Post Graduation in Data Science & Business Analytics - 11th Edition
Data Science Project

Decision Support Model for Determining Appropriate Loan Interest Rates for Customers

Carlos Ramagem | i30386
Luiza Campos | i30387
Thaís Almeida | i30394

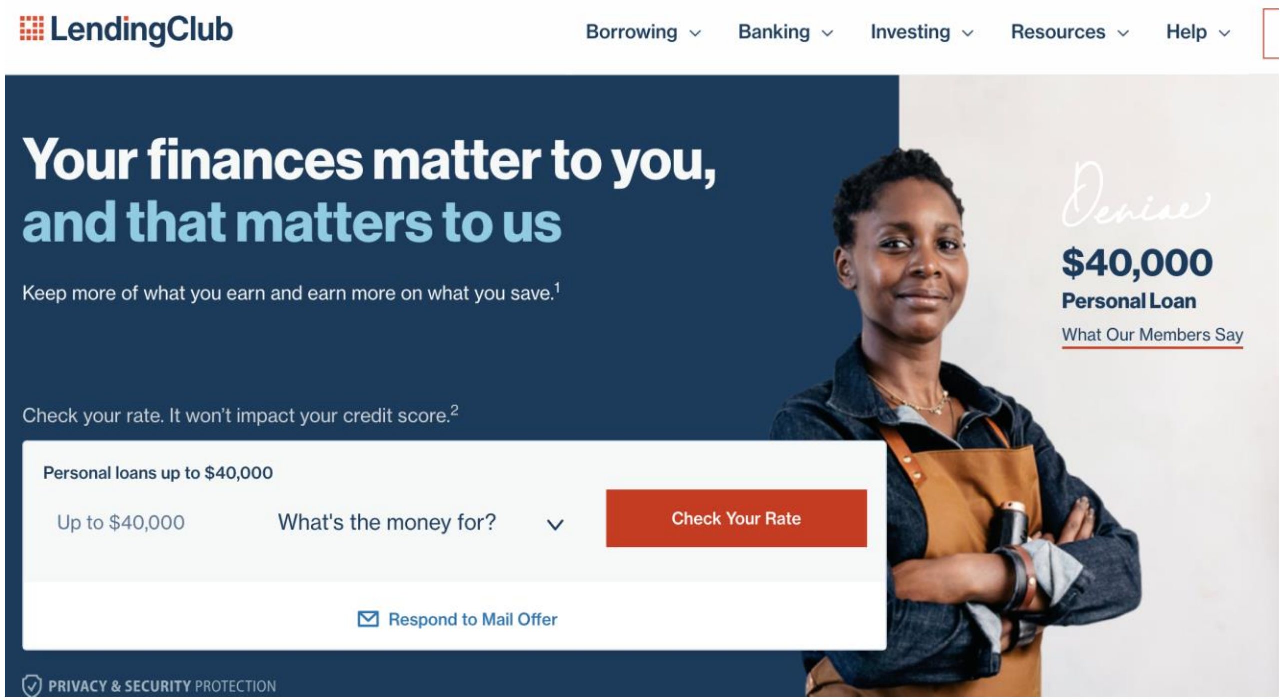
LendingClub

INDEX

- Introduction to the use case and project objectives
- Presentation of the main technical issues implemented
  - 2.1 Dimensional Model
  - 2.2 ETL
  - 2.3 Semantic Model
- Presentation of conclusions obtained through report navigation (Power BI)
- Presentation of conclusions obtained and explanation of the implemented Machine Learning model (ML Studio)
- Conclusions

LendingClub

Use case



www.lendingclub.com

- Online Crowdfunding Loan Platform
- Provides accessible credit to individuals/entities by establishing a "bridge" between investors (who provide the loan) and borrowers.
- Transactions are carried out through direct interactions between the different agents on the platform (borrowers and investors/subscribers of bonds).

Objective:

To provide recommendations for more competitive interest rates in the market, aiming to offer better profit margins for companies while reducing the risk of default.

