
13	<ul style="list-style-type: none"> - Incorporated dashboard feedback - Moved all model code to Databricks - Fixed bugs after initial runs of models on Databricks - Started Pipedrive dashboard - Reviewed feedback given on model code 	Feedback integration and starting of Pipedrive dashboard
14	<ul style="list-style-type: none"> - Made presentation - Made documentation for update process of models - Made list of feedback given during presentation 	Presentation
15	<ul style="list-style-type: none"> - Made dashboard pages for report 	Finishing touches

	<ul style="list-style-type: none"> - Last meeting with colleague to go over the models - Finished documentation - Exit meeting 	
--	-----------------------------------------------------------------------------------------------------------------------------------------------------------------	--

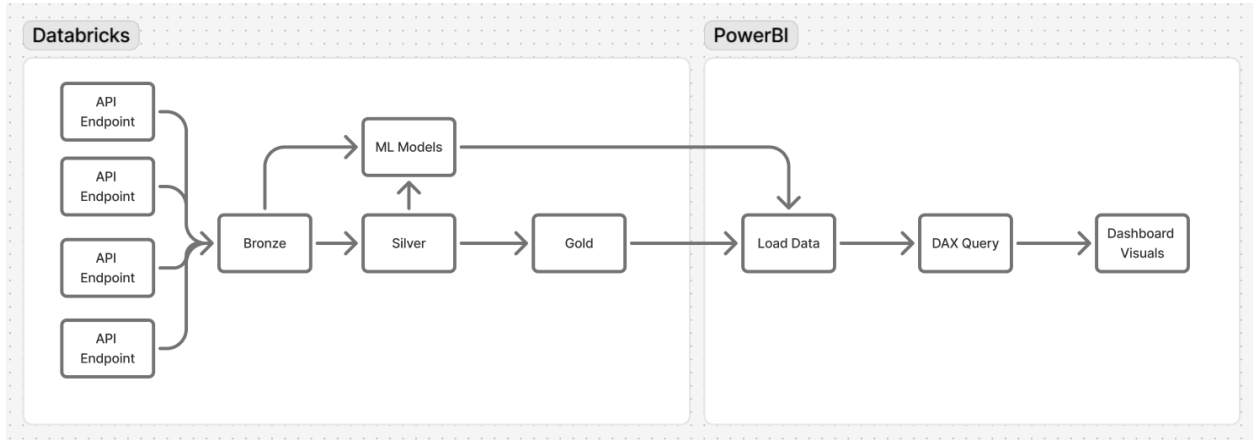
Appendix C. Learning Outcome Table

Category	Learning Outcomes	Evidence/ extent achieved
Functional competence	<ul style="list-style-type: none"> - Demonstrate precision, adaptability, and effective communication skills. 	<ul style="list-style-type: none"> - Regular meetings with external and internal supervisors - Clear progress updates - Adapting to changing data needs
	<ul style="list-style-type: none"> - Handle feedback constructively, incorporating suggestions from supervisors and peers 	<ul style="list-style-type: none"> - Integrated feedback into dashboards design and ML models
Output generation and evaluation	<ul style="list-style-type: none"> - Deliver high-quality outputs such as dashboards and predictive models 	<ul style="list-style-type: none"> - Delivered Pipedrive Sales Analytics Dashboard, 2.0 Employee Dashboard, Classification model, Anomaly Detection

