Otto-von-Guericke-University Magdeburg
Faculty of Computer Science

Project Report

VLBA II: System Architecture

Authors:

Anja Kaminski, Dinesh Babu Gopinath Hemavathi,
Ferdinand Wild, Sai Nithin Reddy Ankireddy, Yashika Balaji

Examiner: Prof. Dr. Klaus Turowski

Supervisors:

Abdulrahman Nahhas, Dirk Dreschel,
Christian Haertel, Christian Daase

Arbeitsgruppe Wirtschaftsinformatik
Institute of Technical and Business Information Systems
Otto-von-Guericke-University Magdeburg

Table of Contents

1 Introduction

In today's business world, the fusion of data analysis and Machine learning has become very crucial in guiding strategic moves as well as remaining competitive. Mobility Worldwide (MWW) is a multinational organization located at Magdeburg headquarters and its activities cover different areas such as transportation innovation, manufacturing, talent development plus support. To ensure they stay number one and ahead of their competitors, MWW have always been quick to adopt new technologies which are integrated seamlessly in the present systems to enhance performance on various fronts.

MWW's Research and Solution Development (RSD) department is keen on being at the forefront of new product and technology developments. This necessitates the management of multiple projects that call for a blend of expertise from professionals with various skill sets and academic backgrounds. However, due to the critical nature of these assignments, RSD's undertakings thrive based on its staff's competency levels as well as qualifications necessary when executing them. For this reason, hiring competent personnel who will foster creativity and realize strategic goals is critical.

MWW's board has acknowledged that there is a need to have a recruitment process that is strong and which depends on data. In effect, their objective is clear-cut; they aim at using internal business information plus external sources to improve their hiring strategy. The purpose of this is to reveal the kinds of candidates who always help in the success of projects, thus enabling the corporation to cut down on the time it uses for acquiring talents. This also seeks to foster global university partnerships so as to guarantee an ongoing stream of high caliber individuals who can be hired.

This paper investigates an extensive and in-depth exploration and analysis of internal and external datasets derived from SAP Datasphere system, as well as other relevant data sources. Its objective is to derive results that provide hands-on information on how best to adopt a strategic data-driven approach to recruitment.
The steps are the following:

1. Data Exploration
2. Predictive Analytics
3. Strategic recommendations

This analysis utilizes a wide array of internal data consisting of employee information and other aspects pertaining to projects such as; contract details, general employee demographics, qualifications, project costs, and membership information. The analysis has benefited from the use of university rankings and other sources, since they offer a framework for assessing the caliber and suitability of academic establishments that could collaborate with MWW on talent acquisition. The study uses this data collection to identify underlying patterns and relationships that support project success.

In conclusion, this report proposes a data-oriented recruitment approach which will not only heighten workforce quality but also guarantee continued success at MWW.

1.1 GCP Integration with SAP Data Warehouse Cloud

The internal data used in the integration between SAP Data Warehouse Cloud (DWC) and Google Cloud Platform (GCP) is typically exported from SAP DWC or the SAP system used

in the backend[1]. For smaller datasets, the data can be exported from SAP in well-structured tables. However, for a seamless transition of data between SAP DWC, it is essential to establish a connection between the interfaces.

To achieve this, SAP DWC can enrich the data by creating an Entity-Relationship (ER) model. This allows the data stored in the backend to be imported into SAP DWC and organized accordingly. New mappings can be created within SAP DWC based on the imported data, enabling the linkage between different datasets. SAP SAC (SAP Analytics Cloud) can then be used to create views and visualization interfaces, facilitating data preparation and initial business result evaluation.

SAP Business Technology Platform, which includes SAP Data Warehouse Cloud and SAP Analytics Cloud, is built upon an on-premise SAP S/4 HANA system in the backend. This on-premise system provides the foundational ABAP-based structure necessary for the typical SAP landscape. SAP Analytics Cloud primarily focuses on creating data reports from both internal and external data sources, including real-time data from on-premise systems and external sources like MS Excel. On the other hand, SAP Data Warehouse Cloud offers multiple levels of data processing and can interact with Google Cloud Platform for enhanced capabilities.

By leveraging the integration between SAP DWC, SAP Analytics Cloud, and GCP, organizations can efficiently manage and analyze their data, enabling better decision-making and insights.

In summary, there are several ways to establish an interconnection between SAP Data Warehouse Cloud (DWC) and Google Cloud Platform (GCP):

a) **Data federation:** Data federation allows multiple databases to be linked and function as one, providing a single data source for further processing. GCP's Dataproc tool can be used to connect SAP DWC to GCP, enabling data manipulation at cloud scale and integration with other GCP services like BigQuery.

b) **Data flow:** SAP DWC can be connected to GCP via Google Cloud Storage Bucket. After setting up a bucket and a service account with the necessary roles, the interconnection can be established using SAP DWC. This option allows for table updates and offers different modes for table filling.

c) **Data federation via REST API:** A REST API can be used to establish a connection between SAP DWC and GCP, specifically using GCP's BigQuery tool. This enables interactions between the two platforms and facilitates data exchange.

These integration options provide flexibility and enable seamless data transfer and processing between SAP DWC and GCP, allowing organizations to leverage the capabilities of both platforms for their data warehousing and analytics needs.

1.2 Services used

GCP (Google Cloud Platform)[2]:

- GCP serves as the hosting provider for our project.
- It offers a wide array of tools and services that cater to diverse use cases, ranging from development and deployment to data analytics and machine learning.

BigQuery:

- BigQuery is a serverless and highly scalable data warehouse that allows for the analysis of large datasets using SQL.

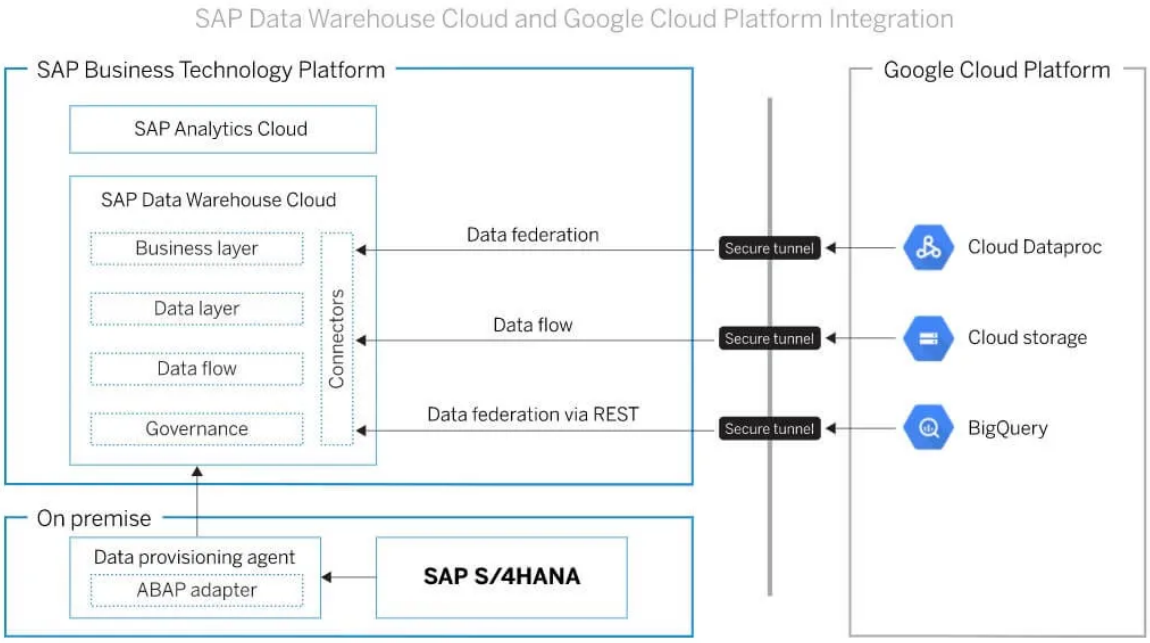


Figure 1: GCP and SAP DWC Integration

- We leveraged BigQuery to create views on our existing data, facilitating efficient data analysis.

BigQueryML:

- BigQueryML enables users to build and execute machine learning models directly within BigQuery using SQL queries, eliminating the need for additional programming or data movement.
- We utilized the views created in BigQuery to construct regressor models, selecting the model with the least mean absolute error and high accuracy.

Vertex AI:

- Vertex AI offers a comprehensive suite of tools and services designed to streamline the machine learning development process.
- It abstracts away the complexities of managing infrastructure and resources, allowing users to focus on model building and deployment.
- Vertex AI encompasses AutoML capabilities, enabling users to build ML models with minimal manual intervention.
- It supports various use cases such as image classification, text classification, tabular data analysis, and video analysis.

Kubeflow:

- Kubeflow is an open-source ML platform built on Kubernetes, a popular container orchestration system.
- It simplifies the deployment, management, and scaling of ML workflows in a cloud-native environment.
- Kubeflow Pipelines enables the creation of reusable and reproducible pipelines for complex ML workflows, encompassing data preprocessing, model training, evaluation, and deployment.
- We utilized Kubeflow SDK to create our ML pipeline.

Looker Studio:

- Looker Studio provides a platform to visualize data through configurable charts and tables.
- It offers seamless connectivity to various data sources.
- Users can easily share insights and collaborate on reports with their teams.
- Built-in sample reports accelerate the report creation process.

1.3 Analytical Insights with BigQuery

To drive our data-driven recruiting strategy and gain valuable insights into the factors influencing project success, our workflow consists of two main steps leveraging the power of BigQuery.

Step 1: we import and integrate relevant data sources, such as the current employee pool and past project outcomes, into BigQuery. Through the creation of views, we organize and structure the data, performing necessary transformations and aggregations to ensure accuracy.

Step 2: we leverage SQL queries and BigQuery's analytical functions to analyze the prepared data. This enables us to generate views capturing employee performance, project outcomes, and recruitment metrics.

To visually represent and communicate these insights, we utilize Looker Studio. This enhances the clarity and accessibility of the insights, allowing us to share and collaborate on reports with our team. By leveraging Looker Studio's capabilities, we can harness the power of data visualization to inform and shape our data-driven recruiting strategy, leading to more informed and effective decision-making.

1.4 Architectural Concept

The following diagram shows the architecture of our project. The base of our project consists of multiple CSV Files which are imported into database tables in BigQuery. There, multiple views and tables are being created. For representation and visualization, Looker Studio has been used. Another important set of used components have been set up by the kubeflow-configuration persisted within the YAML-files and for the machine learning task and the prediction procedure the vertex-ai solution accessible within the cloud has been used. The proposed architectural realization of our approach follows certain consideration of reusability and reliability that are present within the cloud approach. To achieve this we aimed at certain cloud pattern approaches which are explained below.

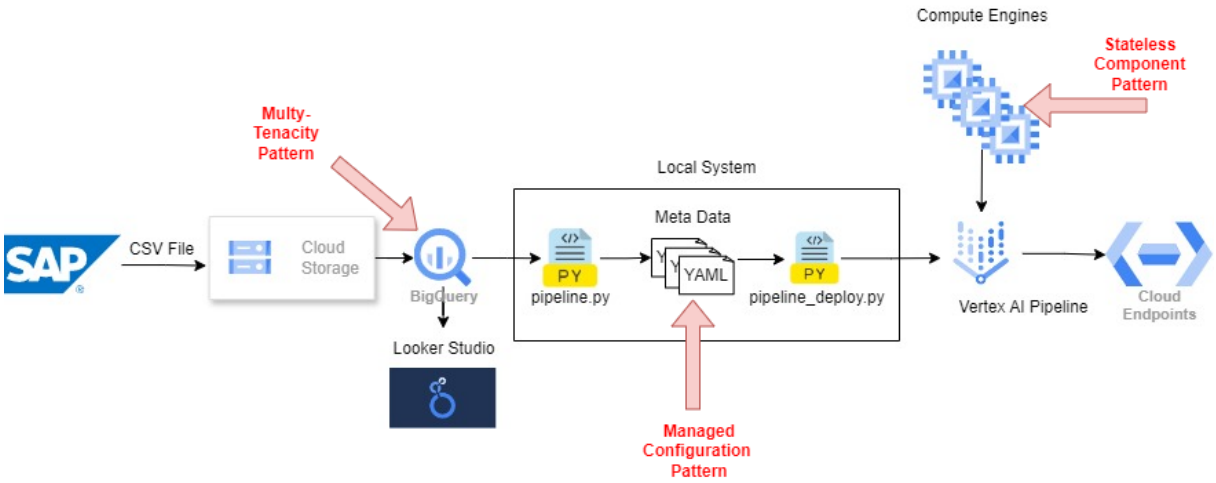


Figure 2: Architecture Model - Pattern Overview

1.4.1 Managed Configuration Pattern

Within our pipeline configuration, we allow the up- and down-scaling of resources like training-instances within Vertex-AI to achieve an optimal setup to balance our resources while also serving our demands. To achieve this, we tried an Infrastructure as Code approach where the YAMLs that define the resources spent and model used can easily be regenerated and stored and later exchanged based on changes within requirements.

1.4.2 Multi-Tenancy Pattern

As we various different data exploration approaches have been performed by different project members on the same storage-bucket where our data is persisted within BigQuery, which is also the component that provides the input data for the machine learning model we would also suggest following the Multi-Tenancy-Pattern here, to reduce costs and inconstancy between the data analysis and the machine learning task while allowing especially for the data exploration, performed by several people comparability and constituency.

1.4.3 Stateless Component Pattern

As for the machine learning task within the vertex-ai-pipeline in contrast to a local setup it was decided to have stateless training resource-instances which are generated and provided at the start of the deployment process before a new pipeline setup. This allows the dynamic recalling of the resources to match the demands considered for each run individually and allows the avoidance of previously performed runs to interfere or interact with the current run.

2 Part 1 - Data Exploration and Descriptive Analysis

The first task given by MWW's RSD department is to implement a data-driven recruiting strategy. For this, we need to analyse the employee and project data need. We should look for general trends and factors for project success.

2.1 Preparation

2.1.1 Goal of data exploration

The task we were given is to help MWW implement a data-driven recruiting strategy. The company would like to know the influence of employee qualification on project success, to recruit candidates that increase project success. To that end, they plan to cooperate with universities to gain access to promising talents. Thus, our task task consists of three steps: First, we have to determine general factors of project success, to distinguish between their effect and the actual influence of employee qualification. Then, we have to determine the employees with high and low project success and determine the general differences in their education. Lastly, and in a different chapter, we have to combine the employee project success with their institution of higher education, to see which universities produce the best candidates.

2.1.2 Data Gathering and Preparation

We did not gather any additional data, as we deemed the the data on the employees and projects of MWW, as well as the ranking of the universities, we received is sufficient to answer the question how the employee qualification affects project success. To prepare the data for analysis, we integrated it into Google Cloud BigQuery. The data we received was already correctly formatted, so no cleaning step for the values was necessary. The things to note during checking for cleaning was, that one of the columns in the data set 'University_ranking' could contain null entries, which we will account for during data exploration and analysis. The column naming isn't consistent (e.g. ID vs. Project_ID, Member_Name but MemberId), but with the low number of columns and few derivation, we can just work with the existing column names. Additionally, looker studio can't process column names with a space, so these have to be renamed for visual analysis.

The relationship of the data sets we received can be seen in the following ER-diagram.

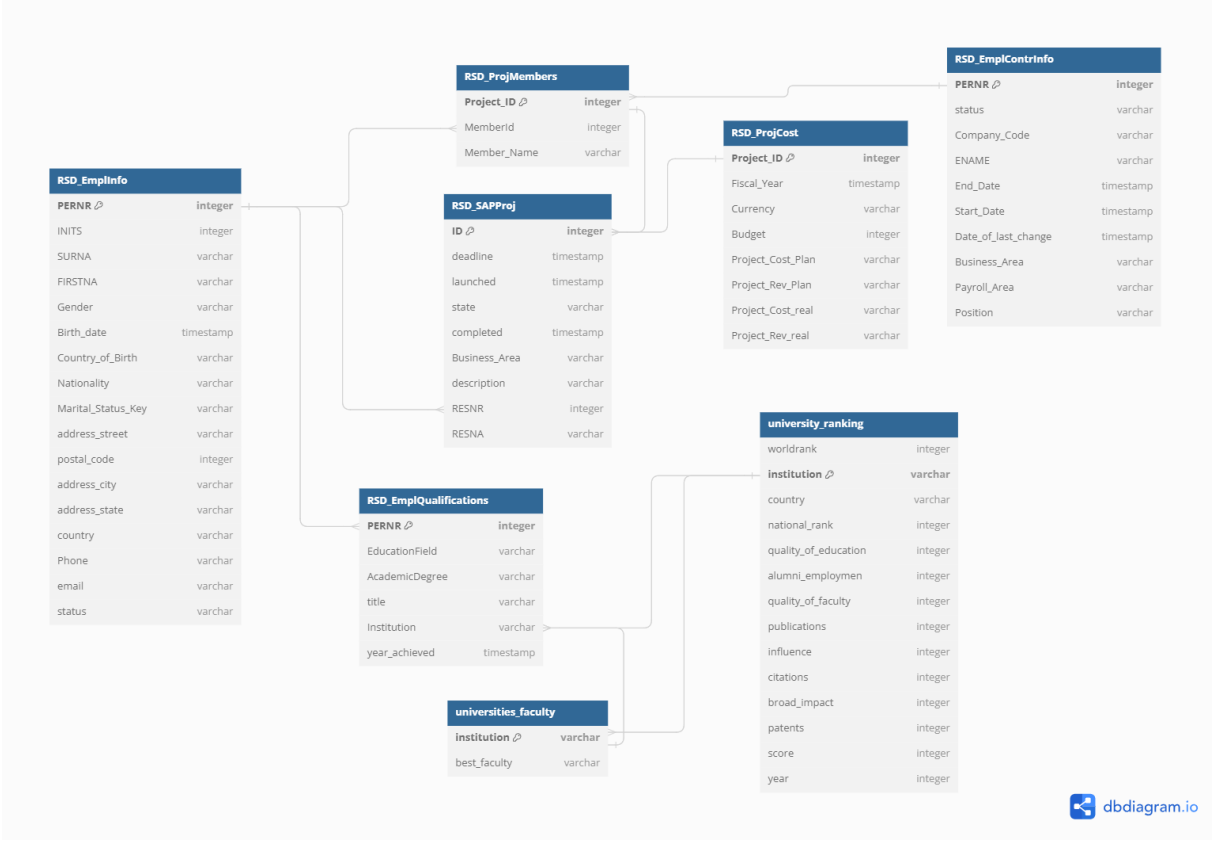


Figure 3: ER-Diagram

2.2 Tooling

Considering the tools that have been used to work with the given dataset, it was required by the task description to use BigQuery for the Data-Exploration and Looker Studio for the visualization. As BigQuery on its own is a fully managed enterprise data warehouse aiming to support managing and analyzing of given data in various forms and helping with the following analysis as well, it's already a tool that provided sufficient capabilities to fulfill the given tasks. At first, the given data was added to the BigQuery-storage within our project to allow direct access via the integrated query-system onto the data. Then the decision between plain SQL and the integrated python-notebook, comparable with Jupyter-notebooks, turned out to favor the SQL-approach over the python variant.
The main reason for choosing SQL-queries as the only direct tooling for the given dataset considering the exploration task was that it's simple to use and doesn't require additional APIs to be activated like the notebook would. In addition to that, the main benefit that the notebook variant provided would be that python allows directly different ways to not only explore the data but also visualize it directly. But as the Task description requires us to use a designated tool for the visualization, called Looker Studio this advantage loses its impact as multiple tools

for visualization seemed rather unnecessary. Finally, also the Project-Costs had been impacting this decision, as BigQuery offers a monthly budget of 1TB of free data-transfer within its built-in-system, which, considering that by average the processed SQL-queries maximum that we performed ranged with-in 15-50MB of size would mean it's would hardly be possible to exceed the free-use-budget from Google which then would mean the project resources can be invested in different tasks elsewhere.