LendingClub

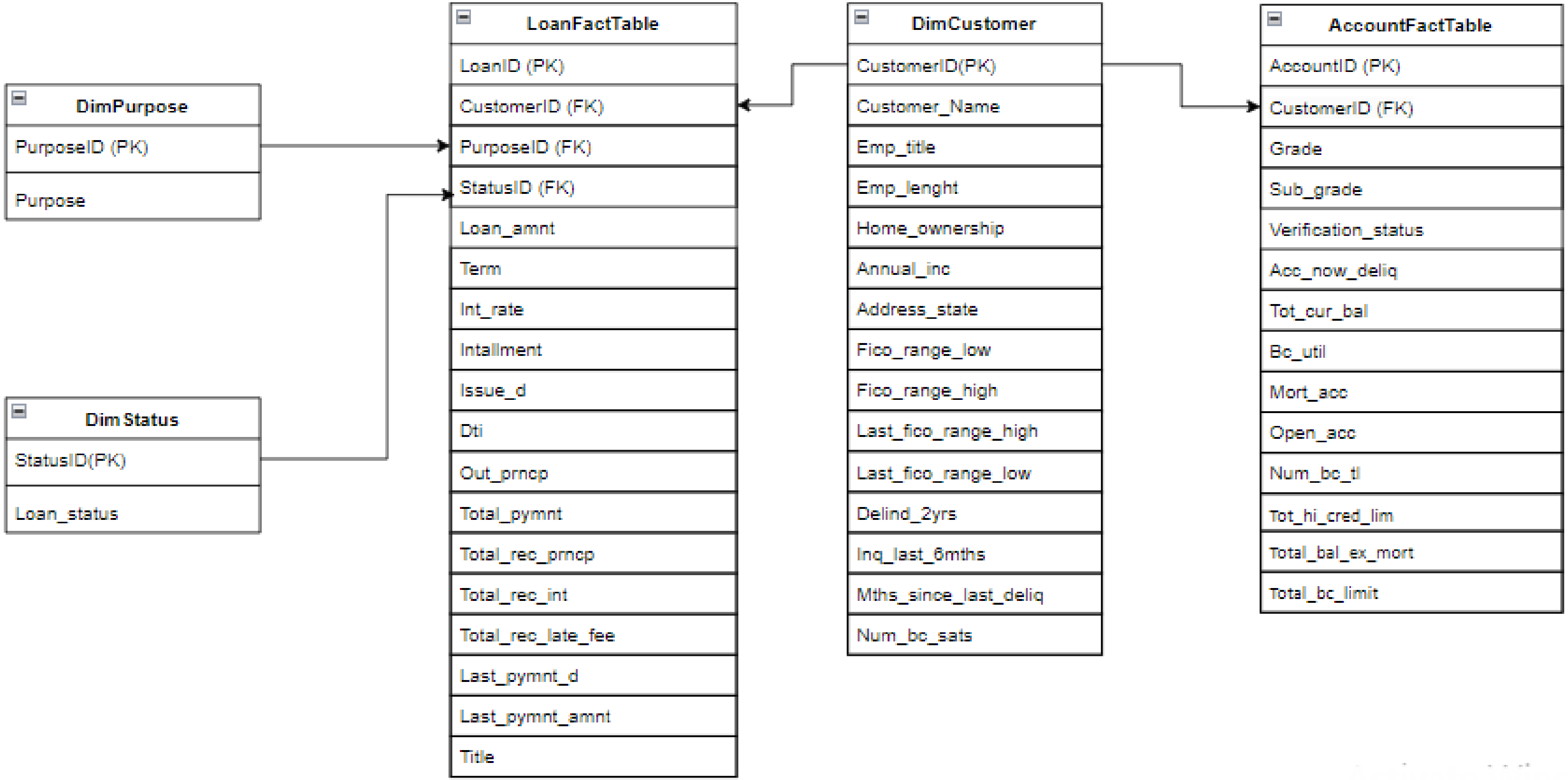
Script to create the star schema

SQL script defining all variables contained in the star schema model

```sql
/*Criando a tabela DSP11_GR10.DimCustomer */
Create table DSP11_GR10.DimCustomer(
[CustomerID] [BIGINT]PRIMARY KEY,
[Customer_name][varchar](100) NULL,
[Emp_title] [varchar](100) NULL,
[Emp_length] [int] NULL,
[Home_ownership] [varchar](100) NULL,
[Annual_inc] [float] NULL,
[Addr_state] [varchar](100) NULL,
[Fico_range_low] [int] NULL,
[Fico_range_high] [int] NULL,
[Last_fico_range_high][int] NULL,
[Last_fico_range_low][int] NULL,
[Delinq_2yrs] [int] NULL,
[Inq_last_6mths] [int] NULL,
[Mths_since_last_delinq] [int] NULL,
[Num_bc_sats][int] NULL,
);

DROP TABLE DSP11_GR10.DimCustomer

select * from DSP11_GR10.DimCustomer

GO

/*Criando a tabela DSP11_GR10.DimPurpose */
Create table DSP11_GR10.DimPurpose(
[PurposeID] [BIGINT]PRIMARY KEY,
[Purpose] [varchar](100) NULL,
);

GO

DROP TABLE DSP11_GR10.DimPurpose

select * from DSP11_GR10.DimPurpose

/*Criando a tabela DSP11_GR10.DimStatus */
Create table DSP11_GR10.DimStatus(
[StatusID][BIGINT]PRIMARY KEY,
[Loan_status] [varchar](100) NULL,
);
select * from DSP11_GR10.DimStatus