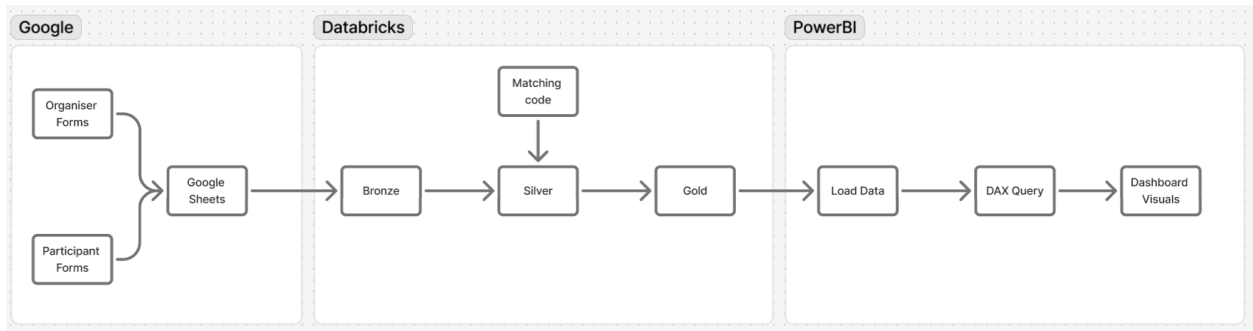
		model
	<ul style="list-style-type: none"> - Ensure outputs meet organisational needs and support decision-making 	<ul style="list-style-type: none"> - Requirements gathered from users - Dashboard in house style - Validated with users
Internship report proficiency	<ul style="list-style-type: none"> - Compile a comprehensive, clear, and accessible report communicating findings effectively 	<ul style="list-style-type: none"> - Structured sections - Explanations of models and dashboard
Reflective analysis	<ul style="list-style-type: none"> - Critically evaluate own skills and learning in relation to DSS programme outcomes 	<ul style="list-style-type: none"> - Evaluation section linking tasks to DSS learning lines and personal learning outcomes
	<ul style="list-style-type: none"> - Identify links between internship tasks and personal learning goals 	<ul style="list-style-type: none"> - Reflection on end-to-end pipelines, ML skills, and data storytelling
Achievement of personal learning objectives	<ul style="list-style-type: none"> - Gain practical competence in ML, data engineering, visualisation, and communication skills 	<ul style="list-style-type: none"> - Built Databricks pipeline, ML models, PowerBI dashboards - Presented and discussed results with stakeholders