2.3 Data Analysis

To develop a data-driven recruiting strategy and predict the outcome of MPD projects, we received both internal and external data sets. The internal data sets consists of each three tables of employee and historical project data of MWW, the external data sets of two tables with information on institutions of higher education. As a first step, we familiarized ourselves with this data and try to determine general trends and pattern in the internal data sets with regards to project states.
To that end, we look at the meaning and possible values of the different columns and try to find single or combinations of columns that could possibly influence the project state.
As the possible predictors for project success, we identified 3 areas for further analysis:

1. Employees
2. Financing
3. Environment

Each of these areas consists of different factors we will look at in detail when we analyse the given data. For the employees, the possible influence factors we identified are success rate of project responsible, success rate of all project members, the factor of employees working outside their business unit and potential benefits of consistent teams. Financing is the influence of budget and planned duration of project success. Environment includes possible influence of the year the project took place, especially influence of the global pandemic between 2020-2022, or the Business Unit responsible for the project. We begin by analysing the possible influence of factors outside of employee qualification.
The results of this analysis are described in the following chapter, with texts describing the most important conclusions, diagrams highlighting our findings and a table summarizing all results for each internal data set. The possible influence on the project state we assume for each attribute (column) is described under 'Influence' in the tables. We defined the following possible values for influence:

- None: The attribute can't possibly have any influence on the project state, e.g. phone numbers of project members.
- Unlikely: Data analysis shows no discernible influence or the change in state results are so minimal, they might just be noise.
- Possible: Some attribute values show an deviation from the average project state results that might be large enough to depict an influence of this value, but many don't, so the overall variation in state results could just be noise.
- Likely: Data analysis indicates the attribute influences project success.

- Strong: Data analysis indicates the attribute influences project success to a great extent.
- Not as-is: The attribute in itself can not be used for state prediction, but possibly in combination with other attributes, e.g. only the launch-date of previous projects will not be helpful for predictions about currently active projects, but the planned duration we can calculate with the deadline in relation to the budget might be.

To be able to conclude an influence on the resulting project state, we first need to determine the average project states of all MWW projects 4, 5. The result is, that the majority of project at MWW fail.

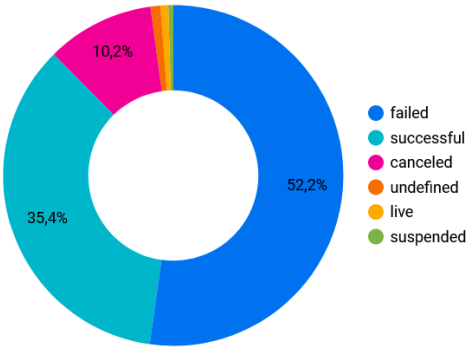


Figure 4: All project states

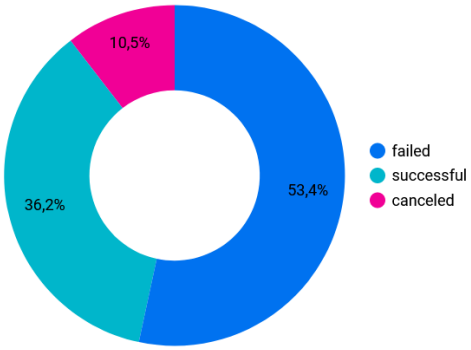


Figure 5: Finished project states

2.3.1 Internal data: MWW employees

RSD EmplContrInfo

From the contract information of all employees at MWW, we first identified PERNR, ENAME and 'Date of last change' as attributed that have no influence on project state, as they are a numerical identifier for employees, the name of an employee and (we assume) the date the employee data was last changed. Furthermore, the Company-Code has the same value, 'MWW-US', for all employees, so it cannot be used for state predictions with the given data sets either.

There is a slight raise in project failure for inactive (former) employees (see 6, but we have to keep in mind that the project teams that are the basis for this diagram can contain both active and inactive members, so its difficult to determine an influence. Additionally, no contract ended before a project deadline and whether employees left between the deadline and actual project can only be determined for successful projects, as no other projects have a date for the actual project end given. With that, we cannot say if project members leaving before the project completion negatively impacts the success, but we can say that there are successful projects that lost at least one project member, in some cases even the employee responsible for the project, after the deadline and before the completion, so it is unlikely to impact the project state significantly. To determine an influence of start date, we could put the start date in relation to the project

start and thus group employees into experience groups depending on the time they already work for MWW, but as the number of projects within that time varies from employee to employee and previous experience in other companies as well as individual learning curves influence results, even this data does not seem to be an adequate factor for predicting project success. Thus, we don't consider 'Start Date' a likely influence to project state.
For 'Business Area' (appendix: 21), 'Payroll Area' (appendix: 22) and Position (appendix: 23), our analysis showed no correlation to project success, so we assume an influence is unlikely.

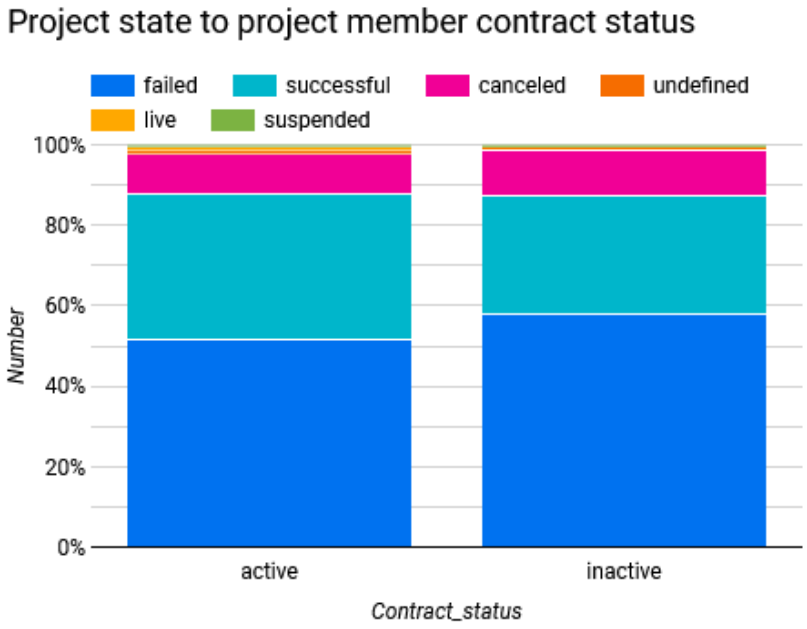


Figure 6: Project state in relation to project member status

RSD EmplInfo

The table RSD_EmplInfo contains data that can be described as personal information of the 62345 employees MWW has ever employed. Given the nature of this data, a lot of columns can be excluded from possible influencing factors right away, like attributes relating to the employee's name or address.
When analysing the remaining attributes, we came to the conclusion that 'Gender' has no influence on project state and 'Marital Status Key' is unlikely to influence the project state, with the possibility of slightly more failed projects for widowed employees (see appendix: 24, 25). For 'Country of Birth', an influence seems questionable, especially with all employees living in the United States, but several of the values diverge significantly from the average project states, like the 23,76% successful projects of Angola which are almost only two thirds the average project success rate of 35,4% (from 4), so we cannot fully exclude this attribute. The results for 'address_state' are similar, but there are only a few values with relatively large divergence (compare 8 to 7), so we deem this attribute less likely to influence the project state while not excluding it from possible influencing factors just yet.

RSD EmplContrInfo			
Table column	Values	Description	Influence
PERNR	Integer	The unique identifier of each employee within MWW	None
status	String	{'active', 'inactive'}, for currently employed (43711) and former (18634) employees	Unlikely
Company-Code	String	'MWW-US'	None
ENAME	String	Employee name	None
End Date	Date	Date the contract of an employee ended (only for 'inactive')	None/ Unlikely
Start Date	Date	Date the employment at MWW began	None
Date of last change	Date	Date employee data was last changed?	None
Business Area	String	{'MS', 'MPD', 'RSD', 'CCR'} - Business Area the employee is employed at	Unlikely
Payroll Area	String	{'BW', 'MP', 'WP'}	Unlikely
Position	String	Employee job position, {'Staff', 'Developer', 'Product Manager', 'Project Manager', 'Service Manager', 'Business Manager', 'Senior Developer', 'Warehouse Manager', 'Production Manager', 'Business Consultant', 'Field Service Representative', 'Business Relationship Manager'}	Unlikely

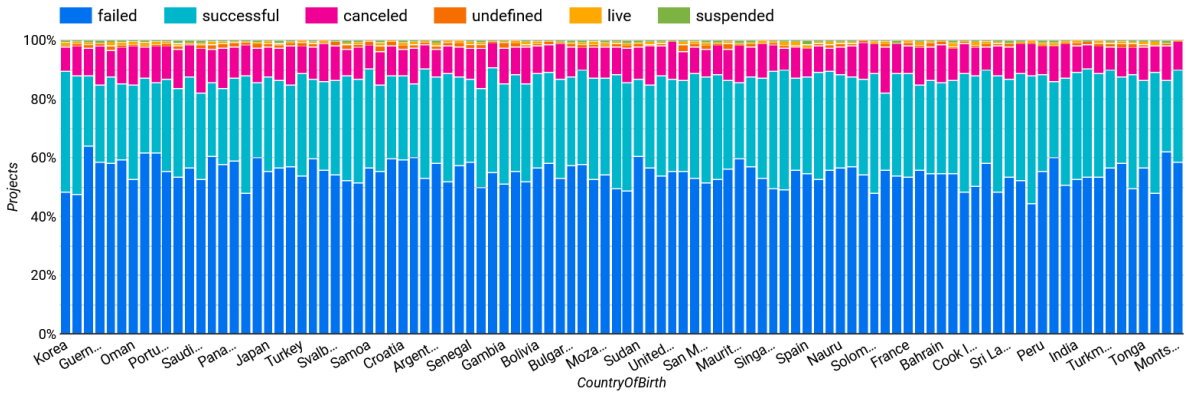


Figure 7: Relation of employee's country of birth to project state

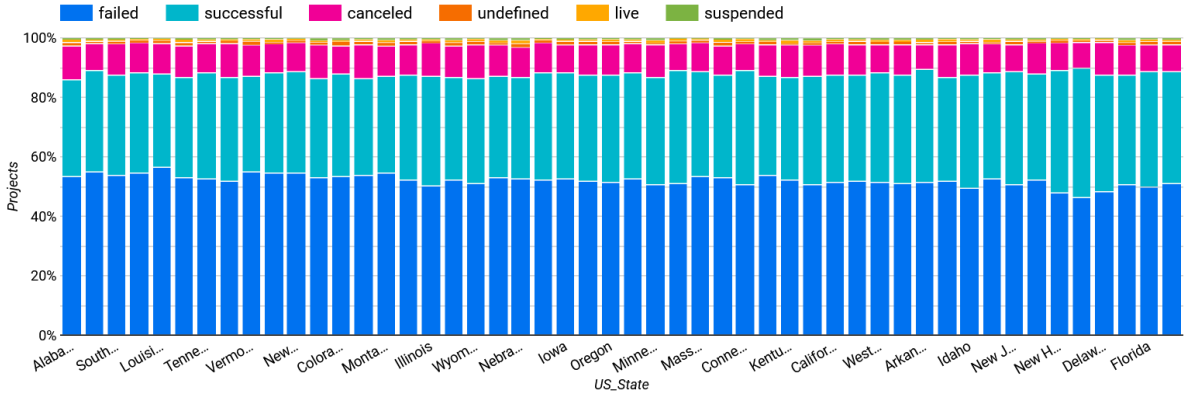


Figure 8: Relation of state the employees live in to project state

RSD EmplQualifications

Contains information on the employees academic degree (at most one degree per employee). Important to note in this data set is, that each employee has an institution listed, even if they don't have a degree.
The field an employee was educated in has no apparent connection to project success (appendix: 26). Whether employees hold an academic degree, which degree it is and when it was achieved does influence the project state, with the employees success rate being higher the higher their degree and the longer they already hold the degree (see 9, 10, appendix: 11, with an alternative diagram on academic titles in appendix: 27).
Connecting the project result with employees' institution, we can see that different institution produce attendees and graduates with vastly different success rates (see appendix: 28, 29). Many universities have only a small number of projects and attendees/graduates that took part in these projects, so the data might not be very reliable for them, as one untypically successful or unsuccessful graduate or attendee with many projects could significantly affect the overall result of the institution, making this a less desirable attribute for project state prediction.

RSD EmplInfo			
Table column	Values	Description	Influence
PERNR	Integer	The unique identifier of each employee within MWW	None
INITS	String	Initials of the employee	None
SURNA	String	Employee's surname	None
FIRSTNA	String	Employee's first name	None
GENDER	String	{'F', 'M', 'D'} with {29938, 29969, 2438} employees each	None
Birth date	Date	Date the employees was born	Not as-is
Country of Birth	String	Country the employee was born in (223-295 per country, except for Congo and Korea with 502 each)	Possibly
Nationality	String	'US'	None
Marital Status Key	String	{'Married', 'Single', 'Divorced', 'Widowed'} with {32372, 20412, 5708, 3853} employees each	Unlikely
address_street	String	Street the employee lives in (all with number in front, occasionally with apartment number at the end)	None
postal code	Integer	postal code of employee's address	None
address_city	String	city the employee lives in	None
address_state	String	state the employee lives in, one of the 50 states of the US	Possibly/ Unlikely
country	String	Always 'US'	None
phone	String	Employee's phone number	None
email	String	Employee's email address (at MWW)	None
status	String	{'active', 'inactive'}, for currently employed (43711), and former (18634) employees	Unlikely

RSD EmplQualifications			
Table column	Values	Description	Influence
PERNR	Integer	The unique identifier of each MWW employee	None
EducationField	String	{'Other', 'Engineering', 'Economics', 'Computer Science'} with {8483, 11609, 17558, 24695} each	Unlikely
AcademicDegree	Boolean	Wether the employee has a degree or not (true=43162, false=19183)	Likely
title	String	{null, 'Bachelor Degree', 'Master Degree', 'Doctoral Degree'} with {19183, 25947, 15079, 2136} employees each	Possible
Institution	String	1024 institutions, having 9 to 6240 attendees each (only 799 universities have graduates)	Likely
year_achieved	Float	Year the highest academic degree was achieved (if one was achieved), 1973-2022	Likely

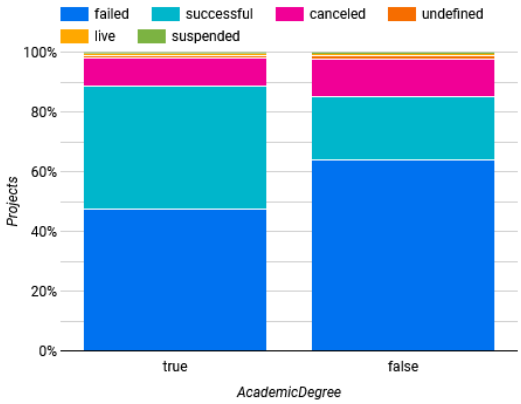


Figure 9: Relation of employees having received a degree to project state

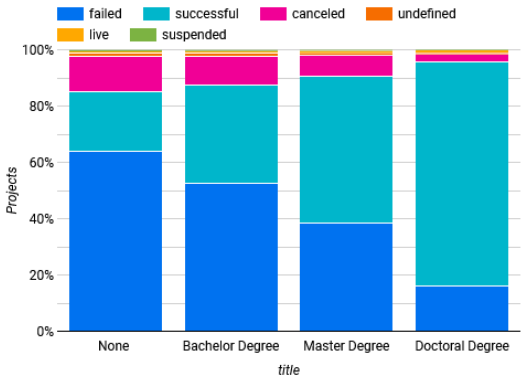


Figure 10: Relation of academic title employee holds to project state

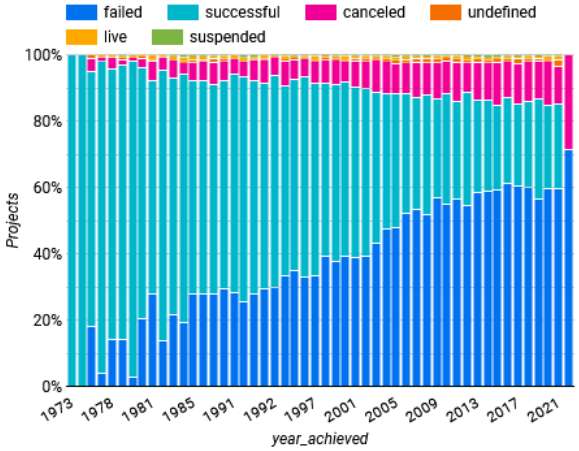


Figure 11: Relation of the year the employees received their degree to project state

2.3.2 Internal data: MWW projects

RSD ProjCost

Information on the costs of all 10000 projects of MWW. Like before, the attribute 'Project-ID' cannot be used to make a prediction of project state, as it is unique for each project. Neither can 'Currency', as it is 'USD' for every project and 'Project Rev_real', which is only greater than 0 for successful projects, so using this attribute for predicting the final state of suspended and live projects would increase the likelihood of the prediction result being failed or cancelled, without any real connection of the attribute value to this outcome.