```

```sql
/*Criando a tabela DSP11_GR10.LoanFactTable */
CREATE TABLE DSP11_GR10.LoanFactTable(
 [LoanID][BIGINT]PRIMARY KEY,
 [CustomerID] [int] NULL,
 [PurposeID] [int] NULL,
 [StatusID] [int] NULL,
 [Loan_amnt][int] NULL,
 [Term][int] NULL,
 [Int_rate] [float] NULL,
 [Installment] [float] NULL,
 [Issue_d][date],
 [Dti][float] NULL,
 [Out_prncp] [float] NULL,
 [Total_pymnt] [float] NULL,
 [Total_rec_prncp] [float] NULL,
 [Total_rec_int] [float] NULL,
 [Total_rec_late_fee] [float] NULL,
 [Last_pymnt_d][date],
 [Last_pymnt_amnt][float] NULL,
 [Title] [varchar](100) NULL);

select * from DSP11_GR10.LoanFactTable

DROP TABLE DSP11_GR10.LoanFactTable

/*Criando a tabela DSP11_GR10.AccountFactTable */
CREATE TABLE DSP11_GR10.AccountFactTable(
[AccountID][BIGINT]PRIMARY KEY,
[CustomerID] [int] NULL,
[Grade] [varchar](50) NULL,
[Sub_grade] [varchar](50) NULL,
[Verification_status][varchar](100) NULL,
[Acc_now_delinq][int] NULL,
[Tot_cur_bal][int] NULL,
[Bc_util][float] NULL,
[Mort_acc][int] NULL,
[Open_acc] [int] NULL,
[Num_bc_tl][int] NULL,
[Tot_hi_cred_lim][int] NULL,
[Total_bal_ex_mort][int] NULL,
[Total_bc_limit][int] NULL);

DROP TABLE DSP11_GR10.AccountFactTable
```