Appendix D. Workflows

Pipedrive Sales Analytics Dashboard



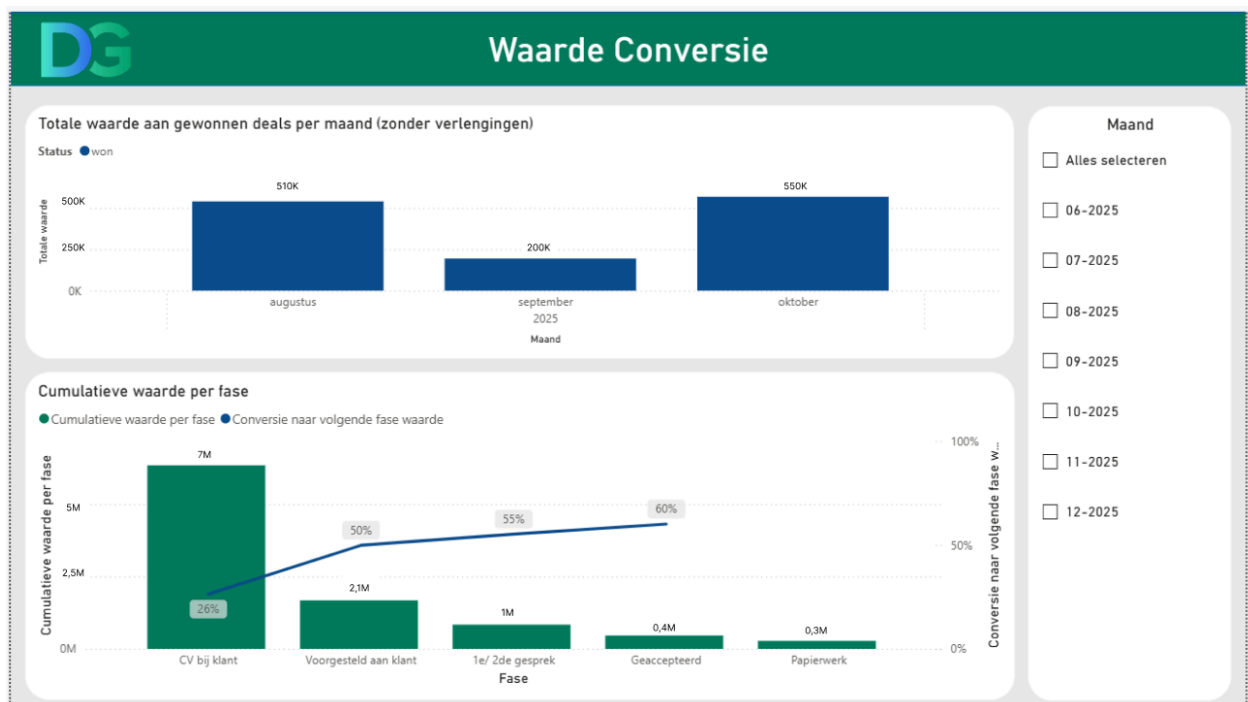
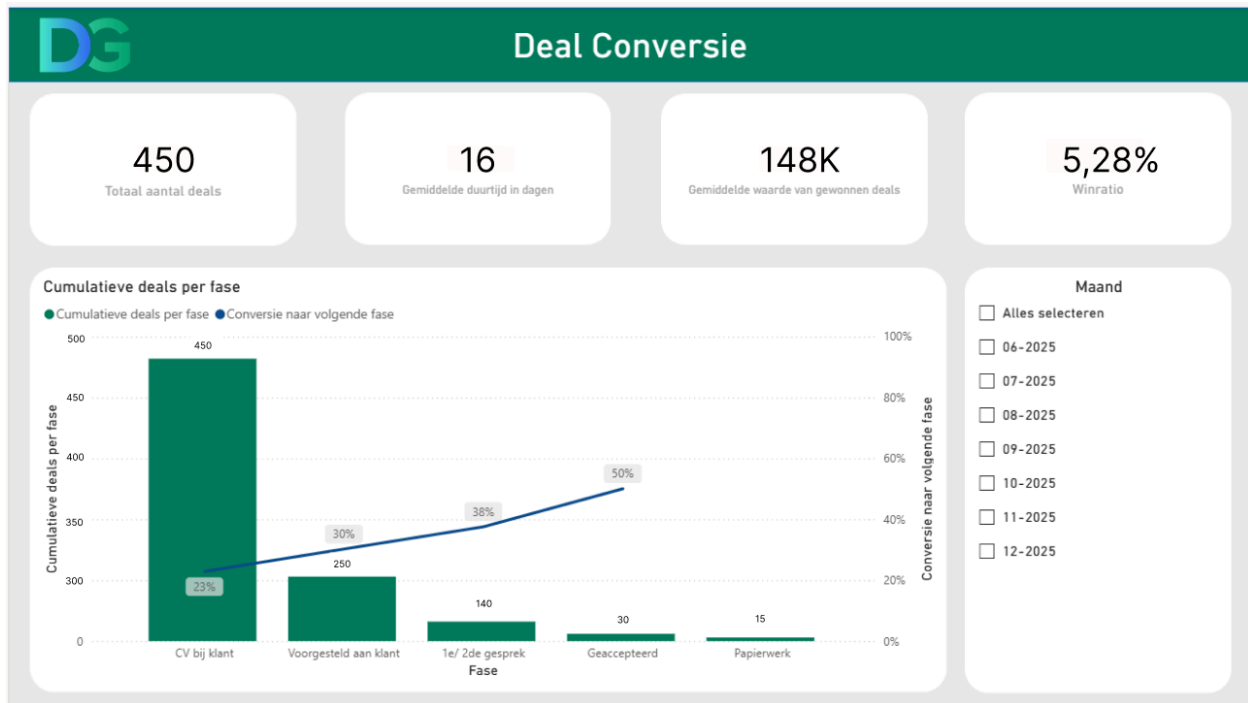
2.0 Employee Engagement Dashboard



Appendix E. Product

The next pictures are of the product as delivered to the company. For privacy reasons names and amounts have been changed.

Pipedrive Sales Analytics Dashboard



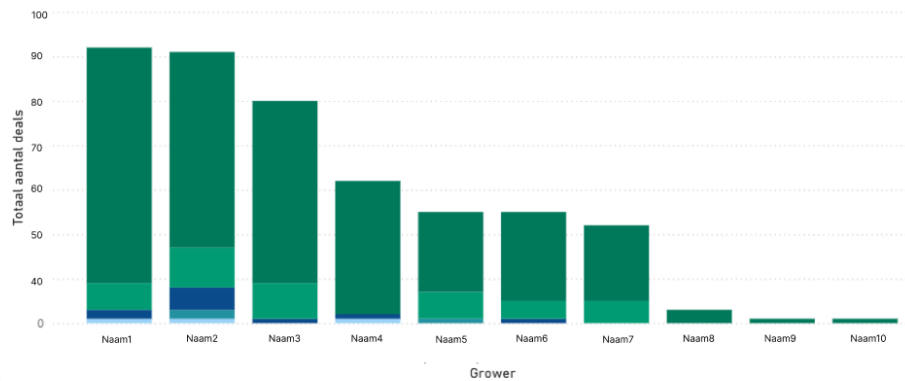
Naam1	Naam2	Naam3	Naam4	Naam5	Naam6	Naam7	Naam8	Naam9	Naam10
-------	-------	-------	-------	-------	-------	-------	-------	-------	--------