The values for 'Project Rev_Plan' seem to be depending on 'Project Cost_Plan' or 'Budget' or vice versa, and we decided focus on these attributes instead and were not able to find a correlation of 'Project Rev_Plan' to project state that is independent of these attributes. Between the 'Project Cost_Plan' and 'Budget', the former cannot exceed the latter, so we will focus on 'Budget' as a possible influence to project state.

For 'Fiscal Year', we excluded 1974 for questionable integrity (based on the deadline, 2014 seems correct, but we can't say for sure) and the current 2022 as it seems to be the year the data was extracted and thus in an incomplete state (i.e. it is not the data of a typical year, thus it would

falsify the analysis results). Whether the year has some influence is difficult to determine, but there does not seem to be any negative impact from the Covid19-pandemic. There seems to be a raised probability for failure for projects started in 2013 (see 12), possibly because they just started project work and thus lacked experience, but also possibly because the much lower number of data points for this year is shifting the result. The difference between the lowest failure rate in 2014 of 42,13% and the highest of 57,2% in 2018 is 15,07% which is no absolute factor, but can't be completely ignored as well. We thus consider 'Fiscal Year' a possible weak influence on project state, that could be used but is most likely not essential for correct predictions.
'Project Cost_real' could give a weak indication to final project state in combination with 'Budget', as cancelled and failed projects all exceeded their budget, but roughly half of the successful projects exceeded their budget as well, so project costs below the budget would indicate project success. As roughly half of the successful project exceed their budget as well, it would however not be a good indicator of project state in case costs rise above the budget. Additionally, none of the live or suspended projects currently exceeded their budget, with live projects not even having any cost given, but we can't conclude from the data how much completing the project would cost and thus it is impossible to tell whether the projects will exceed their budget or not.

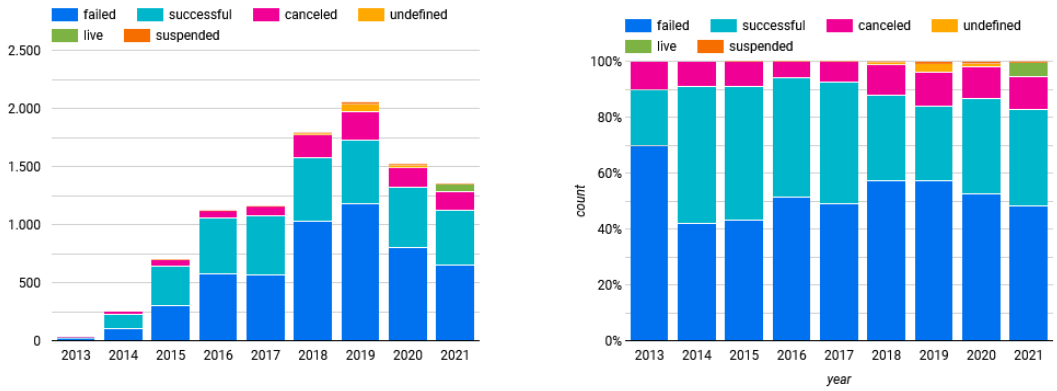


Figure 12: Project state of projects of the fiscal years

As projects have vastly different Budgets, we tried to group the budgets into groups to see whether it they influence project state. As such a grouping is rather arbitrary, we decided to use two different ranges, to reduce the possibility of indicated influence on the chosen ranges.
The groups we've chosen for the first grouping are 'low' with a budget up to 50000 USD, 'mid' with a budget between 50000 and 300000 USD and 'high' with a budget above 300000 USD. The result (see 13) indicates that there might be a higher success rate the lower the project budget, but as we defined the budget ranges, this isn't preferred for predicting the project state.

The second group consists of:

- larger: → over 100.000
- large: → larger than 50.000 and below 100.000
- mid: → larger than 10.000 and below 50.000
- small: → larger than 5.000 and below 10.000

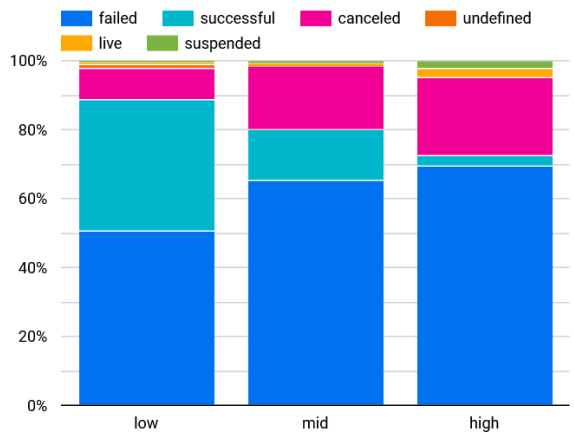
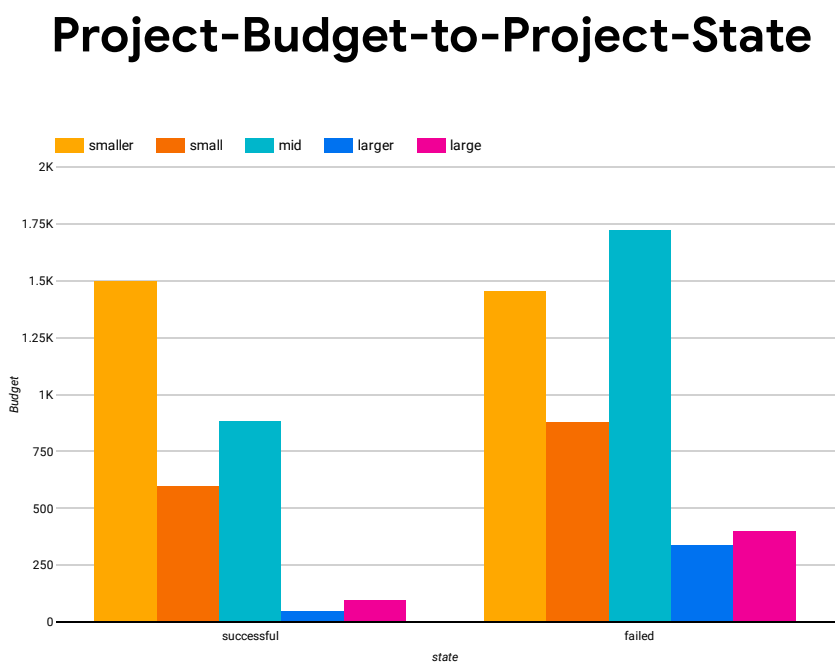


Figure 13: Project states of the different budget ranges

- smaller: → smaller than 5.000

Based on the performed Budget-Grouping a representation of each project-set within a group and its corresponding known outcome being successful or failed is shown in the diagram below.



Exploring the results of this representation shows that while there isn't really a significant difference for the smaller project-group considering the project-outcome, a slight trend can be derived. As by comparing the ratios of the different chunks for both of the project-states in the above graph, the larger the Project-Budget is, the lower the ratio between successful to failed appears to be. This highlights the importance of severely improving the impact or quality of other successful-driving factors from other domains, especially on large and larger projects, as the majority of these projects seem to fail according to the given data. This is also crucial as of course with a larger invested project budget the financial burden of a failed project's impact is severely increased.

RSD ProjCost			
Table column	Values	Description	Influence
Project-ID	Integer	Unique identifier of the project	None
Fiscal Year	Integer	1 project in 1974, otherwise 2013-2022	Possible/ Unlikely
Currency	String	'USD'	None
Budget	Integer	0 to 65601030	Not as-is, Possibly
Project Cost_Plan	Integer	Planned project cost, at most the full budget (39 projects)	None
Project Rev_Plan	Float	Planned revenue, from 0.74 to 58000000.0	None
Project Cost_real	Integer	Actual project costs (8021 above, 9 equal to, 1970 below Budget)	Unlikely
Project Rev_real	Integer	Actual revenue, from 0 to 1118475	None

RSD ProjMembers
Information on the MWW employees that participated in the projects. Interestingly, only 30072 employees have participated in projects so far. While none of the values itself can possibly affect the project state, the employee behind the PERNR and PERNA, in other words, the columns/attributes of other tables that can be connect with the key PERNR, for example their qualification, certainly do have an affect. This table is therefor necessary to link the employees to the projects they worked on, to determine the influence of their attribute values on the project state.

RSD ProjMembers			
Table column	Values	Description	Influence
Project-ID	Integer	Unique identifier of the project	None
MemberId	Integer	PERNR of the employee participating in the project	None
Member-Name	String	ENAME (RSD EmplContrInfo) of the employee participating in the project	None

RSD SAPProj
Information on MWW projects. This table contains the project state, the attribute whose values we want to predict for task 2. As it is our target value, it's excluded from possible influencing factors.
As we're analysing the source data set for project state, it makes sense to determine the possible outcomes we want to predict in task 2. With 'live' and 'suspended' already marked as temporary, we have 4 more states as possible outcomes. Of these, only successful projects have a date for completion. Despite this, we regard the final states of a project as successful, failed and canceled, as it goes against common sense to consider a failed or cancelled project as ongoing. As for 'undefined', from visually inspecting the combined data of all three project tables, those 94 projects where fully prepared with members, start date and deadline, planned cost and revenues. But there is no date for completion, and no real cost and revenue. So the data might have not been reported, or the projects might have never even started. We could exclude them, because we don't know what it means. We could use it, as other projects by those project members could end up in limbo as well. We could build two models, and compare.
Other non-influencing factors beside state are the unique 'ID' and 'completed', which is only given for successful projects and thus would negatively impact the accuracy of the prediction.
It can be seen (appendix: 30) that the different business areas have a very similar rate of successful, failed and cancelled projects, as well as projects in general. We thus conclude that 'Business Area' does not influence the project success. The 'description' of the projects show a similar relation to project state (14). There is a bit more variation in the resulting states, but not significantly, so this attribute could be excluded from possible influencing factors or kept as a low-priority influence. The attributes 'deadline' and 'launched' are not usable on their own, but can be combined to determine the planned duration of projects. The planned duration lie between 60 and 152 days, with the majority of the projects having a duration between 170 and 122 days, so project state ratios for very high or very low duration are less dependable. The project states vary greatly (15, with a slight indication of longer projects having reduced success, but given the varying dependability of the data and the duration being planned, not actual, we

consider this attribute to be a possible but not deciding influence on project state.

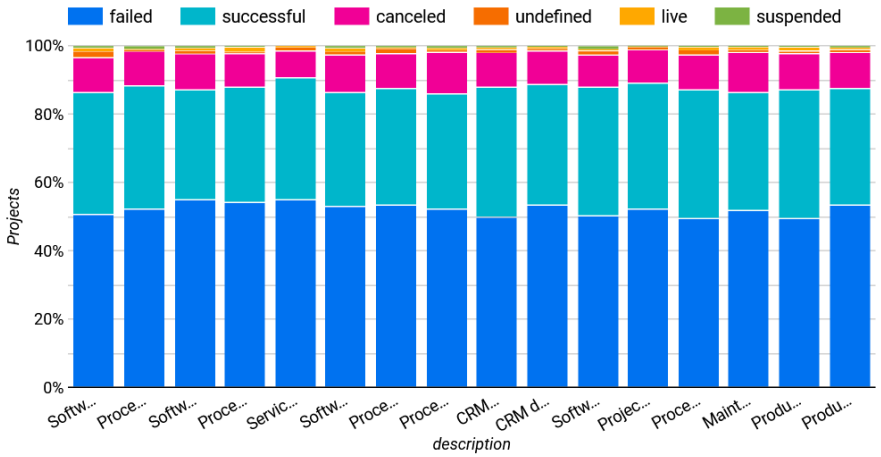


Figure 14: Project states per project type

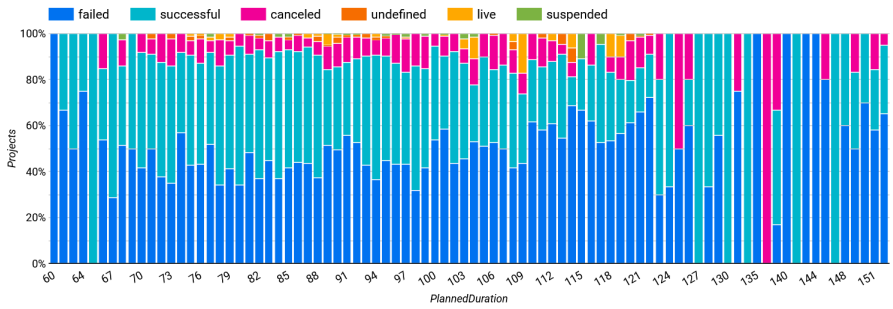


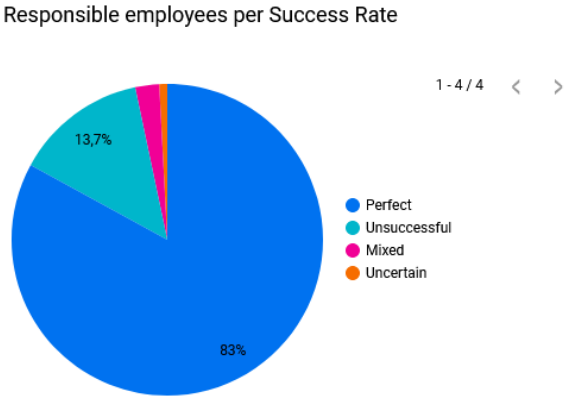
Figure 15: Project Duration to project states

The 'RESNR' and 'RESNA' have no inherent influence on project state, but they correspond to an employee responsible for and participating in the project. It thus seems sensible to look at the influence of the employee responsible for the project (ERP), as they most likely are project leaders and thus have a stronger influence than regular team members on project success.
In a first Query (appendix: 5.1.6), we determine the number of projects each employee participated in as the project responsible, as a regular team member and in total. Based on these numbers, we further calculate the rate of responsibility, as number of projects the employee was responsible for divided by total projects they participated in.
With the given data, we see that each employee has so far been either a regular project member or a project responsible for every project they participated in, which makes sense in that a quick check of the ERPs shows that all of them hold the position of Project Manager (but not all Project Managers had projects). This means the employees will not get promoted and be responsible for projects based on successful participation in projects nor is there a way to demote bad project leaders. Which simplifies our data analysis task, as we don't have to account for employees project success depending on their role within the project.

RSD SAPProj			
Table column	Values	Description	Influence
ID	Integer	Called 'Project-ID' in other tables.	None
deadline	date	Planned end of the project	Not as-is, Possibly
launched	date	Start of the project	Not as-is, Possibly
state	String	{'successful', 'failed', 'canceled', 'live', 'undefined', 'suspended'} with {3538, 5222, 1024, 74, 94, 48} projects each	-
completed	date	Actual project end. Only for 'successful' projects.(Ending later than planned does not immediately lead to a failure)	None
Business Area	String	{'MS', 'MPD', 'RSD', 'CCR'} with {2522, 2438, 2525, 2515}	None
description	String	One of 16 possible project descriptions, each belonging to one exactly one Business Area	Possibly
RESNR	Integer	PERNR of the employee responsible for the project	Not as-is, Strong
RESNA	String	ENAME of the employee responsible for the project	None

The 1930 ERPs of MWW have each been responsible for between 1-38 projects. As a first step, we need to determine the number of projects in each state for each ERP. With the view we generated from that query (appendix: SQL 1), we can then determine the success rate of the project responsible, by dividing the number of their projects of each state by the number of the projects there were responsible for in total (appendix: SQL 2). We can divide the ERPs into 4 groups: always successful, always unsuccessful, mixed results and no finished project yet (meaning only suspended or live projects). There are 1601 responsible employees with perfect project success score, 265 with only failed and/or cancelled projects, 48 with success rates between 0 and 1 and 16 who only have live or suspended projects. All the ERPs with mixed result however, have exactly one live project beside their otherwise completely successful projects (see appendix 31), so there is no group with confirmed mixed results. (On a site note: all unknown projects belong to unsuccessful ERPs with a number of failed projects (at least 9), so that state does not hinder prediction of project success).
We can see that the overwhelming number on project responsible has a flawless project result score. However, the 14,56% of responsible with a bad project success error were responsible for almost twice as many projects, leading to the overall bad project success of MWW. It should be noted, that the 'Projects' number in the table and lower tight chart refers to the number of project under those responsible, not the project success. Meaning, the 154 projects of 'Mixed' results are the previously established 48 live projects (1 per responsible) and 106 successful projects of those responsible, that have not finished all there projects but were only successful in previous projects.