Star schema

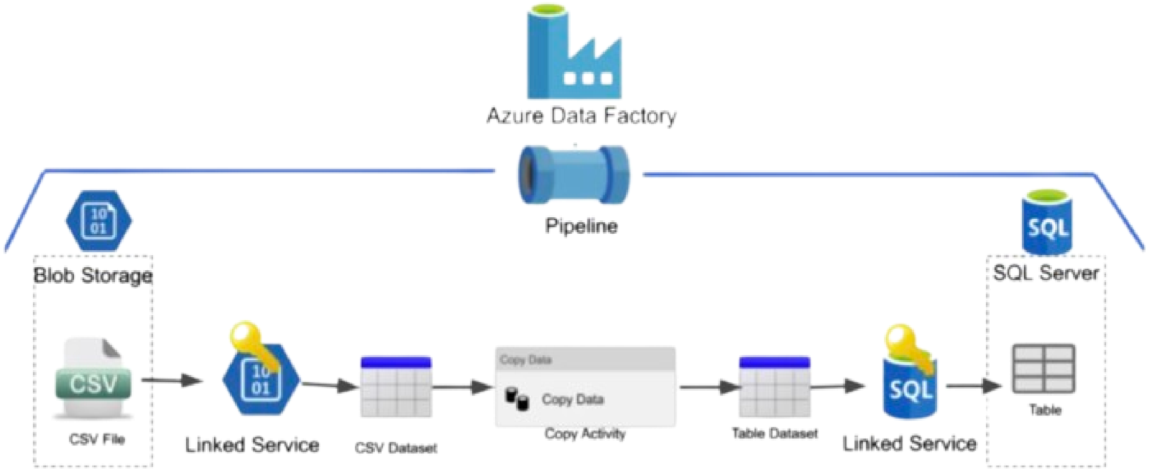

DimPurpose
PurposeID (PK)
Purpose
Dim Status
StatusID(PK)
Loan_status
LoanFactTable
LoanID (PK)
CustomerID (FK)
PurposeID (FK)
StatusID (FK)
Loan_amnt
Term
Int_rate
Intallment
Issue_d
Dti
Out_prncp
Total_pymnt
Total_rec_prncp
Total_rec_int
Total_rec_late_fee
Last_pymnt_d
Last_pymnt_amnt
Title
DimCustomer
CustomerID(PK)
Customer_Name
Emp_title
Emp_lenght
Home_ownership
Annual_inc
Address_state
Fico_range_low
Fico_range_high
Last_fico_range_high
Last_fico_range_low
Delind_2yrs
Inq_last_6mths
Mths_since_last_deliq
Num_bc_sats
AccountFactTable
AccountID (PK)
CustomerID (FK)
Grade
Sub_grade
Verification_status
Acc_now_deliq
Tot_cur_bal
Bc_util
Mort_acc
Open_acc
Num_bc_tl
Tot_hi_cred_lim
Total_bal_ex_mort
Total_bc_limit

2.2 ETL

Objective: Automate the process of data ingestion, transformation, and importation using Azure Data Factory and Azure Data Studio.

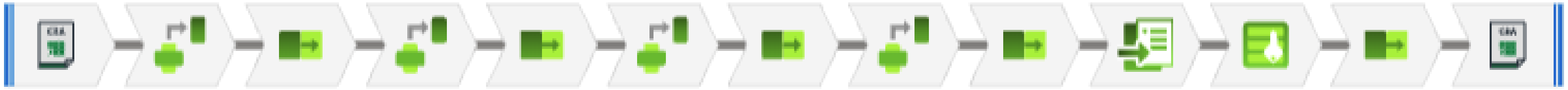
Strategy Used:

- Ingestion: Extraction of data from a blob container in Azure.
- Transformation: Use of the "FlowData" artifact to process and prepare the data.
- Importation: Loading the transformed data into Azure SQL for analysis and centralized storage



Transformation

- DimCustumer

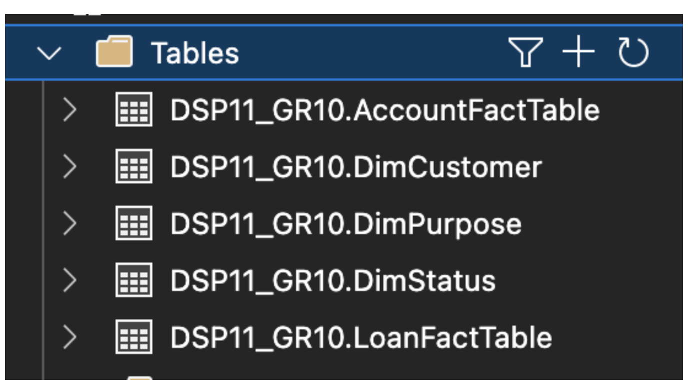


- FactAccount

LendingClub

Importation

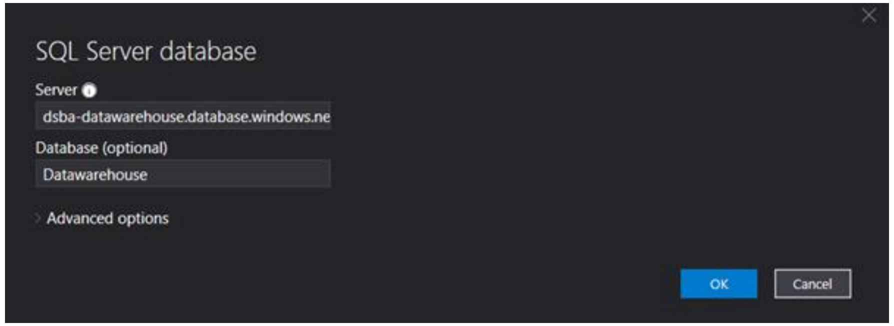
- Transformed data from each DataFlow in Azure Data Factory, related to each entity, was loaded into the Azure SQL account created for this project.
- Pipeline was used to load the data into Azure SQL. For each table, SQL code was executed to set up the corresponding structure according to the data schema.




```sql
/*Criando a tabela DSP11_GR10.DimCustomer */
Create table DSP11_GR10.DimCustomer(
[CustomerId] int identity(1000,1),
[emp_title] [varchar](100) NULL,
[emp_length] [varchar](100) NULL,
[home_ownership] [varchar](100) NULL,
[annual_inc] [float] NULL,
[addr_state] [varchar](100) NULL,
[fico_range_low] [int] NULL,
[fico_range_high] [int] NULL,
[last_fico_range_high][int] NULL,
[last_fico_range_low][int] NULL,
[delinq_2yrs] [int] NULL,
[inq_last_6mths] [int] NULL,
[mths_since_last_delinq] [int] NULL,
[num_bc_sats][int] NULL,
);

GO
```