7,48%

Winratio

Totaal aantal deals per grower en fase

Fase ● CV bij klant ● Voorgesteld aan klant ● 1e/ 2de gesprek ● Geaccepteerd ● Papierwerk



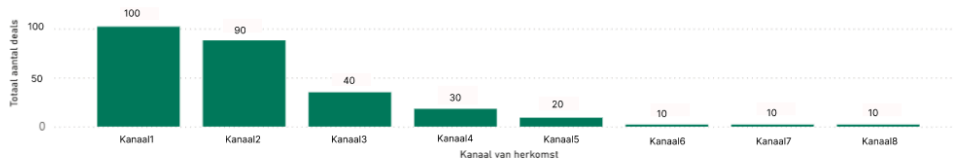
Fase

- ☐ CV bij klant
- ☐ Voorgesteld aan klant
- ☐ 1e/ 2de gesprek
- ☐ Geaccepteerd
- ☐ Papierwerk

Maand

- ☐ Alles selecteren
- ☐ 06-2025
- ☐ 07-2025
- ☐ 08-2025
- ☐ 09-2025
- ☐ 10-2025
- ☐ 11-2025
- ☐ 12-2025

Totaal aantal deals per kanaal van herkomst

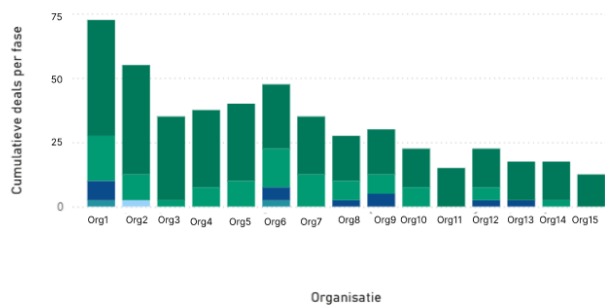


Fase

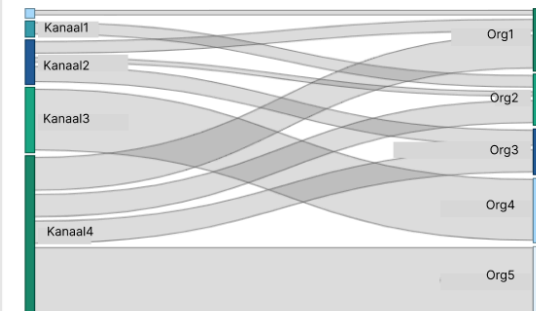
- ☐ CV bij klant
- ☐ Voorgesteld aan klant
- ☐ 1e/ 2de gesprek
- ☐ Geaccepteerd
- ☐ Papierwerk

Cumulative deals per fase per organisatie

Fase ● CV bij klant ● Voorgesteld aan klant ● 1e/ 2de gesprek ● Geaccepteerd ● Papierwerk

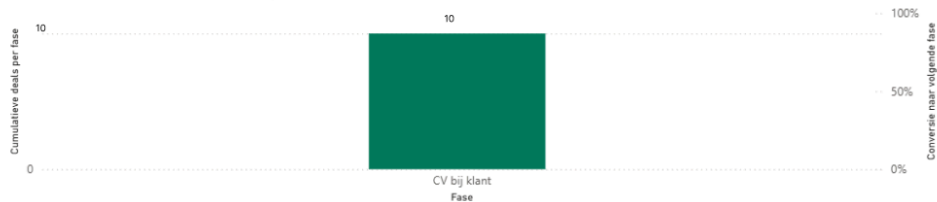


Deal stroming van kanaal naar organisatie



Cumulative deals per fase organisatie 1

● Cumulative deals per fase ● Conversie naar volgende fase



Organisatie 1

- ☐ Organisatie 1
- ☐ Organisatie 2
- ☐ Organisatie 3
- ☐ Organisatie 4
- ☐ Organisatie 5
- ☒ Organisatie 6
- ☐ Organisatie 7
- ☐ Organisatie 8