	Success_Factor	Responsibles	Projects
1.	Perfect	1.601	3.432
2.	Unsuccessful	265	6.396
3.	Mixed	48	154
4.	Uncertain	16	18



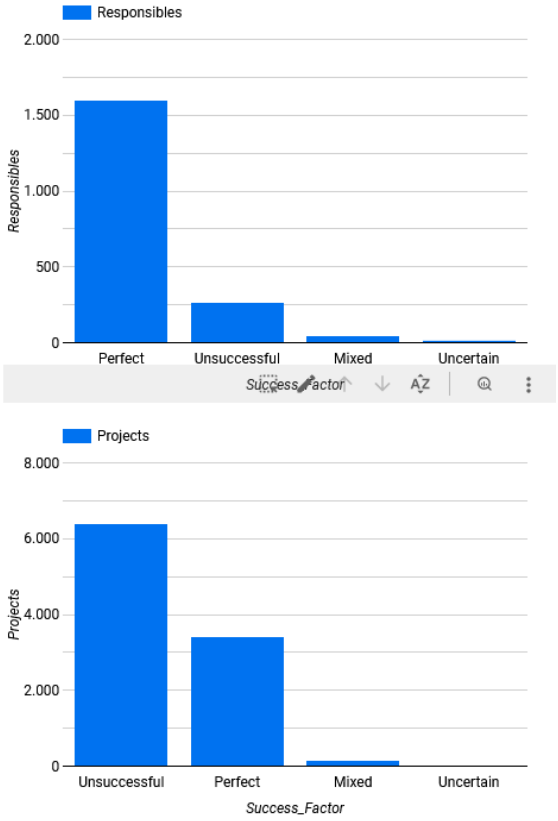


Figure 16: Responsible employees and project success

Now we need to extract the same information for regular project members, to compare them for differences (appendix: SQL 3).
For the 3513 project members with mixed results, the results are actually mixed, having successful, failed and/or cancelled projects (appendix: result of 4). There also comparatively a lot more employees with bad project results than there were ERPs with bad results, which might be connect to the unsuccessful ERPs having more projects and the regular employees participating in these projects, leading to their bas results.
But as we have previously seen, the project success of the responsible employee seems to never have changed so far, which indicates that the success of a project depends significantly on the responsible employee, with regular project members howing only a minor (if any) influence.

Because of the significant influence of responsible employees we determined, we re-examine and look more in-depth at the influence of their qualification on their success rate.
All employees responsible for projects (ERPs) are Project Managers contracted to RSD. All Project Managers are contracted to RSD in general, but not all of them have had projects. Position and Business Area are thus irrelevant to ERPs project success. All have an academic degree, but the exact degrees vary. If we look at the success of all ERPs per academic degree, we can see that a higher academic degree increase the likelihood of having perfect project success (see appendix: 32), as indicated by previous analysis (compare 10).

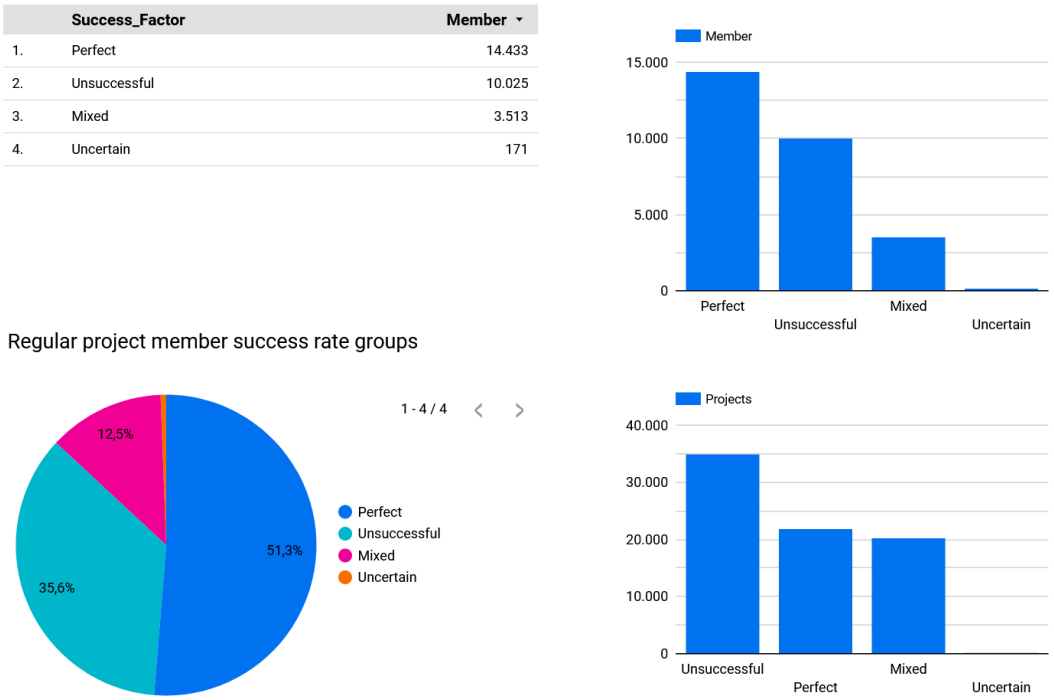


Figure 17: Regular project members and project success

In which education field the degree was received does not matter, as the success factor rate of all education fields is very similar (see appendix: 33).

2.3.3 External data: Institutions of higher education

Universities_faculty
The 1024 universities that were attended by the MWW employees with their best faculty. Best faculties are pretty evenly distributed.

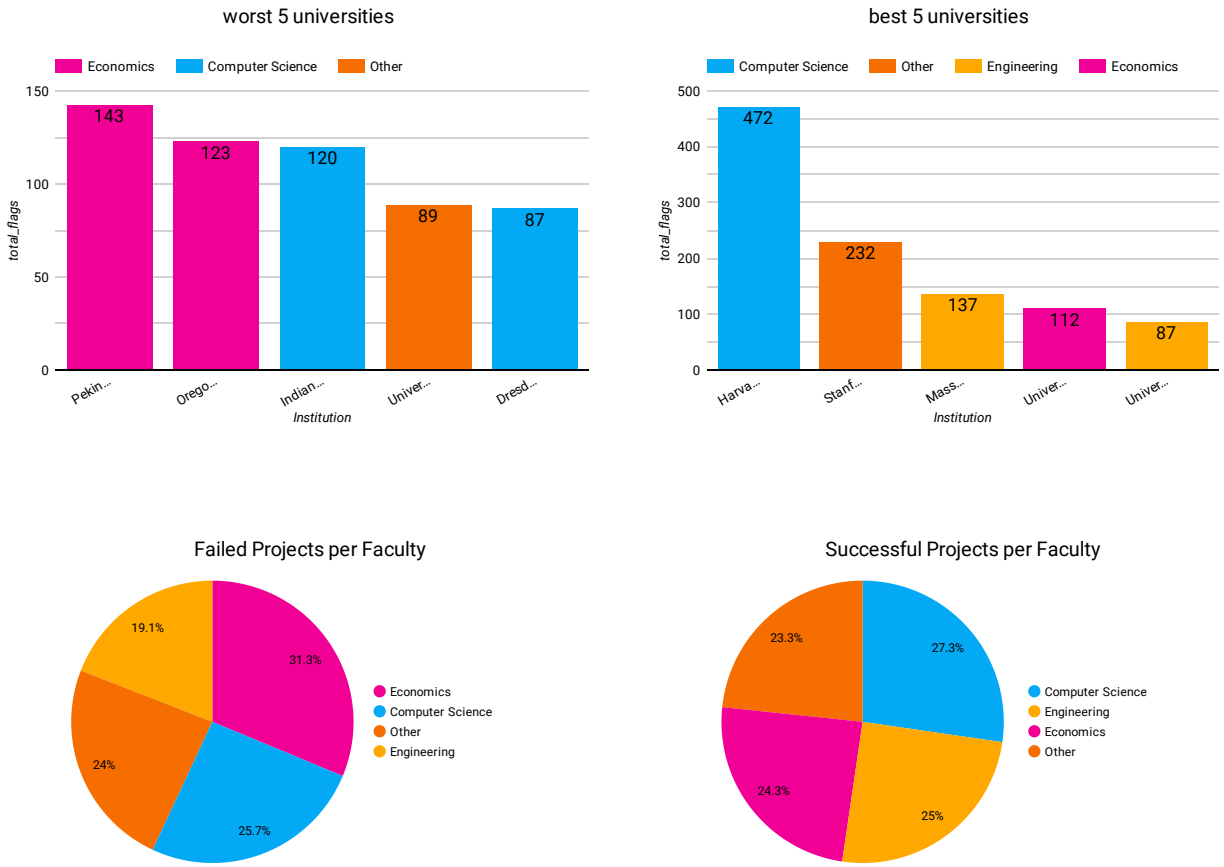
- institution: name of the institution
- best_faculty: best faculty of the institution, possible values are Other, Engineering, Economics, Computer Science, of 256, 246, 270, 252 institutions

Universities_ranking
Contains 2200 entries, the top 100 universities in 2019 and 2020, and the top 1000 for 2021 and 2022.

- world_rank:
- institution: name of the institution
- country: location of the university

- national_rank: rank within the country
- attributes the institutions are ranked in (globally)
  - quality_of_education
  - alumni_employment
  - quality_of_faculty
  - publications
  - influence
  - citations
  - broad_impact: not for 2019, 2020
  - patents
- score: cumulative result of the ranking attributes, with rank 1 having 100%
- year: year the ranking was determined

Qualification of employees
With the external data on the education institution employees attended, we can do a more detailed analysis of the influence of qualification of employees on project outcomes.
At first, we focused on a broader analysis as we considered that broader or ideally general or publicly available information which is then of course easier to access that could already allow estimating the project's outcome to some degree, would simplify the decision-making process for executives and therefor could allow saving of resources and time.
This allows us to reduce the potential time and resource allocation of data gathering processes to keep the findings within this report up-to-date to a minimum. Therefore, we firstly concentrated on a broader University level analysis, which can be seen on the first 2 graphs below.
There it can be seen that on the right side a ranking Graph is shown that presents the most data points of successful projects at a certain university, distributed by the faculties. So for example, the Harvard University providing the most data of projects that in the end have been successful, when the Project-Members came from the faculty of Computer Science which is shown by the color coding of the chunk. On the contrary, as the graph on the left side shows the results for the faculties of each university for failed projects, it can be seen that for example the university of Peking and within that institution especially the faculty of economics ranks the highest of failed projects within the provided dataset. This can potentially be a reasonable first indication that the selection of Project Members in general when considering someone within a university degree seems to benefit from filtering by set institution.
Now of course this is still a very abstract generalizing overview that for example doesn't take the employer's position into consideration within a Project-Group, but the underlying relation between successful projects and the certain universities seems eminent.

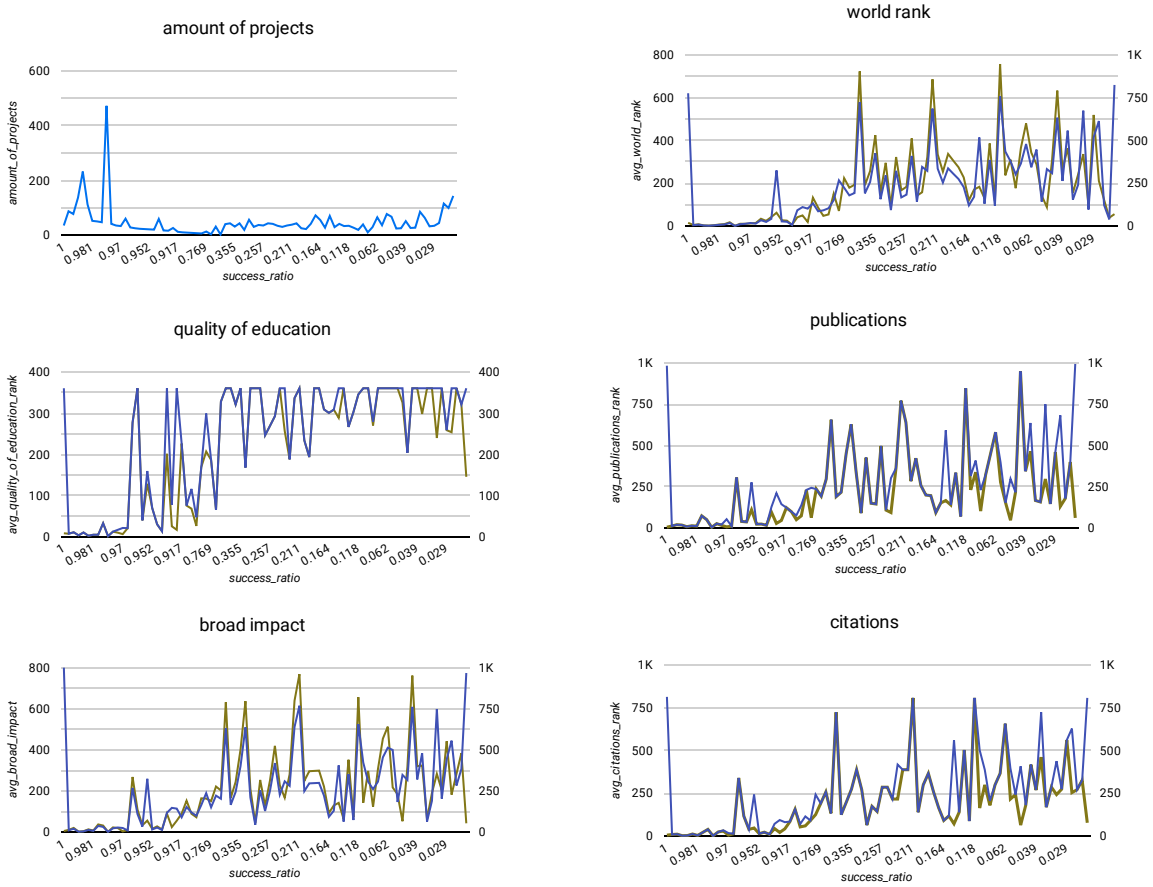


Universities and faculty performance Overview

Now in the bottom Graphs seen in the image above on the left side the ratio of the overall faculty, the Project-Member studied at to the project state of being a failed Project is shown. The graph on the right side shows the same overview but for successful Projects. It can be seen that it might be slightly more beneficial for the Project outcome to be successful if the corresponding Members studied at a technical department as the ratios of successful to failed seem to be slightly larger for both the Computer Science and Engineering faculties compared to the Economics faculty where it's the opposite.
The fourth faculty, being a set of unknown faculties combined, has almost no influence for the Project-outcome based on the graph above.This could of course be based on the fact that there are less distinct data points of some of the subgroups within this set, or that some of the subgroups impact is evened out by the others within the group, so it's hardly possible to allow some suggestions on this faculty subgroups in this context, and therefore we decided to label the impact as being insignificant here.
In general a distinctive conclusion based on the visualized relation between faculty and project outcome cant be drawn here, as even if there might be a slight advantages impact on choosing Members from technical faculties, the overall significance of this positive factor seems to be very insignificant as for both project-states, successful and failed, a large proportion of the involved

Project-Members can be found within each faculty.

In the following graphs, a more detailed view on the educational institution of employees within the provided data is presented. Each diagram shows a comparison of the relation between the columns from the university-tables, providing the detailed information about the employer's education for each institution of each employer being a project-Member to the project-success. As seen below these columns are the broad impact, world rank, quality of education, citations and publications. The provided values for these columns are all comparable with each other, as they are simply resembled as a score value ranking from 1 to last. The olive line in the graphs represents the minimum value of the chosen metric or column, while the blue line presents the maximum value. So in general can be stated that if the both lines, olive and blue are locally separated from one of each other the information gain on the impact for that column at that range of the value of the project-success-rate is dubious and therefore not abstractable.



Comparison of University ranks and project success rates

This is especially the case for broad-impact and to some degree for all the others as well, quality of education, publications, citations and world-rank.
So there isn't properly a single causal relation between the project-success and one of these score-values, hence still there might be information of an underlying impact abusable in f.e.

the edge cases of the graphs.Regarding this it's interesting to see that the overall top scoring universities, being always the same also seem to generally provide excellent project-success-rates no matter what metric is chosen. This just underlines in general that if the selection-process of Project-Members takes a university-filtering approach into consideration, these universities would be the most likely the most promising considering project-success-rates.
It's also interesting to see that the quality of education with its very frequent changes of low-point-values and high-point-values after first few data points seems to be compared to the others less persistent in its shape and therefore most likely less visible for a prediction on whenever a project might be successful or not. Now the other graphs suffer partially from this state as well, but the general trend of success-rate decreasing on increasing of the corresponding rank-values can still be seen for the whole graphs. In general, it can be concluded that the reliable information gathered from these employer-education based columns consisting of university specific data allows only for a consideration of the very best rated universities to have significantly higher chances of providing something to the project which results in it being more likely successful.
The best five ranking universities among all categories are University of Oxford, University of Cambridge, Massachusetts Institute of Technology, Stanford University, Harvard University.

2.4 Analysis Outcome

At the beginning of our analysis, we defined three categories of possible influencing factors to analyse. These areas where employees, financing and environment.
We come to the conclusion that employees are the biggest influence on project success, with the attributes making a significant impact being who the responsible employees is and the degree employees (both responsible and regular team member) received. Other attributes of employees, like their personal data plays almost no role in deciding the outcome of a project.
For financing, there is an indication of project budget and duration being related to project outcome, but it seems to be a weaker influence than employees.
Finally, for environment, the year of a project plays an insignificant role in determining the project outcome, in addition to lack of data for recent projects and other factors like project business area having no discernable influence on the project outcome. Thus, it should be possible to exclude this category from the attributes used for project state prediction.
Apart from the importance of attribute for training a prediction model, we need to consider the following circumstances for selecting the training data:

- Institution Listing: It was observed that every employee has an institution listed, regardless of whether they graduated or not. This suggests that when evaluating the success of projects involving universities, it may be necessary to distinguish between graduates and non-graduates.

- Undefined outcomes: Project state being undefined needs to handled during training for state prediction, both excluding the data and including it are possible, as it could be a data entry error as well as an internal problem of project teams not reporting, which they could repeat for new projects.