2.3 Modelo Semântico

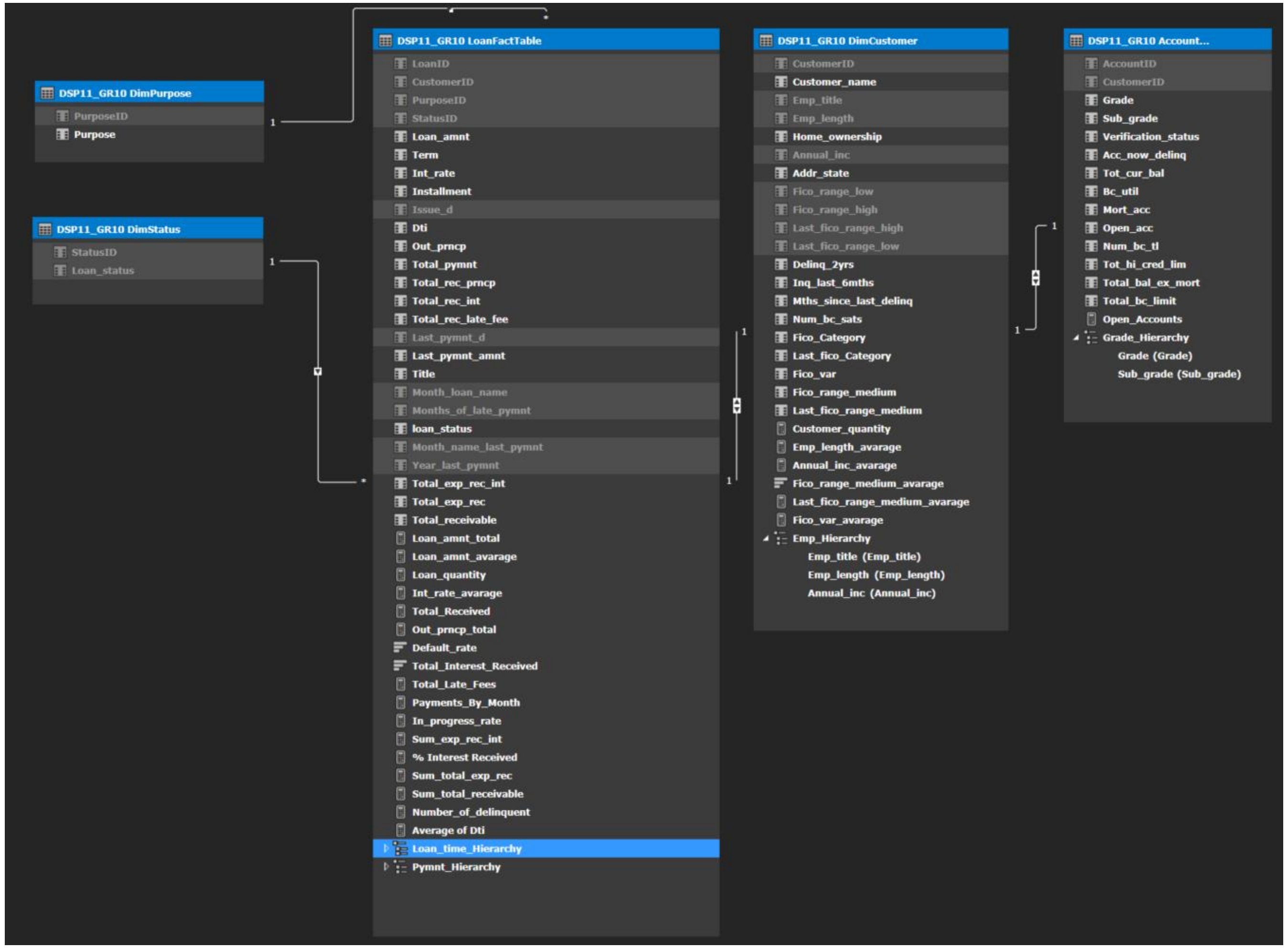
- The transformed and stored data tables in Azure were connected to Visual Studio using the Microsoft Analysis Services template.
- This integration allowed for the import and organization of data tables within the Visual Studio development environment, preparing them for the construction of the semantic model.