Cumulative deals per fase organisatie 2

● Cumulative deals per fase ● Conversie naar volgende fase



Organisatie 2

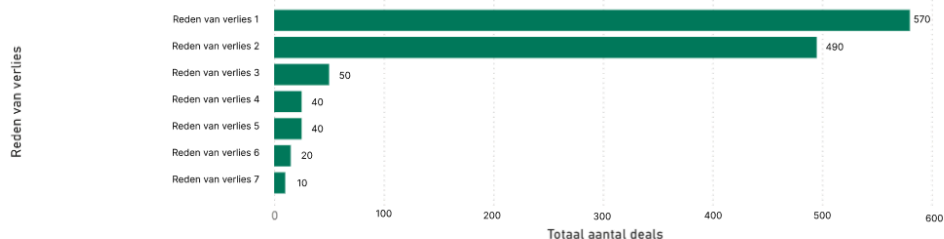
- ☐ Organisatie 1
- ☐ Organisatie 2
- ☐ Organisatie 3
- ☐ Organisatie 4
- ☒ Organisatie 5
- ☐ Organisatie 6
- ☐ Organisatie 7
- ☐ Organisatie 8

Organisatie Conversie

Fase	CV bij klant		Voorgesteld aan klant		1e/ 2de gesprek		Geaccepteerd		Papierwerk	
Organisatie	Aantal deals	Conversie (%)	Aantal deals	Conversie (%)	Aantal deals	Conversie (%)	Aantal deals	Conversie (%)	Aantal deals	Conversie (%)
Organisatie 1	25	100%	25	100%	25	50%	13	100%	13	
Organisatie 2	74	100%	74							
Organisatie 3	10	100%	10	100%	10					
Organisatie 4	62	100%	62							
Organisatie 5	89	100%	89							
Organisatie 6	74	60%	40	33%	12	50%	6			
Organisatie 7	46	56%	29							
Organisatie 8	75	50%	36							
Organisatie 9	14	50%	7	100%	7	100%	7			
Organisatie 10	12	43%	8	33%	6					
Organisatie 11	21	43%	7	67%	4					
Organisatie 12	25	39%	4	43%	2	33%	1			
Organisatie 13	36	33%	12							
Organisatie 14	10	33%	3	100%	3	100%	3	100%	3	
Organisatie 15	25	33%	3							
Organisatie 16	41	33%	16	50%	8					
Organisatie 17	85	33%	32							
Organisatie 18	10	25%	3							
Organisatie 19	20	24%	4	25%	1	100%	1	100%	1	
Organisatie 20	30	17%	3	100%	3					
Organisatie 21	100	17%	17							
Organisatie 22	25	8%	2							
Organisatie 23	75									
Organisatie 24	65									
Organisatie 25	45									
Organisatie 26	40									

Reden van Verlies

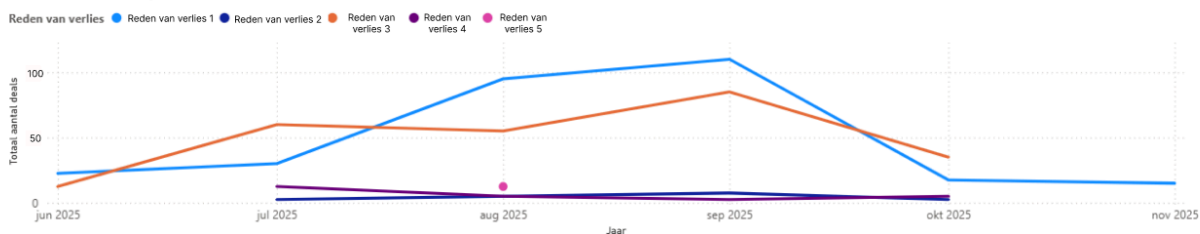
Totaal aantal deals per reden van verlies



Fase

- ☐ CV bij klant
- ☐ Voorgesteld aan klant
- ☐ 1e/ 2de gesprek
- ☐ Geaccepteerd
- ☐ Papierwerk