- Project Data Time-frame: The project data spans from 2013 to 2022. However, there is an anomaly with a single cancelled project recorded in 1974 which may be due to data entry errors (deadline and project duration of all projects indicate the project is from 2014) or inconsistencies.

- Incomplete Data for 2022: The dataset indicates that there are only nine projects recorded for the year 2022, and all of them are still active. This limited number of projects in 2022 suggests that the data for that year is incomplete. It is likely that the entire dataset was extracted in 2022, resulting in a lack of comprehensive project information for that year.
- University ranking: We have the ranking for 4 years, with the first two years showing the top 100 and the latter years the top 1000 universties, along with values for the 'broad_impact'.

3 Part 2: Predicting the project outcome

Based on the insights from the data exploration of task 1, use data analytics and machine learning methods to predict the outcome of MPD projects. Use Kubeflow to orchestrate the Data Science lifecycle (including e.g., data preparation, model training, deployment, and monitoring). Apply the model on "suspended" and "live" projects from the MPD unit.

3.1 Determine optimal Model

In the process of finding the most suitable model for predicting outcomes using BigQueryML, we first analyzed the success rates of employees during data preparation. Now, our goal is to forecast the success rate of a new candidate by leveraging their information. To achieve this, we need to train the model using both historical project data and current employee data. By employing this trained model, we can then predict the success rate for each new candidate.

BigQueryML offers a range of algorithms that can be utilized for numerical value prediction. In our case, we employed regression models and assessed their respective errors. Ultimately, we selected the model with the lowest error rate. Here are the steps we followed:

1. We created a BigQuery view of the training data to make it accessible for the models.
2. Utilizing BigQueryML, we constructed several machine learning models.
3. Finally, we identified the model with the least mean absolute error as our finalized choice.

3.1.1 Generating BigQuery Training Data View

To create a BigQuery view for training data, we have compiled an integrated dataset consisting of various employee information sources, including RSD_EmplInfo, RSD_EmplContrInfo, RSD_EmplQualifications, RSD_SAPProj, and RSD_ProjMembers. By merging these datasets, we have derived the necessary features to train our machine learning model.Since our dataset encompasses multiple Business Units, we are specifically filtering to isolate only the relevant data for the "MPD" Business Unit. The schema of the dataset has been formed based on the features selected to train the model as follows:

1. Gender
2. Age
3. Nationality
4. Marital_Status
5. Business_Area
6. Payroll_Area
7. Position_of_Employee
8. Education_Field
9. Education_Institution

These features have been selected to allow the prediction model to be based on general individual employer-information like age and gender, as this seems reasonable for the later-on performed HR-Recommendation because these feature-values will be easily accessible. We also tried to integrate the most promising analysis results from the data exploration and therefore added Education Institution and Position of Employee.

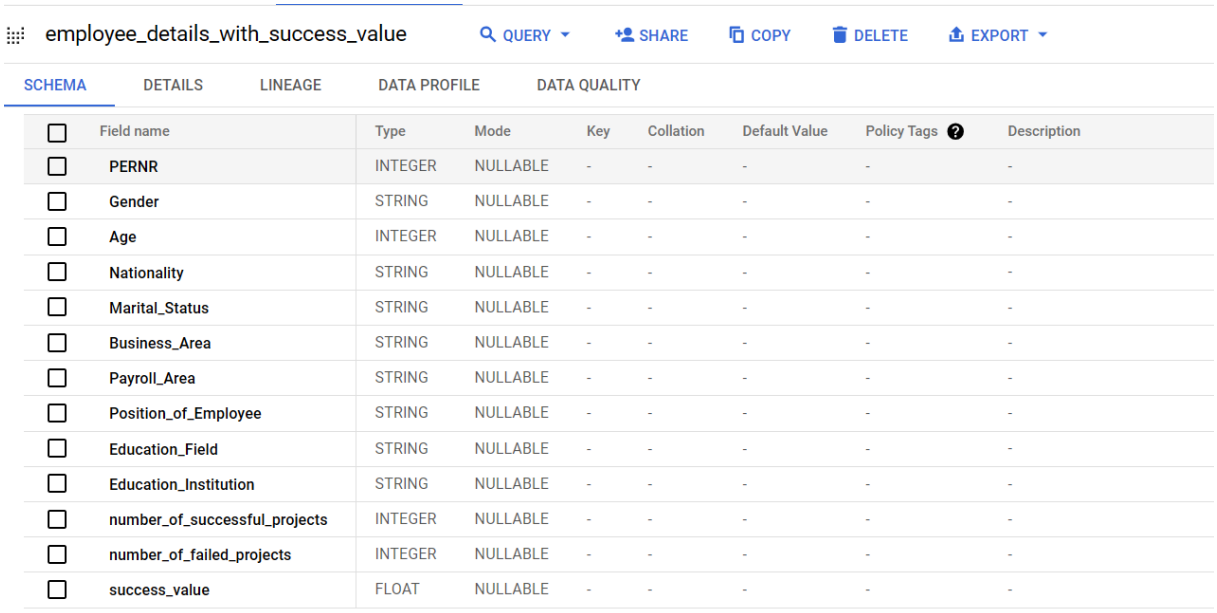
The cumulative dataset has been created by joining necessary tables:

```sql
CREATE OR REPLACE VIEW `vlba-rsd-grp3.RSD_department.employee_details_with_success_value` AS
SELECT
    emp_info.PERNR,
    emp_info.Gender,
    DATE_DIFF(CURRENT_DATE(), DATE(emp_info.`Birth date`), YEAR) AS Age,
    emp_info.Nationality,
    emp_info.`Marital Status Key` AS Marital_Status,
    emp_contract_info.`Business Area` AS Business_Area,
    emp_contract_info.`Payroll Area` AS Payroll_Area,
    emp_contract_info.`Position` AS Position_of_Employee,
    emp_qualifications.`EducationField` AS Education_Field,
    emp_qualifications.`Institution` AS Education_Institution,
    COALESCE(emp_success_projects.number_of_successful_projects, 0) AS number_of_successful_projects,
    COALESCE(emp_failed_projects.number_of_failed_projects, 0) AS number_of_failed_projects,
    CASE
        WHEN COALESCE(emp_success_projects.number_of_successful_projects, 0) + COALESCE(emp_failed_projects.number_of_failed_projects, 0) = 0
        THEN 0.0
        ELSE ROUND(
            COALESCE(emp_success_projects.number_of_successful_projects, 0) /
            (COALESCE(emp_success_projects.number_of_successful_projects, 0) + COALESCE(emp_failed_projects.number_of_failed_projects, 0)),
            2
        )
    END AS success_value
FROM
    `vlba-rsd-grp3.RSD_department.RSD_EmplInfo` emp_info
JOIN
    `vlba-rsd-grp3.RSD_department.RSD_EmplContrInfo` emp_contract_info
ON
    emp_info.PERNR = emp_contract_info.PERNR
JOIN
    `vlba-rsd-grp3.RSD_department.RSD_EmplQualifications` emp_qualifications
ON
    emp_info.PERNR = emp_qualifications.PERNR
```

```sql
LEFT JOIN
    (SELECT
        project_members.MemberId,
        COUNT(sapproj.ID) AS number_of_successful_projects
    FROM
        `vlba-rsd-grp3.RSD_department.RSD_SAPProj` sapproj
    JOIN
        `vlba-rsd-grp3.RSD_department.RSD_ProjMembers` project_members
    ON
        sapproj.ID = project_members.Project_ID
    WHERE
        sapproj.state = 'successful'
    GROUP BY
        project_members.MemberId
    ) emp_success_projects
ON
    emp_info.PERNR = emp_success_projects.MemberId
```

```
LEFT JOIN
    (SELECT
        project_members.MemberId,
        COUNT(sapproj.ID) AS number_of_failed_projects
    FROM
        `vlba-rsd-grp3.RSD_department.RSD_SAPProj` sapproj
    JOIN
        `vlba-rsd-grp3.RSD_department.RSD_ProjMembers` project_members
    ON
        sapproj.ID = project_members.Project_ID
    WHERE
        sapproj.state = 'failed'
    GROUP BY
        project_members.MemberId
    ) emp_failed_projects
ON
    emp_info.PERNR = emp_failed_projects.MemberId
WHERE
    emp_contract_info.`Business Area` = 'MPD'
ORDER BY
    emp_info.PERNR;
```

The features from the cumulative dataset will be trained to predict the target value of the Machine Learning model that we intend to build. The schema of the dataset is :



Field name	Type	Mode	Key	Collation	Default Value	Policy Tags	Description
PERNR	INTEGER	NULLABLE	-	-	-	-	-
Gender	STRING	NULLABLE	-	-	-	-	-
Age	INTEGER	NULLABLE	-	-	-	-	-
Nationality	STRING	NULLABLE	-	-	-	-	-
Marital_Status	STRING	NULLABLE	-	-	-	-	-
Business_Area	STRING	NULLABLE	-	-	-	-	-
Payroll_Area	STRING	NULLABLE	-	-	-	-	-
Position_of_Employee	STRING	NULLABLE	-	-	-	-	-
Education_Field	STRING	NULLABLE	-	-	-	-	-
Education_Institution	STRING	NULLABLE	-	-	-	-	-
number_of_successful_projects	INTEGER	NULLABLE	-	-	-	-	-
number_of_failed_projects	INTEGER	NULLABLE	-	-	-	-	-
success_value	FLOAT	NULLABLE	-	-	-	-	-

3.1.2 Building BigQueryML Models for Prediction

BigQuery ML empowers users to develop and execute machine learning models directly within BigQuery using SQL queries. This approach aims to democratize machine learning by enabling SQL practitioners to leverage their existing skills and tools. By eliminating the need for data movement, the development speed is enhanced.

Here we have experimented with two different models:

1. DNN_REGRESSOR
2. LINEAR_REGRESSION

1. DNN_REGRESSOR

The 'DNN_REGRESSOR' model is a type of Deep Neural Network Regression model that predicts the success value of employees based on selected features. This model leverages the power of deep learning, using multiple layers of neurons to capture complex and nonlinear relationships within the data. By adjusting hyperparameters such as the number of layers, the number of neurons per layer, activation functions, and dropout rates, this model aims to achieve high prediction performance. This makes it especially effective for uncovering intricate patterns and interactions in the data, which are crucial for accurately predicting success rates.

The query of the model is:

```
CREATE OR REPLACE MODEL `vlba-rsd-grp3.RSD_department.DNN_predict_success_score`
OPTIONS(
    MODEL_TYPE='DNN_REGRESSOR',
    ACTIVATION_FN='RELU',
    BATCH_SIZE=60,
    DROPOUT=0.1,
    EARLY_STOP=TRUE,
    HIDDEN_UNITS=[68, 8],
    INPUT_LABEL_COLS=['success_value'],
    LEARN_RATE=0.1,
    MAX_ITERATIONS=100,
    OPTIMIZER='ADAGRAD',
    DATA_SPLIT_METHOD='RANDOM',
    DATA_SPLIT_EVAL_FRACTION=0.2
)
AS SELECT * EXCEPT (PERNR) FROM `vlba-rsd-grp3.RSD_department.employee_details_with_success_value`;
```

Evaluation Metric:

Mean absolute error	0.5531
Mean squared error	0.3302
Mean squared log error	0.1853
Median absolute error	0.6567
R squared	-0.6651

2. LINEAR_REGRESSION

The 'LINEAR_REG' model, predicts the success value of employees using linear regression. This model establishes a linear relationship between input features and the target variable, making it straightforward to interpret and efficient to train. It is particularly useful for understanding the direct influence of each feature on success rates, providing a simple yet effective prediction approach for scenarios where relationships are assumed to be linear.

The query of the model is:

```
CREATE OR REPLACE MODEL `vlba-rsd-grp3.RSD_department.LR_predict_success_percentage`
OPTIONS(
  MODEL_TYPE='LINEAR_REG',
  INPUT_LABEL_COLS = ['success_value'],
  L2_REG=0.1,
  L1_REG=0.1,
  DATA_SPLIT_EVAL_FRACTION=0.1,
  DATA_SPLIT_METHOD="RANDOM",
  MAX_ITERATIONS = 50
) AS
SELECT * EXCEPT (PERNR)
FROM `vlba-rsd-grp3.RSD_department.employee_details_with_success_value`;
```

Training Metrics:

Evaluation Metrics:

Mean absolute error	0.2053
Mean squared error	0.072
Mean squared log error	0.0318
Median absolute error	0.1713
R squared	0.6266

COMPARISON OF THE MODELS

Based on the evaluation metrics, we have chosen Linear Regression (LR) over the Deep Neural Network (DNN) Regressor to build the pipeline in GCP due to its superior performance and practical advantages. The LR model demonstrated significantly lower Mean Absolute Error (0.2053 vs. 0.5531) and Mean Squared Error (0.072 vs. 0.3302), indicating it provides more accurate and reliable predictions.

LINEAR REGRESSION MODEL VS DNN REGRESSION MODEL

Mean absolute error	0.2053
Mean squared error	0.072
Mean squared log error	0.0318
Median absolute error	0.1713
R squared	0.6266

Mean absolute error	0.5531
Mean squared error	0.3302
Mean squared log error	0.1853
Median absolute error	0.6567
R squared	-0.6651

Additionally, the LR model's R-squared value of 0.6266 reflects a decent fit to the data, whereas the DNN Regressor's negative R-squared (-0.6651) suggests it fails to capture the underlying data patterns effectively. Furthermore, the simplicity and interpretability of LR make it easier to deploy and maintain, especially in a GCP environment where computational efficiency and resource utilization are critical. Considering the complexity and poorer performance of the DNN model , the LR model offers a more robust and cost-effective solution, ensuring better scalability and reduced risk of overfitting.
The Linear Regression has fewer hyperparameters to tune, simplifying the process of model management and updates. On the other hand, DNNs require extensive tuning of multiple hyperparameters, such as learning rates, batch sizes, and network architectures, which can be time-consuming.

3.1.3 Selecting the Final BigQueryML Model for Prediction

Based on the evaluation metrics, it is evident that the Linear Regression model exhibits the lowest error rate. Consequently, we have opted to employ the Linear Regression model for training in the Vertex AI platform as part of our machine learning pipeline.

3.2 Building ML Pipeline with Vertex AI

Vertex AI is a comprehensive machine learning (ML) platform that empowers users to train, deploy, and customize ML models and AI applications. It also offers the capability to customize large language models (LLMs) for enhanced AI-powered applications. By integrating data engineering, data science, and ML engineering workflows, Vertex AI provides a unified toolset for collaborative development and scalable deployment on Google Cloud. In our ML pipeline, we utilized both the Vertex AI Python SDK and Kubeflow to streamline the process.

When running a pipeline, certain resources, including the metadata store utilized by Vertex ML Metadata, are generated within your Google Cloud project for the initial time [4].

3.2.1 Configure your project:

The project is setup using th GCP documentation:[3]

i. Run the following command to create a service account.

gcloud iam service-accounts create service-acc-grp-3 –description="Service-account" –display-name="service-acc-grp-3" –project=vlba-rsd-grp3

ii. Grant your service account access to Vertex AI.

gcloud projects add-iam-policy-binding vlba-rsd-grp3 –member="serviceAccount:service-acc-grp-3@vlba-rsd-grp3.iam.gserviceaccount.com" –role="roles/aiplatform.user"

iii. To use Vertex AI Pipelines to run pipelines with this service account, run the following command to grant your user account the **roles/iam.serviceAccountUser** role for your service account.

gcloud iam service-accounts add-iam-policy-binding service-acc-grp-3@vlba-rsd-grp3.iam.gserviceaccount.com –member="user:dinesh.gopi.hema@gmail.com" –role="roles/iam.serviceAccountUser"

iv. Run the following command to grant your service account the **Vertex AI User** role.

gcloud projects add-iam-policy-binding vlba-rsd-grp3 –member="serviceAccount:service-acc-grp-3@vlba-rsd-grp3.iam.gserviceaccount.com" –role="roles/vertexai.user"

v. Run the following command to create a Cloud Storage bucket in the region that you want to run your pipelines in.

gsutil mb -p vlba-rsd-grp3 -l europe-west1 gs://pipeline-vlba-grp-3

vi. Run the following commands to grant your service account access to read and write pipeline artifacts in the bucket that you created in the previous step.

gsutil iam ch serviceAccount:service-acc-grp-3@vlba-rsd-grp3.iam.gserviceaccount.com:roles/storage.objectCreator,objectViewer gs://pipeline-vlba-grp-3

3.2.2 Set up authentication

To set up authentication, you must create a service account key, and set an environment variable for the path to the service account key. [4]

1. Create a service account key for authentication:

- In the Google Cloud console, click the email address for the service account that you created.
- Click **Keys**.
- Click **Add key**, then **Create new key**.
- Click **Create**. A JSON key file is downloaded to your computer.
- Click **Close**.

2. Grant your new service account access to the service account that you use to run pipelines.

- Click arrow_back to return to the list of service accounts.
- Click the name of the service account that you use to run pipelines. The **Service account details** page appears. The Compute Engine default service account is named like the following: `234211414315-compute@developer.gserviceaccount.com`
- Click the **Permissions** tab.
- Click **Grant access**. The **Add principals** panel appears.
- In the **New principals** box, enter the email address for the service account you created in a previous step.
- In the **Role** drop-down list, select **Service accounts ¿ Service account user**.
- Click **Save**

3. Set the environment variable GOOGLE_APPLICATION_CREDENTIALS to the path of the JSON file that contains your service account key.
$env:GOOGLE_APPLICATION_CREDENTIALS="[PATH]"
Replace [PATH] with the path of the JSON file that contains your service account key.