2.3 Semantic Model

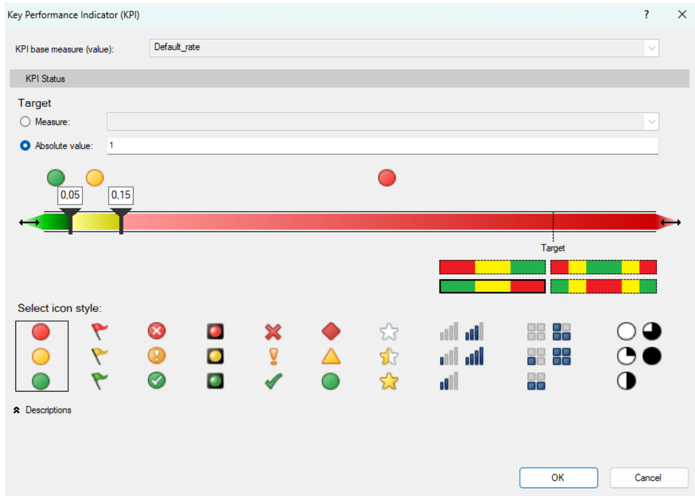
LendingClub

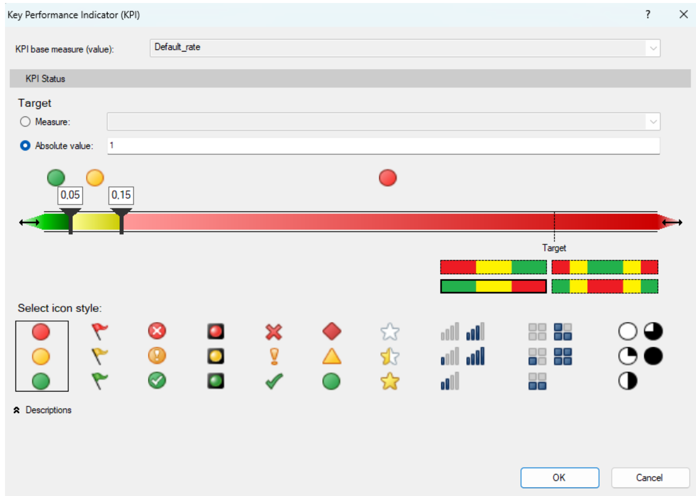
- It was necessary to define the relationships between the tables to facilitate data analysis, including the addition of calculated columns and the definition of measures for business evaluation.