Totaal aantal deals per maand en reden van verlies





Anomaly Detection

ID	Titel	Grower	Fase	Model Prediction
327	Titel van deal 1	Naam 1	Voorgesteld aan klant	-1
337	Titel van deal 2	Naam 1	Voorgesteld aan klant	-1
338	Titel van deal 3	Naam 1	Voorgesteld aan klant	-1
286	Titel van deal 4	Naam 2	CV bij klant	1
321	Titel van deal 5	Naam 3	CV bij klant	1
322	Titel van deal 6	Naam 1	CV bij klant	1
323	Titel van deal 7	Naam 1	CV bij klant	1
324	Titel van deal 8	Naam 2	CV bij klant	1
325	Titel van deal 9	Naam 1	CV bij klant	1
326	Titel van deal 10	Naam 1	CV bij klant	1
329	Titel van deal 11	Naam 3	CV bij klant	1
330	Titel van deal 12	Naam 1	CV bij klant	1
335	Titel van deal 13	Naam 1	CV bij klant	1
335	Titel van deal 13	Naam 2	CV bij klant	1
336	Titel van deal 15	Naam 1	CV bij klant	1
339	Titel van deal 16	Naam 1	CV bij klant	1
340	Titel van deal 17	Naam 1	CV bij klant	1
342	Titel van deal 18	Naam 3	CV bij klant	1
343	Titel van deal 19	Naam 1	CV bij klant	1
344	Titel van deal 21	Naam 1	CV bij klant	1
345	Titel van deal 22	Naam 1	CV bij klant	1
345	Titel van deal 22	Naam 2	CV bij klant	1
348	Titel van deal 24	Naam 1	Voorgesteld aan klant	1
349	Titel van deal 25	Naam 1	Voorgesteld aan klant	1
350	Titel van deal 26	Naam 2	CV bij klant	1
351	Titel van deal 27	Naam 1	Voorgesteld aan klant	1

Anomaly Detection

De kolom **Model Prediction** geeft aan of de deal normaal is of een afwijking ten opzichte van andere deals. Alleen huidige open deals worden weergegeven.

Labels:

- **Anomaly (-1):** deze deal wijkt af, en gaat waarschijnlijk gewonnen worden.
- **Normaal (1):** deze deal lijkt qua kenmerken op de meeste andere deals en gaat waarschijnlijk verloren.



Classification

ID	Titel	Grower	Fase	Prediction	Win Probability
335	Titel van deal 1	Naam 1	CV bij klant	lost	22%
335	Titel van deal 1	Naam 2	CV bij klant	lost	22%
343	Titel van deal 2	Naam 1	CV bij klant	lost	22%
344	Titel van deal 3	Naam 1	CV bij klant	lost	22%
327	Titel van deal 4	Naam 3	Voorgesteld aan klant	lost	21%
330	Titel van deal 5	Naam 1	CV bij klant	lost	21%
336	Titel van deal 6	Naam 1	CV bij klant	lost	21%
337	Titel van deal 7	Naam 1	Voorgesteld aan klant	lost	21%
338	Titel van deal 7	Naam 1	Voorgesteld aan klant	lost	21%
339	Titel van deal 8	Naam 3	CV bij klant	lost	21%
340	Titel van deal 9	Naam 1	CV bij klant	lost	21%
342	Titel van deal 10	Naam 1	CV bij klant	lost	21%
345	Titel van deal 11	Naam 1	CV bij klant	lost	21%
345	Titel van deal 11	Naam 2	CV bij klant	lost	21%
348	Titel van deal 12	Naam 1	Voorgesteld aan klant	lost	21%
349	Titel van deal 13	Naam 1	Voorgesteld aan klant	lost	21%
350	Titel van deal 14	Naam 1	CV bij klant	lost	21%
351	Titel van deal 15	Naam 3	Voorgesteld aan klant	lost	21%
352	Titel van deal 16	Naam 1	Voorgesteld aan klant	lost	21%
353	Titel van deal 17	Naam 1	CV bij klant	lost	21%
353	Titel van deal 17	Naam 3	CV bij klant	lost	21%
354	Titel van deal 18	Naam 1	CV bij klant	lost	21%
329	Titel van deal 19	Naam 1	CV bij klant	lost	20%
286	Titel van deal 20	Naam 2	CV bij klant	lost	19%
321	Titel van deal 21	Naam 1	CV bij klant	lost	18%
322	Titel van deal 22	Naam 1	CV bij klant	lost	18%