3.2.3 Define your workflow using Kubeflow Pipelines DSL package

The `kfp.dsl` package contains the domain-specific language (DSL) that you can use to define and interact with pipelines and components.[5]

Kubeflow pipeline components are factory functions that create pipeline steps. Each component describes the inputs, outputs, and implementation of the component. For example, in the code sample below, `ds_op` is a component.

```
import kfp
from google_cloud_pipeline_components.v1.bigquery import (BigqueryQueryJobOp)
from google_cloud_pipeline_components.v1.automl.training_job import AutoMLTabularTrainingJobRunOp
from google_cloud_pipeline_components.v1.dataset import TabularDatasetCreateOp
from google_cloud_pipeline_components.v1.endpoint import EndpointCreateOp , ModelDeployOp

project_id = "vlba-rsd-grp3"
pipeline_root_path = "gs://pipeline-vlba-grp-3"

# [START aiplatform_sdk_create_and_import_dataset_tabular_bigquery_sample]
def create_and_import_dataset_tabular_bigquery_sample(
    display_name: str,
    project: str,
    bigquery_source: str,
):
    ds_op = TabularDatasetCreateOp(
        display_name=display_name,
        bq_source=bigquery_source,
        project=project,
    )
    print(ds_op.outputs['dataset'])
    return ds_op
```

```python
# Define the workflow of the pipeline.
@kfp.dsl.pipeline(
    name="pipeline-vlba-grp-3",
    pipeline_root=pipeline_root_path)
def pipeline(project_id: str):
    # The first step of your workflow is a dataset generator.
    ds_op = create_and_import_dataset_tabular_bigquery_sample("enployee_success", project_id,
            "bq://vlba-rsd-grp3.RSD_department.employee_details_with_project_outcome_score")
    # The second step is a model training component. It takes the dataset
    # outputted from the first step, supplies it as an input argument to the
    # component
    training_job_run_op = AutoMLTabularTrainingJobRunOp(
        project=project_id,
        display_name="regression_model_training",
        optimization_prediction_type="regression",
        dataset=ds_op.outputs["dataset"],
        model_display_name="Linear_Regression_model",
        target_column="success_value",
        training_fraction_split=0.6,
        validation_fraction_split=0.2,
        test_fraction_split=0.2,
        budget_milli_node_hours=2000,
    )

    # The third and fourth step are for deploying the model.
    create_endpoint_op = EndpointCreateOp(
        project=project_id,
        display_name = "Success_regression_endpoint",
    )

    model_deploy_op = ModelDeployOp(
        model=training_job_run_op.outputs["model"],
        endpoint=create_endpoint_op.outputs['endpoint'],
        dedicated_resources_machine_type="n1-standard-16",
        dedicated_resources_min_replica_count=1,
        dedicated_resources_max_replica_count=1,
    )
```

3.2.4 Compile the pipeline into a YAML file

After the workflow of the pipeline is defined, we can proceed to compile the pipeline into YAML format[6]. The YAML file includes all the information for executing the pipeline on Vertex AI Pipelines.

```python
from kfp import compiler
compiler.Compiler().compile(
    pipeline_func=pipeline,
    package_path='pipeline-vlba-grp-3.yaml'
)
```

3.2.5 Submit the pipeline run

After the workflow of pipeline is compiled into the YAML format, we can use the Vertex AI Python client to submit and run our pipeline.[7]

```python
import google.cloud.aiplatform as aip

project_id = "vlba-rsd-grp3"
PROJECT_REGION = "europe-west1"
pipeline_root_path = "gs://pipeline-vlba-grp-3"

# Before initializing, make sure to set the GOOGLE_APPLICATION_CREDENTIALS
# environment variable to the path of your service account.
aip.init(
    project=project_id,
    location=PROJECT_REGION,
)

# Prepare the pipeline job
job = aip.PipelineJob(
    display_name="Linear-regression-employee-success-value-training",
    template_path="pipeline-vlba-grp-3.yaml",
    pipeline_root=pipeline_root_path,
    parameter_values={
        'project_id': project_id
    }
)

job.submit()
```

Once the pipeline is deployed using the provided code in the Vertex AI pipeline, the necessary services are initialized, and the model begins its learning process. After successful deployment, the pipeline is as below:

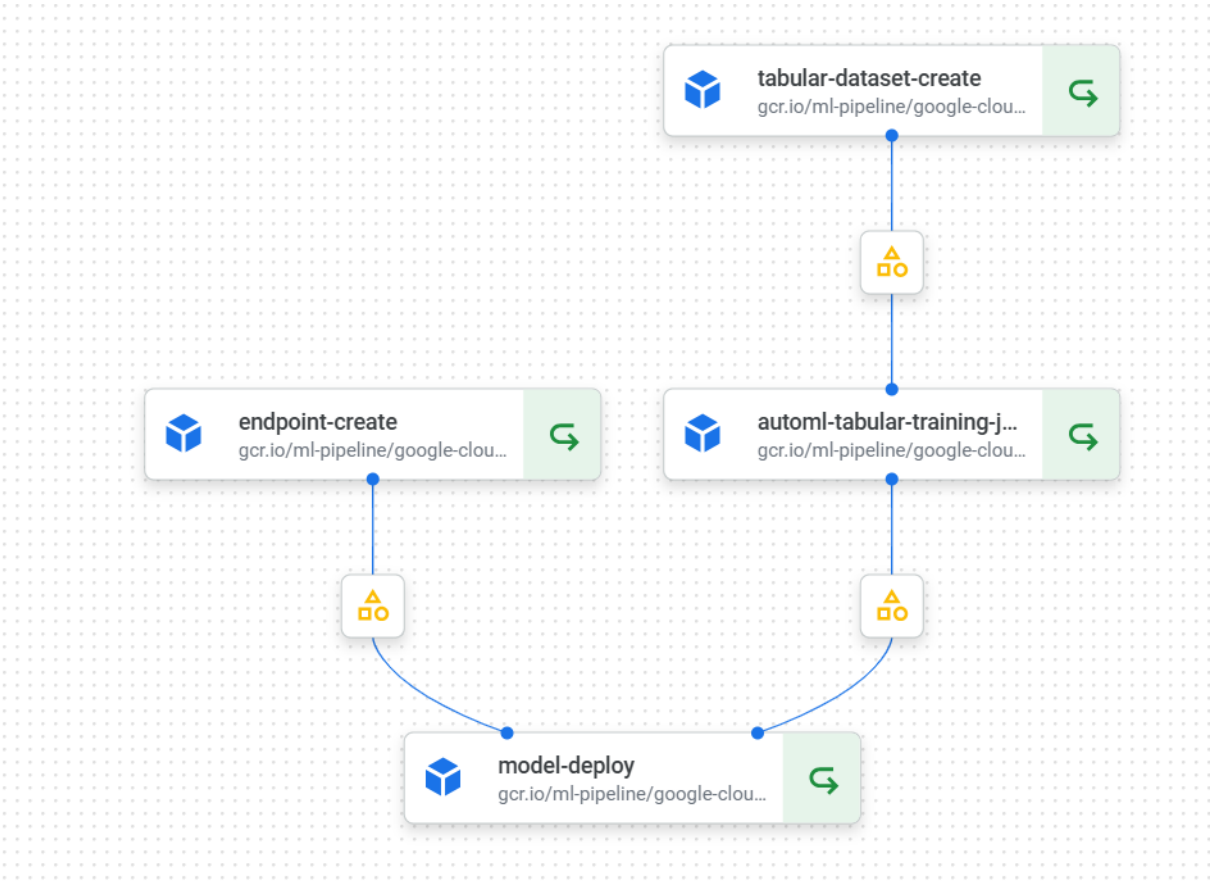


Figure 18: Deployed pipeline in Vertex AI using Kubeflow SDK

4 Part 3: Optimising talent recruitment and project oucomes

3. Summarize your findings and recommendations (e.g., recommended partner universities) for achieving better access to promising talents to improve project outcomes in the form of a dashboard and/or report.

4.1 Introduction

As we described in the first section above, we analyzed the provided datasets and each of the columns individually and if feasible combined to allow us to get the information of this column's impact on the success-rate of a project individually or within a set of other columns. In the second section, we used various ML-techniques to allow a prediction to be performed based on our selected features from the dataset. Then we analyzed the chosen final model about its insights, on why certain predictions are performed. For this, we used simple feature-importance-analysis and ranked them accordingly. This was done as we concluded that the resulting top-ranked features would be promising candidates for an HR-Recruitment filtering-process.

4.2 Task Conclusion

The data analysis and exploration process revealed several key findings that impact project success and have implications for improving the HR hiring process for future projects.

1. Impact of Project Size and Budget: Larger projects with higher budgets are more likely to fail. This highlights the need to improve project performance, as the failure of such projects can result in significant resource wastage.

2. Importance of Project Responsibilities: The project responsible, compared to other project members, has a significantly greater impact on project success. This suggests that selecting the right person for this role is crucial for project outcomes.

3. Educational Background: The educational institution where an individual was educated has a substantial impact on project success. Considering the data points, it is evident that project responsibles, who all have a university degree, greatly influence project outcomes. Therefore, filtering based on the educational institution becomes crucial.

4. Faculty-level Analysis: A broader investigation at the faculty level shows that project members from economics faculties tend to have lower overall performance compared to graduates from engineering or computer science faculties. This indicates the importance of considering the faculty when filtering project members based on their educational background.

5. Performance Differences Among Universities: Some universities consistently outperform others in terms of project success rates. This aligns with their rankings and reputation for providing high-quality education. Graduates from well-ranked universities, such as University of Oxford, University of Cambridge, Massachusetts Institute of Technology, Stanford University, and Harvard University, have a higher chance of positively impacting project outcomes.

Based on these findings, it is recommended, especially for project responsible roles, to prioritize hiring project members who have graduated from the aforementioned universities. This can increase the likelihood of project success by leveraging the knowledge and skills gained from these prestigious institutions.

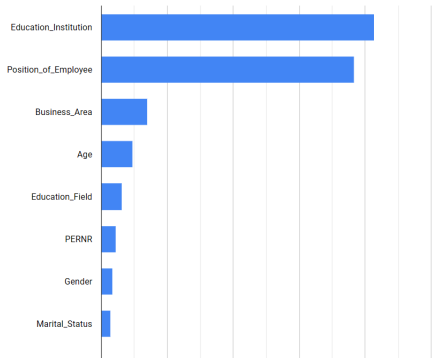


Figure 19: Feature Importance from the trained Linear regression model

Additionally, as can be seen in the diagram above, a feature analysis on the resulting model of task2 showed similar results compared to the exploration task. As it can be seen that the institution and position of project-members provided the most usable information for our model and this of course matches the outcome of the data analysis performed before, which therefore underlines the importance of set columns or features. So it can be concluded that the project-position and the educational background, here the university of graduation for the employee, is of critical importance to increase the chance of a project being successful or not.

4.3 Successful universities

Given the project outcomes of responsible employees, we can aggregate the projects lead by graduates of each university, to see which universities have a high success rate for a significant number of graduates.
The universities with the highest number of projects and thus reliability and very good success rates are Harvard University, Stanford University, MIT, University of Cambridge and University of Oxford. The graduates of Peking University, Oregon Health & Science University and Indiana University have reliably bad project outcome measured by project, thus are not desirable partner universities.34
The success rates can also be mapped to the number of graduates however, in which case the best universities don't change, but the worst have a lower reliability, because a few graduates lead many bad projects. 35

4.4 Business approach recommendation

Based on the accomplished gathering of insight information of impacting columns/features for improving the chances of a new SAP-project for this company being successful, we propose the following recruitment strategy for project members shown in the image below.

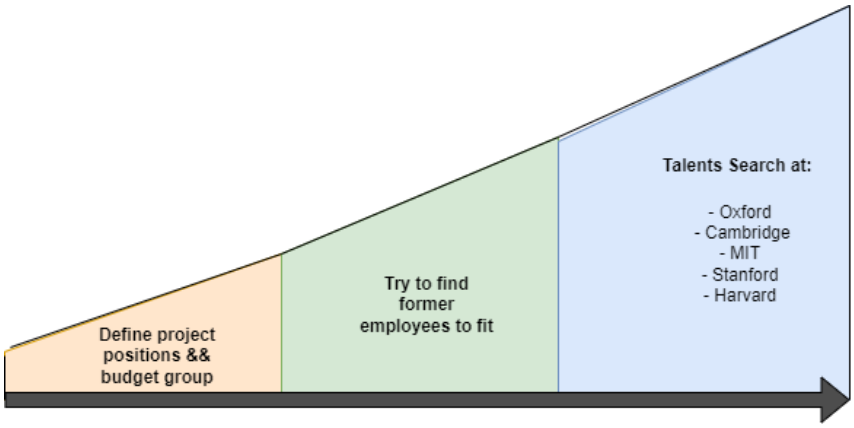


Figure 20: Recruitment Strategy

Proposed Step-by-Step Approach for Recruiting Based on Project Demands:

1. Consider Project Budget: Recognize that larger budget projects are more likely to face higher risk of failure. Therefore, prioritize a well-thought-out and well-performed talent selection process for these projects. For smaller budget projects, initial filtering may be sufficient due to lower demands.

2. Focus on Project Responsibilities: Emphasize finding the best fit for the project responsible role, as our analysis indicates its significant impact on project outcomes. Allocate more resources and attention to identifying candidates who possess the skills and expertise required for this critical role.

3. Consider Former Project Members: Leverage the experience and expertise of former project members who have a track record of successful projects. This not only reduces the recruitment process but also increases the chances of project success by relying on individuals who have already demonstrated their capabilities.

4. Explore Top-Ranked Universities: If the position remains unfilled, prioritize recruitment from top-ranked universities, such as University of Oxford, University of Cambridge, Massachusetts Institute of Technology, Stanford University, and Harvard University. These institutions have shown a clear positive impact on project success rates based on our data analysis.

5. Establish Partnerships and Hiring Events: Consider establishing partnerships with these top-ranked universities or organizing hiring events specifically targeted towards graduates from these institutions. This strategic approach provides the company with a competitive advantage in attracting the most promising talents from these renowned universities.

By following this step-by-step approach, recruiters can enhance the hiring process by focusing efforts on critical roles, leveraging the experience of former project members, and targeting candidates from universities with a proven positive impact on project success. This approach increases the likelihood of selecting candidates who are well-suited for the demands of the project and can contribute to its successful outcome.

4.5 Final Conclusion

Finally, we want to clarify that because the here performed data analysis and prediction procedure is based on an older dataset (2020-2023) with partly very low density of its provided values for some data-tuples, like for example for certain universities or project-states only a very small set of values is available, we suggest further investigations on updated or additional datasets to be performed to verify the here proposed conclusion even for future uses and potentially extend the usability even further by including additional findings.

References

[1] "Gcp integration with sap dwc." https://applexus.medium.com/google-bigquery-integration-with-sap -data-warehouse-cloud-2b95a62ab771.

[2] "Gcp documentation." https://cloud.google.com/docs.

[3] "Configure a service account with granular permissions." https://cloud.google.com/vertex-ai/docs/pipelines/configure-projectservice-account.

[4] "Gcp service account authentication." https://cloud.google.com/vertex-ai/docs/pipelines/build-pipelinestarted.

[5] "kubeflow pipeline workflow." https://cloud.google.com/vertex-ai/docs/pipelines/build-pipelinedefine$_{y}our_{w}orkflow_{u}sing_{k}ubeflow_{p}ipelines_{d}sl_{p}ackage$.

[6] "Compile pipeline workflow to yaml file." https://cloud.google.com/vertex-ai/docs/pipelines/build-pipelinecompile$_{y}our_{p}ipeline_{i}nto_{a}yaml_{f}ile$.

[7] "Submit pipeline workflow to gcp." https://cloud.google.com/vertex-ai/docs/pipelines /build-pipelinesubmit$_{y}our_{p}ipeline_{r}un$.

5 Appendix

5.1 Supplementary Material for task 1

An alternative view of figure 13 10, showing the doctoral and master degree holders have a much larger proportion of successful projects compared to the other states, but overall harder to read.

5.1.1 RSD EmplContrInfo

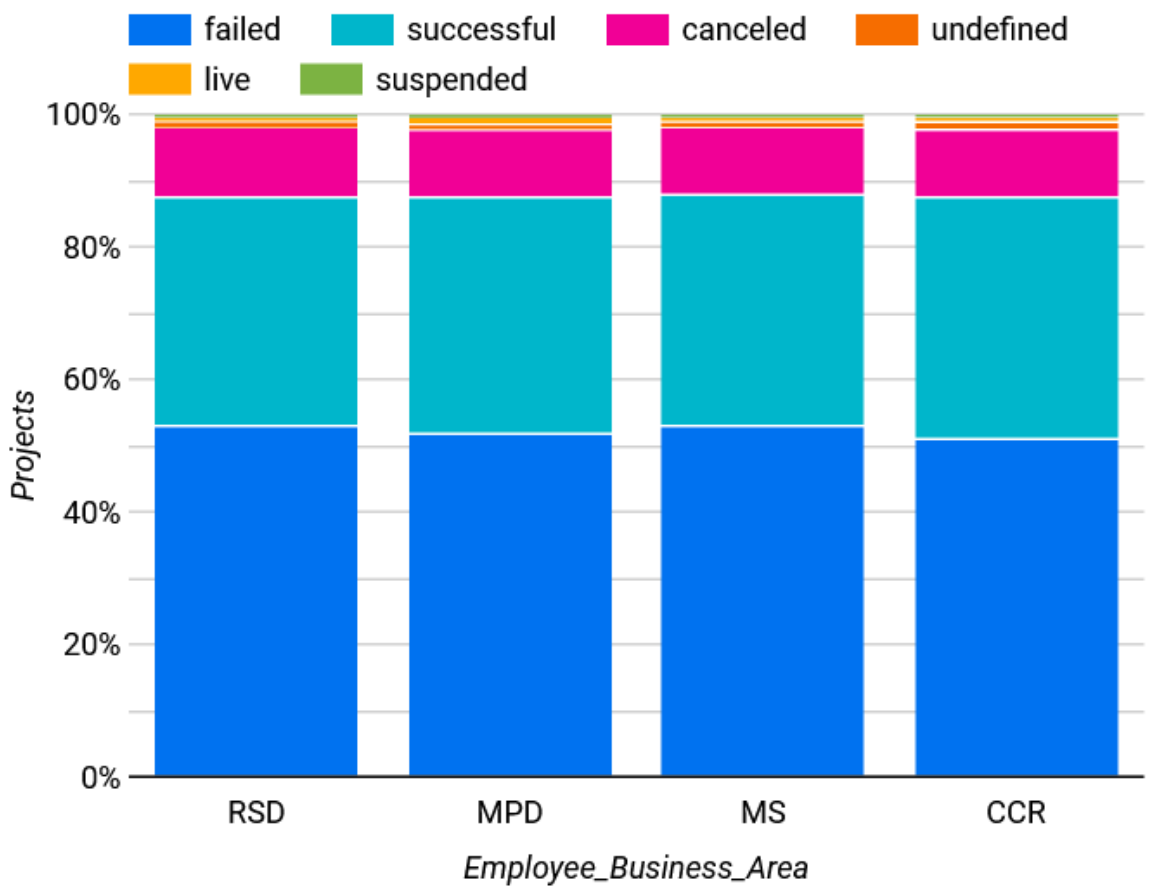


Figure 21: Relation of employee business area to project states

5.1.2 RSD EmplInfo

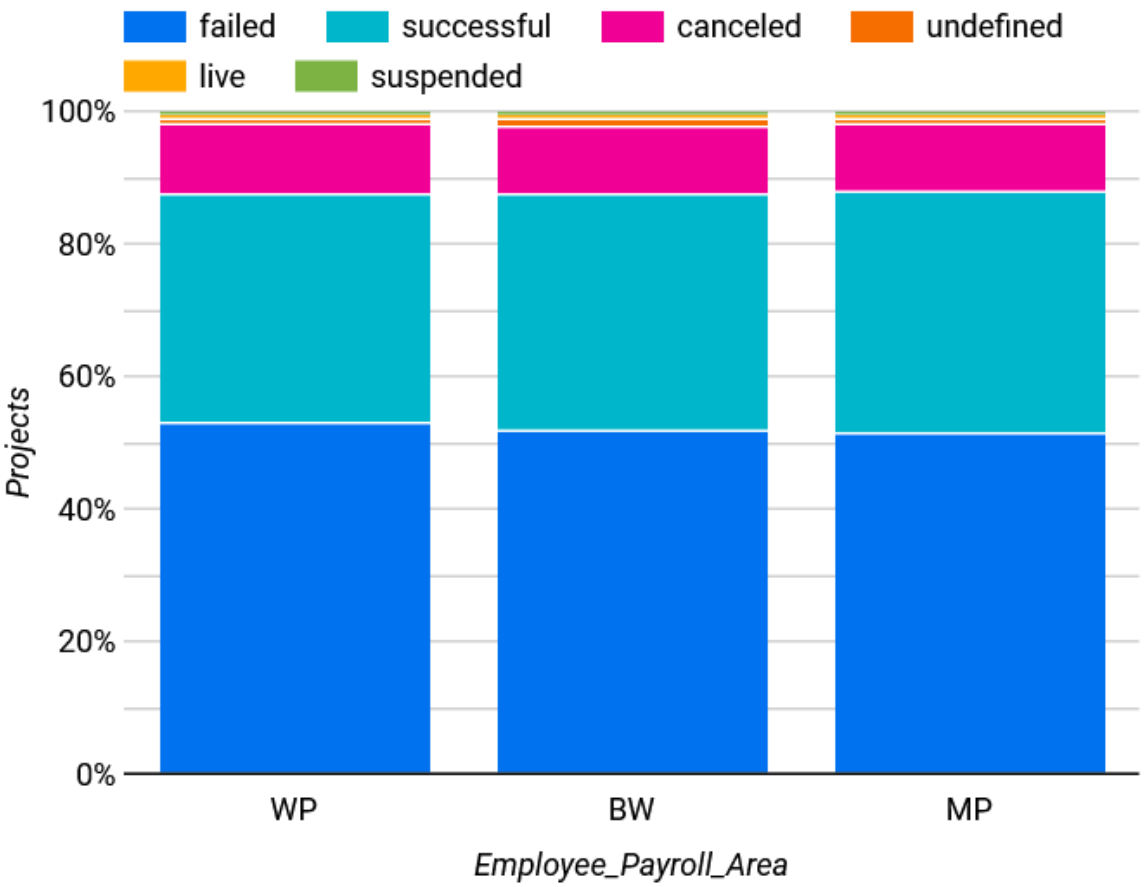


Figure 22: Relation of employee payroll area to project states

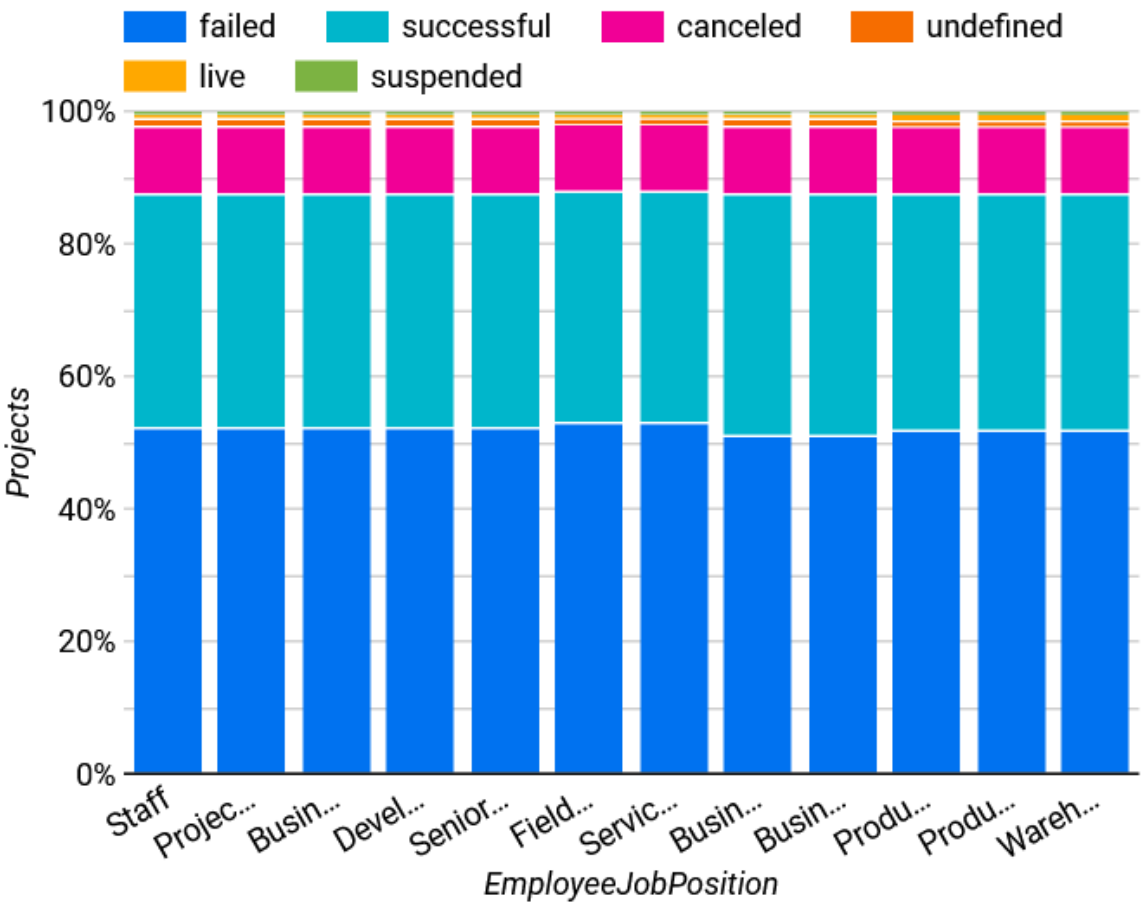


Figure 23: Relation of employee position to project states

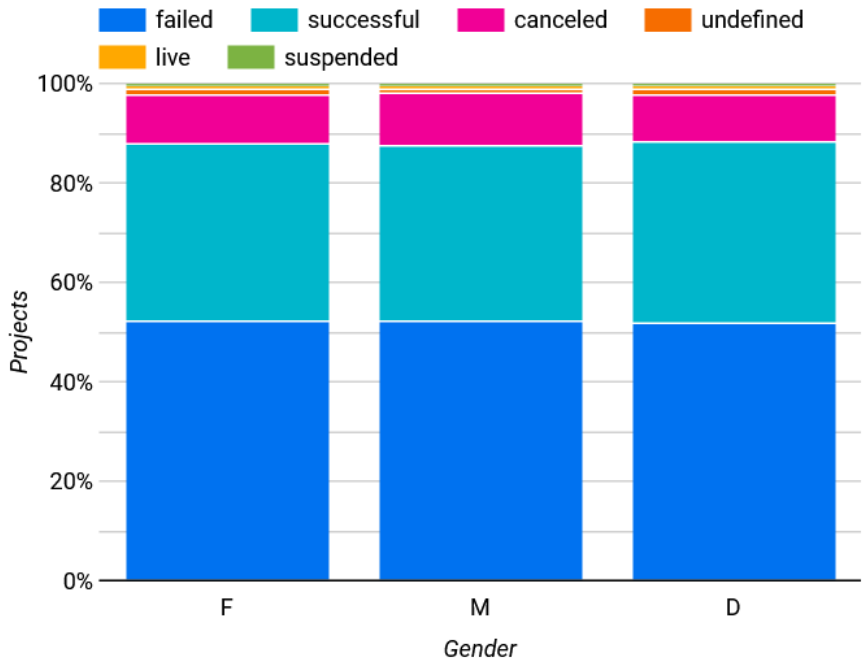


Figure 24: Relation of employee's gender to project state

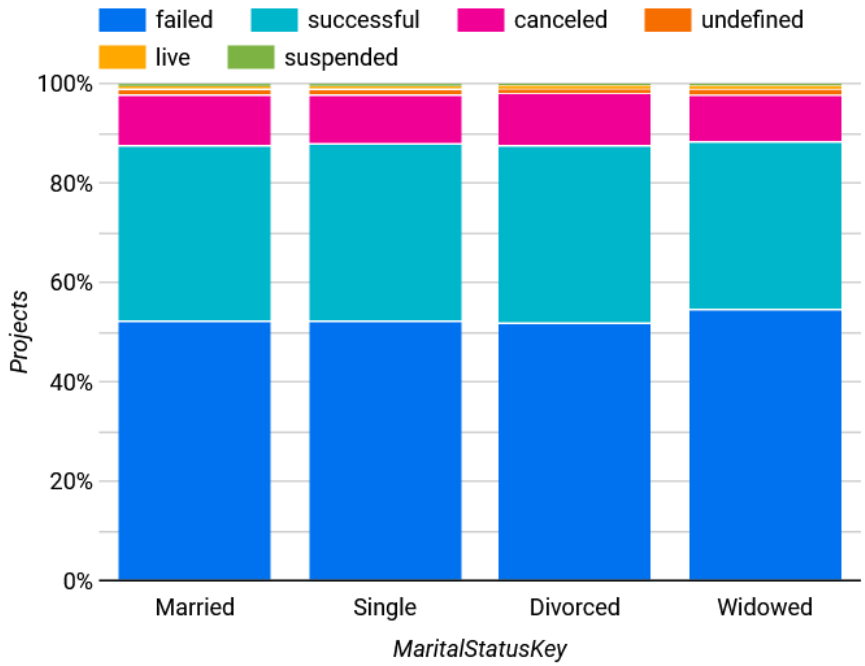


Figure 25: Relation of employee's marital status to project state

5.1.3 RSD EmplQualifications

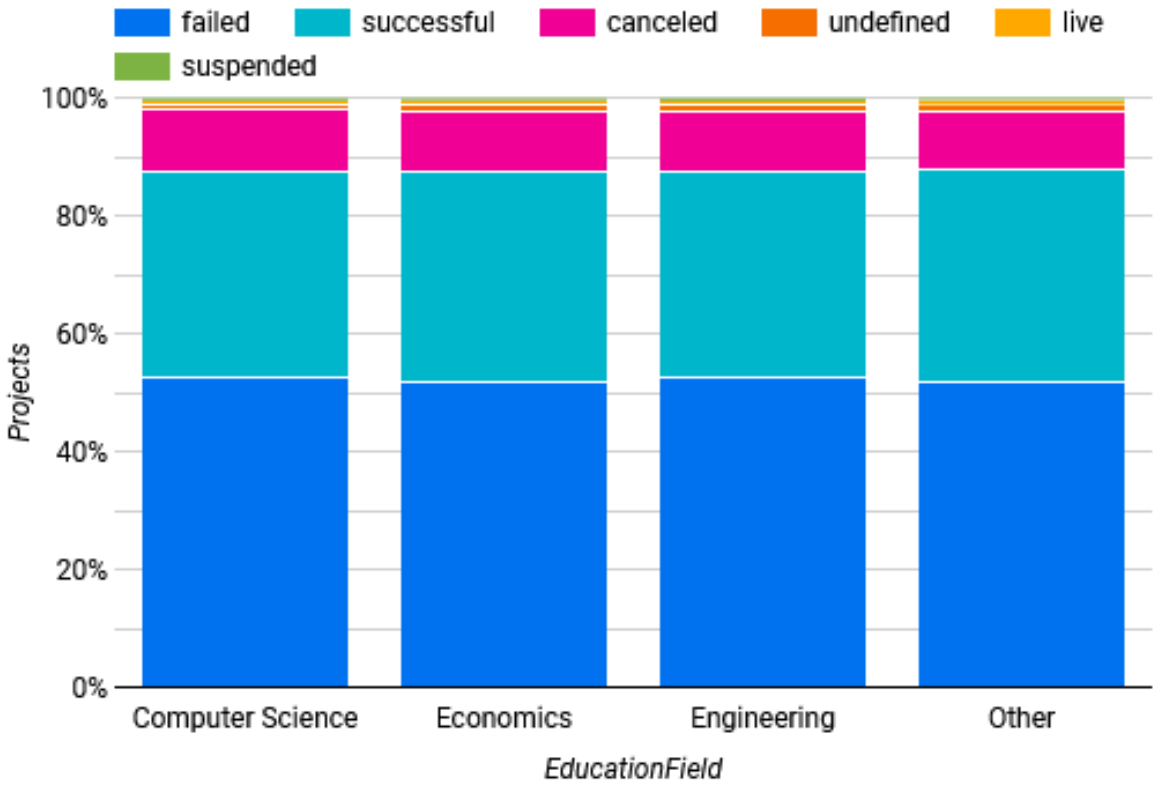


Figure 26: Relation of field the employee was educated in to project state

5.1.4 RSD ProjCost

```
SELECT *, COUNT(*) AS Projects FROM (
  select state, (CASE
  WHEN Budget <= 50000 THEN 'low'
  WHEN Budget BETWEEN 50000 AND 300000 THEN 'mid'
  WHEN Budget > 300000 THEN 'high' END) AS BudgetRange
  FROM `vlba-rsd-grp3.RSD_department.RSD_ProjCost` C JOIN
      `vlba-rsd-grp3.RSD_department.RSD_SAPProj` P ON P.ID = C.Project_ID
) GROUP BY state, BudgetRange;
```

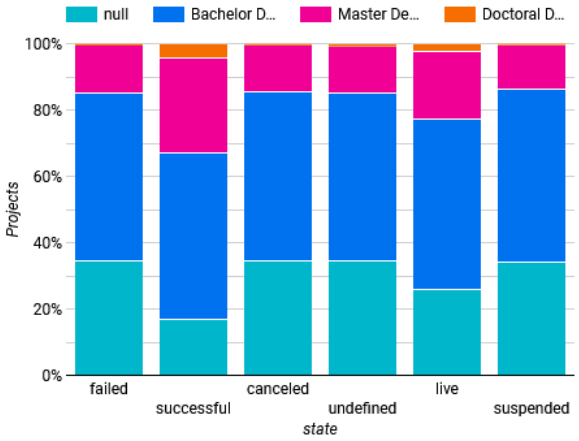


Figure 27: Appendix: The titles employees in each project state held.