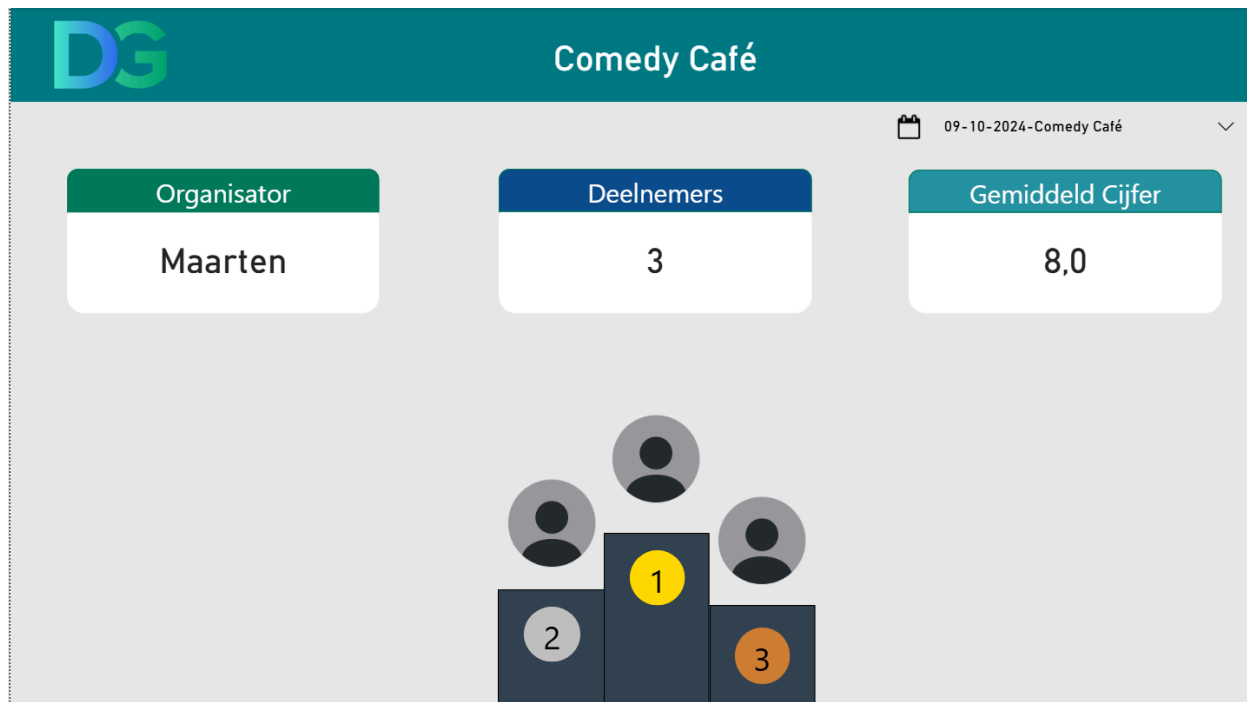
Classification

De kolom **Prediction** geeft aan 'won' of 'lost', en de kolom **Win Probability** geeft aan hoe groot de kans is dat de deal gewonnen wordt. Alleen huidige open deals worden weergegeven.

Labels:

- **Lost:** als de Win Probability onder de 50% is dan geeft het model aan dat deze deal verloren zal gaan.
- **Won:** als de Win Probability boven de 50% is dan geeft het model aan dat deze deal gewonnen zal worden.

2.0 Employee Engagement Dashboard



Alle Activiteiten				
Datum	Activiteit	Deelnemers	Gemiddelde	Organisator
vrijdag 12 juli 2024	Ping Pong	2	7,5	Jesse
woensdag 7 augustus 2024	Poolen	2	7,0	Bram
vrijdag 6 september 2024	Play-in arcade games	2	6,5	Thomas
woensdag 9 oktober 2024	Comedy Café	3	8,0	Maarten
vrijdag 8 november 2024	FC Utrecht	3	7,3	Sam
woensdag 4 december 2024	Karten	2	7,0	Sarah
woensdag 5 februari 2025	Nox room	4	8,8	Lars
donderdag 10 april 2025	RaceSquare	4	7,3	Mees
woensdag 11 juni 2025	Chi Chi	7	8,6	Thomas
donderdag 10 juli 2025	Jeu de Boule	7	7,4	Maarten
vrijdag 8 augustus 2025	VR-Gaming	5	8,8	Bram
woensdag 3 september 2025	Fusion Drift	8	8,1	Olivia
vrijdag 14 november 2025	Cocktailworkshop	7	7,0	Thijs