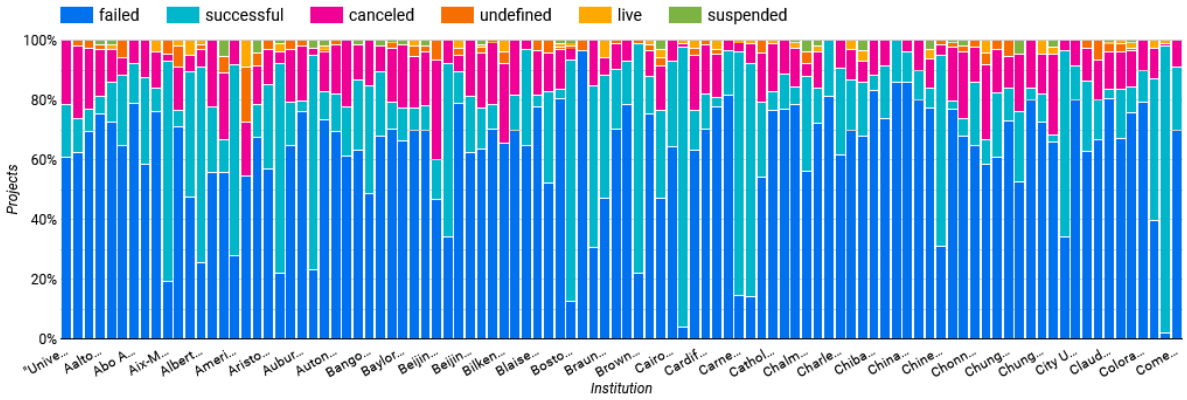


Figure 28: Project state per universities for attendees

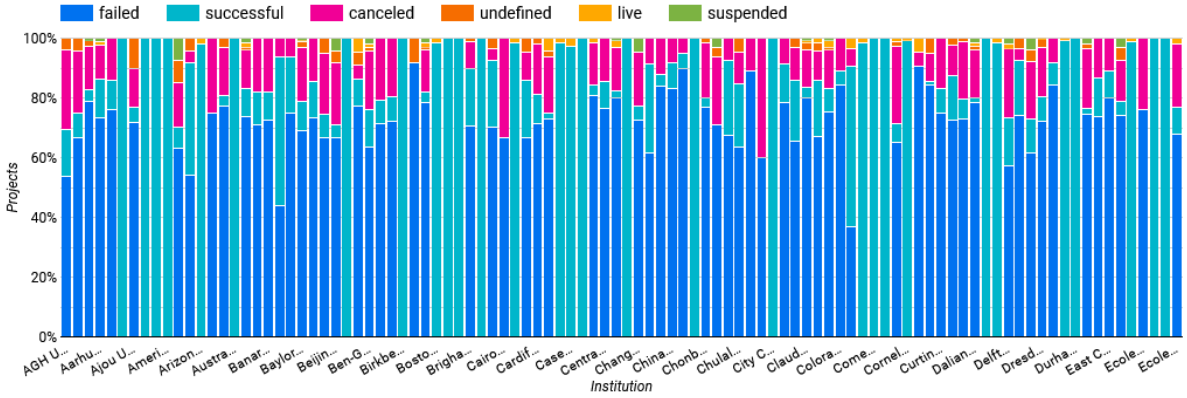


Figure 29: Project state per universities for graduates

5.1.5 RSD SAPProj

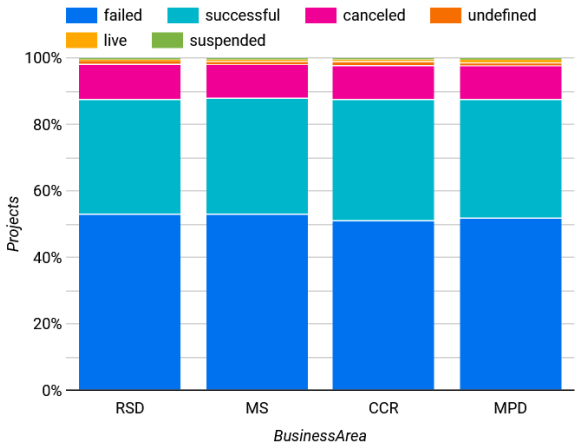


Figure 30: Project states per project business area

	RESNR	Total_Count	Success_Count	Live_Count
1.	49791	9	8	1
2.	22391	8	7	1
3.	36938	8	7	1
4.	19993	5	4	1
5.	67570	5	4	1
6.	21051	5	4	1
7.	11068	4	3	1
8.	24421	4	3	1
9.	10956	4	3	1
10.	23196	4	3	1
11.	32269	4	3	1
12.	37351	4	3	1
13.	29020	4	3	1
14.	42903	4	3	1
15.	58384	3	2	1
16.	49625	3	2	1
17.	58748	3	2	1
18.	41280	3	2	1
19.	43934	3	2	1
20.	28368	3	2	1
21.	38493	3	2	1
22.	42747	3	2	1
23.	35640	3	2	1
24.	23863	3	2	1

	RESNR	Total_Count	Success_Count	Live_Count
25.	68987	3	2	
26.	18973	3	2	
27.	67122	3	2	
28.	18699	3	2	
29.	22073	2	1	
30.	25712	2	1	
31.	24888	2	1	
32.	23324	2	1	
33.	70235	2	1	
34.	39789	2	1	
35.	58508	2	1	
36.	29362	2	1	
37.	27086	2	1	
38.	41099	2	1	
39.	56055	2	1	
40.	21163	2	1	
41.	12741	2	1	
42.	55765	2	1	
43.	49101	2	1	
44.	35449	2	1	
45.	27129	2	1	
46.	21248	2	1	
47.	42516	2	1	
48.	71993	2	1	

Figure 31: List of responsible employees with 'mixed' success

5.1.6 RSD SAPProj - responsible employee

```
WITH EmpResp AS (SELECT MemberId, Member_Name, SUM (CASE WHEN MemberId=RESNR
THEN 1 ELSE 0 END) AS Responsible, SUM (CASE WHEN MemberId=RESNR THEN 0
ELSE 1 END) AS Member, COUNT(*) as Projects
FROM  vlba -rsd-grp3.RSD_department.RSD_SAPProj JOIN
 vlba -rsd-grp3.RSD_department.RSD_ProjMembers ON ID = Project_ID
GROUP BY MemberId, Member_Name ORDER BY Projects)
SELECT Responsibility_Rate, Count(*) AS Employees FROM (
SELECT *, Responsible/Projects AS Responsibility_Rate FROM EmpResp)
GROUP BY Responsibility_Rate
```

Listing 1: Project states per responsible employee

```
SELECT RESNR, RESNA,
    SUM(CASE WHEN state != '-' THEN 1 ELSE 0 END) AS Total_Count,
    SUM(CASE WHEN state = 'successful' THEN 1 ELSE 0 END) AS Success_Count,
    SUM(CASE WHEN state = 'successful' THEN 1 ELSE 0 END)/SUM(CASE WHEN
        state != '-' THEN 1 ELSE 0 END) AS Success_Ratio,
    SUM(CASE WHEN state = 'failed' THEN 1 ELSE 0 END) AS Failure_Count,
```

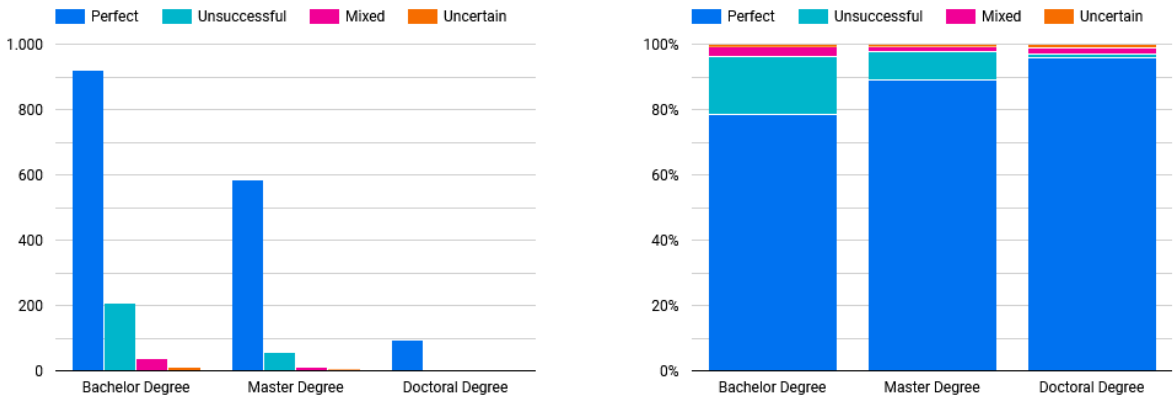


Figure 32: Relation of ERP project success to academic degree

```
    SUM(CASE WHEN state = 'failed' THEN 1 ELSE 0 END)/SUM(CASE WHEN state !=
        '-' THEN 1 ELSE 0 END) AS Failure_Ratio,
    SUM(CASE WHEN state = 'canceled' THEN 1 ELSE 0 END) AS Canceled_Count,
    SUM(CASE WHEN state = 'canceled' THEN 1 ELSE 0 END)/SUM(CASE WHEN state
        != '-' THEN 1 ELSE 0 END) AS Cancel_Ratio,
    SUM(CASE WHEN state = 'live' THEN 1 ELSE 0 END) AS Live_Count,
    SUM(CASE WHEN state = 'live' THEN 1 ELSE 0 END)/SUM(CASE WHEN state !=
        '-' THEN 1 ELSE 0 END) AS Live_Ratio,
    SUM(CASE WHEN state = 'undefined' THEN 1 ELSE 0 END) AS Undefined_Count,
    SUM(CASE WHEN state = 'undefined' THEN 1 ELSE 0 END)/SUM(CASE WHEN state
        != '-' THEN 1 ELSE 0 END) AS Undefined_Ratio,
    SUM(CASE WHEN state = 'suspended' THEN 1 ELSE 0 END) AS Suspended_Count,
    SUM(CASE WHEN state = 'suspended' THEN 1 ELSE 0 END)/SUM(CASE WHEN state
        != '-' THEN 1 ELSE 0 END) AS Suspended_Ratio
FROM `vlba-rsd-grp3.RSD_department.RSD_SAPProj`
GROUP BY RESNR, RESNA;
```

Listing 2: Sucess rates of responsible employees

```
 SELECT RESNR, RESNA,
  (CASE WHEN Success_Ratio = 1.0 THEN 'Perfect' ELSE (CASE WHEN
      Success_Ratio = 0.0 THEN (CASE WHEN Live_Ratio + Suspended_Ratio = 1.0
      THEN 'Uncertain' ELSE 'Unsuccessful' END) ELSE 'Mixed' END) END) AS
      Success_Factor,
  Total_Count, Success_Count, Failure_Count, Canceled_Count, Live_Count,
      Suspended_Count, Undefined_Count FROM
`vlba-rsd-grp3.Generated_views.project-responsible-results`
```

Query for Relation of project success to academic degree of ERP

```
SELECT AcademicDegree, title, Success_Factor, COUNT(*) AS number FROM
    `vlba-rsd-grp3.Generated_views.project-responsible-data` D JOIN
    `vlba-rsd-grp3.Generated_views.project-responsible-success-factor` R ON
    D.RESNR = R.RESNR GROUP BY AcademicDegree, title, Success_Factor ORDER
    BY title
```

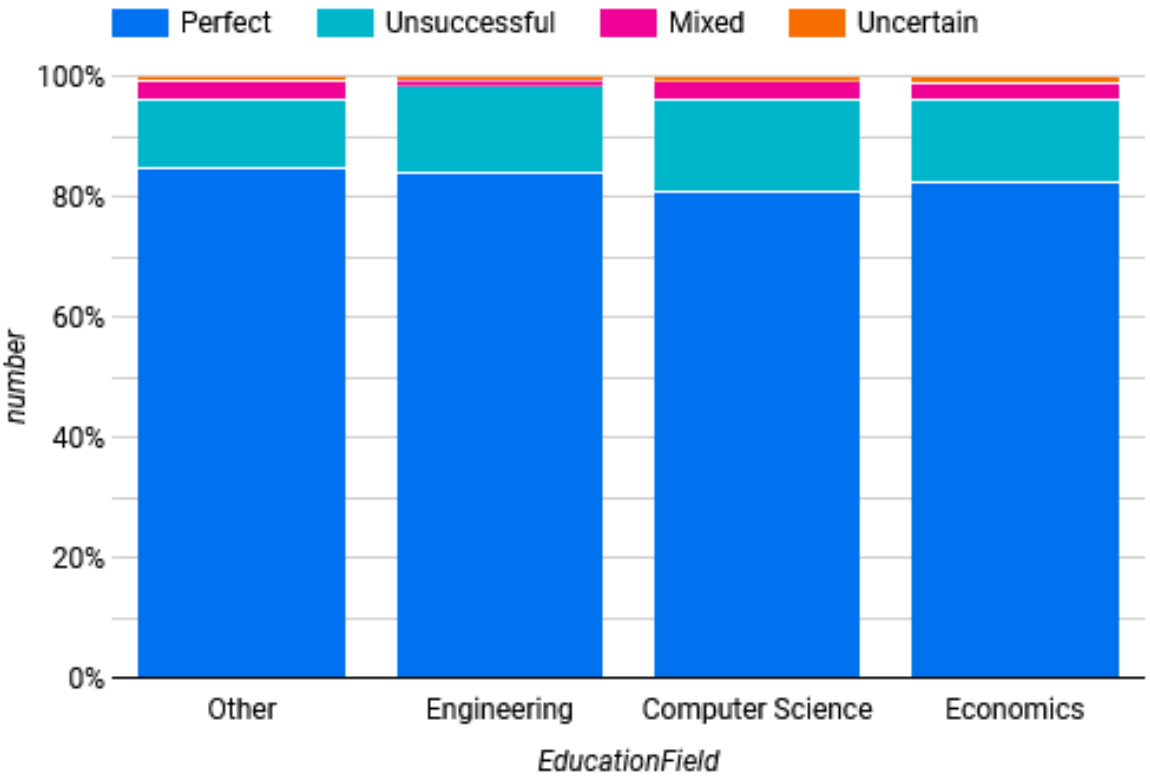


Figure 33: ERP Education Fields with corresponding success factors as 100%

5.1.7 RSD SAPProj - regular project member

Listing 3: Query to group project members according to their cumulative project success

```
SELECT Success_Factor, COUNT(*) AS Member, SUM(Total_Count) AS Projects FROM
    (
SELECT MemberId, Member_Name, (CASE WHEN Success_Ratio = 1.0 THEN 'Perfect'
    ELSE (CASE WHEN Success_Ratio = 0.0 THEN (CASE WHEN Live_Ratio +
    Suspended_Ratio = 1.0 THEN 'Uncertain' ELSE 'Unsuccessful' END) ELSE
    'Mixed' END) END) AS Success_Factor,
  Total_Count FROM
`vlba-rsd-grp3.Generated_views.project-member-results` WHERE MemberId NOT IN
    (SELECT RESNR
    FROM`vlba-rsd-grp3.Generated_views.project-responsible-results`)) GROUP
    BY Success_Factor;
```

Listing 4: Query to list all project members with mixed project success

```
SELECT * FROM (
SELECT *, (CASE WHEN Success_Ratio = 1.0 THEN 'Perfect' ELSE (CASE WHEN
    Success_Ratio = 0.0 THEN 'Unsuccessful' ELSE 'Mixed' END) END) AS
    Success_Factor, Total_Count FROM
`vlba-rsd-grp3.Generated_views.project-member-results` WHERE MemberId NOT IN
    (SELECT RESNR
    FROM`vlba-rsd-grp3.Generated_views.project-responsible-results`)) WHERE
    Success_Ratio > 0 AND Success_Ratio < 1 AND Failure_Ratio >= 0 AND
    Cancel_Ratio >= 0 AND Failure_Ratio < 1 AND Cancel_Ratio < 1
```

5.1.8 Successful universities

Listing 5: Query for university success rates

```
SELECT *, SuccessfulProjects/TotalProjects AS SuccessRate,
    FailedProjects/TotalProjects AS FailedRate,
    CanceledProjects/TotalProjects AS CancelRate FROM (SELECT
    institution, SUM(Total_Count) AS TotalProjects, SUM(Success_Count) AS
    SuccessfulProjects, SUM(Failure_Count) AS FailedProjects,
    SUM(Canceled_Count) AS CanceledProjects, SUM(Live_Count) AS
    LiveProjects, SUM(Suspended_Count) AS SuspendedProjects,
    SUM(Undefined_Count) AS UndefinedProjects, COUNT(*) as RespGraduates
    FROM (SELECT * FROM
    `vlba-rsd-grp3.Generated_views.project-responsible-results` JOIN
    `vlba-rsd-grp3.RSD_department.RSD_EmplQualifications` Q ON RESNR =
    PERNR) GROUP BY institution);
```

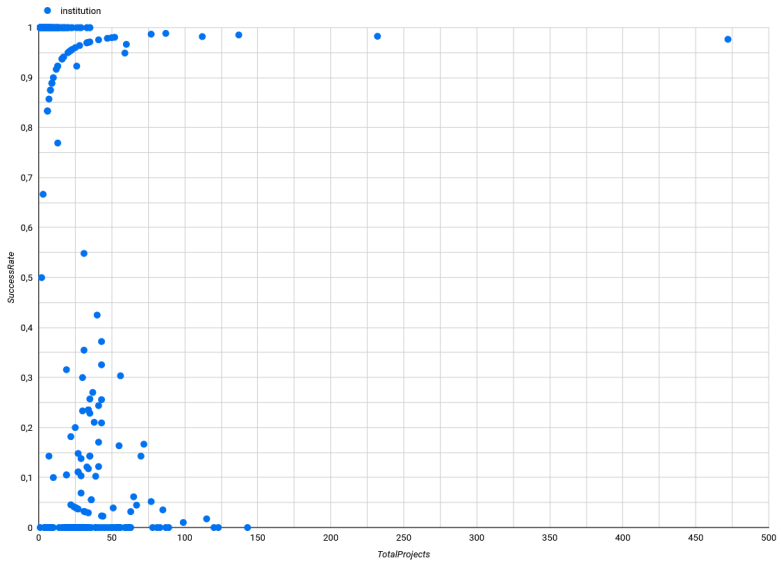


Figure 34: University success rate with projects

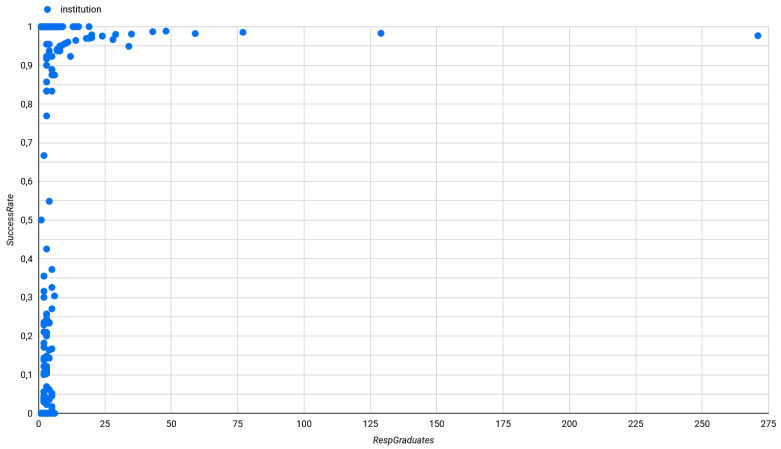


Figure 35: university success rate with graduates