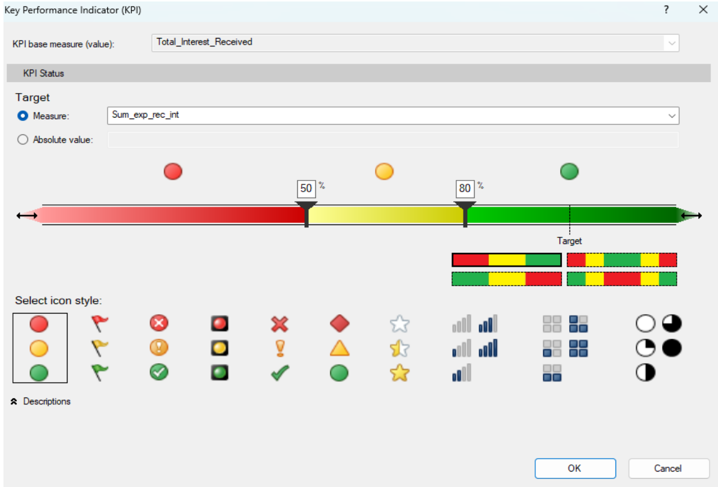


2.3 Semantic Model

Creation of KPIs for Performance Monitoring

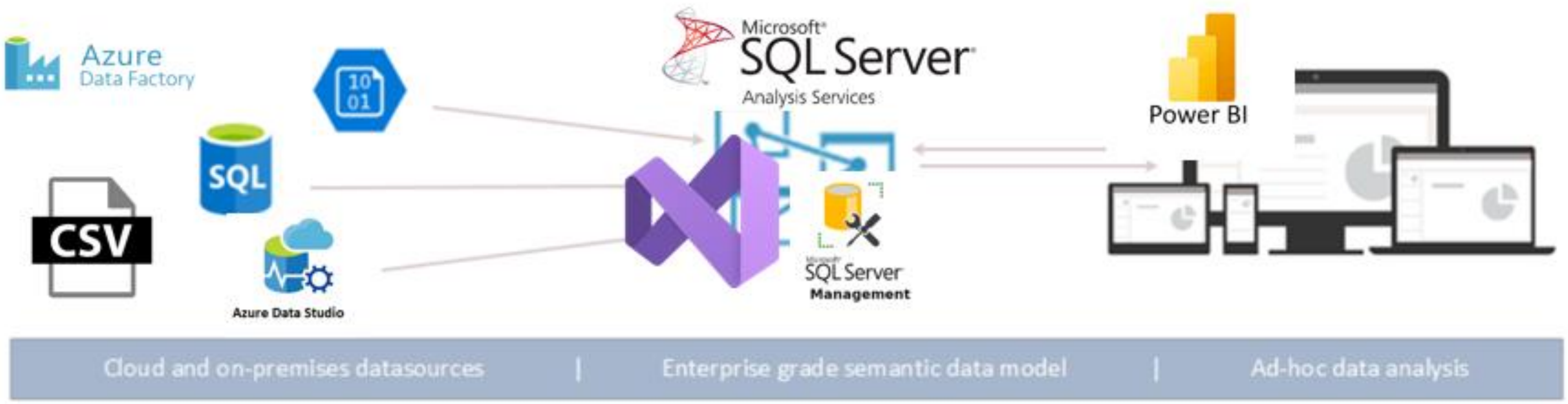




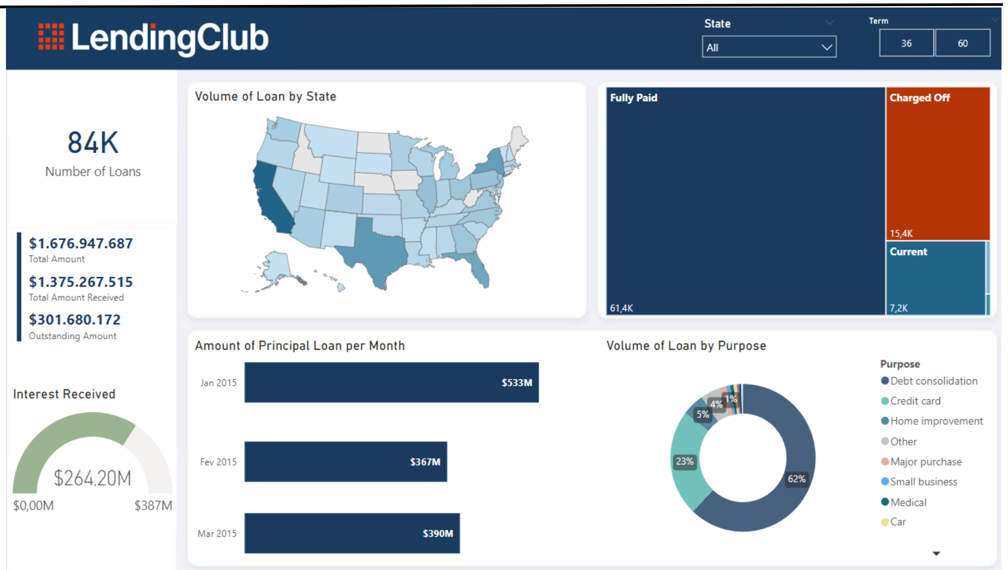


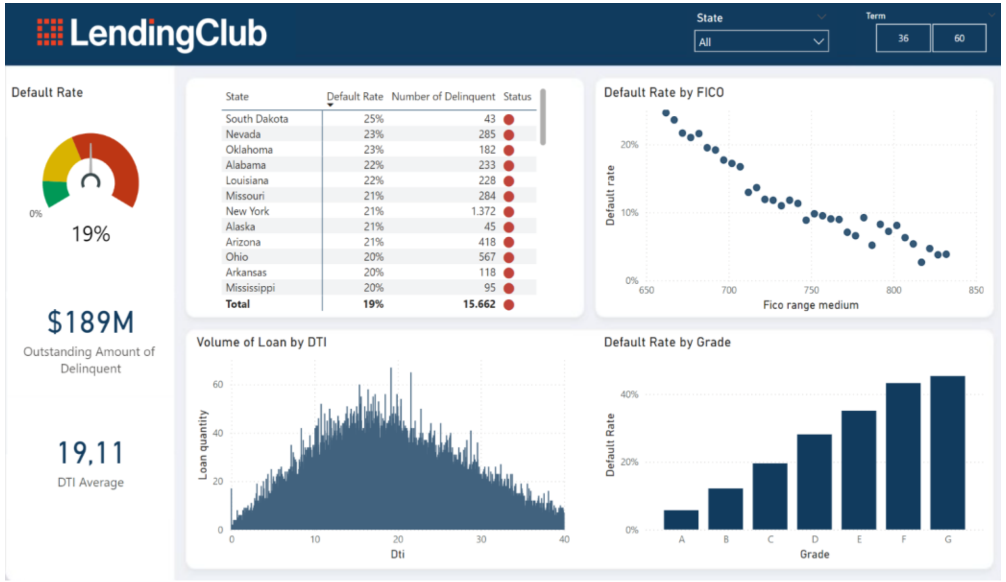
2.3 Semantic Model

PoweBI's deployment



3. Power BI Dashboards





4. Implemented Machine Learning Model (ML Studio)

- **Objective:** Predict the appropriate interest rate for a loan, specifically tailored to each client's profile and history.

- **Strategy Used**: From Simple to Complex

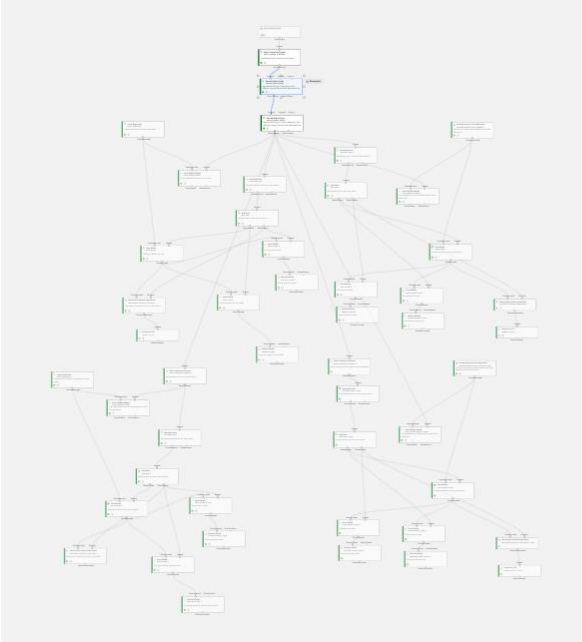
An iterative approach to developing the Machine Learning model.

- **Chosen Feature:** Loan Interest Rate (Int_rate)

- The methods used were **Linear Regression** and **Boosted Decision Tree Regression.**

- For Linear Regression, **L2 regularization** and feature selection strategies were applied as model tuning processes to find the best fit.

- Still not satisfied with the results, **Boosted Decision Tree Regression** was tested with hyperparameter tuning and feature selection to find the optimal model.



4. Implemented Machine Learning Model (ML Studio)

LendingClub

MAE (Mean Absolute Error):
Lowest MAE of 0.90 ± 0.01, indicating higher model accuracy.

RMSE (Root Mean Square Error):
Lowest RMSE of 1.29 ± 0.03, suggesting a better model fit.

RSE (Relative Squared Error):
Lowest RSE of 0.09 ± 0.00, indicating that 9% of the variance in interest rates is explained, highlighting a superior model fit compared to Linear Regression.

RAE (Relative Absolute Error):
Lowest RAE of 0.26 ± 0.00, indicating minimal deviation from actual interest rates compared to other models.

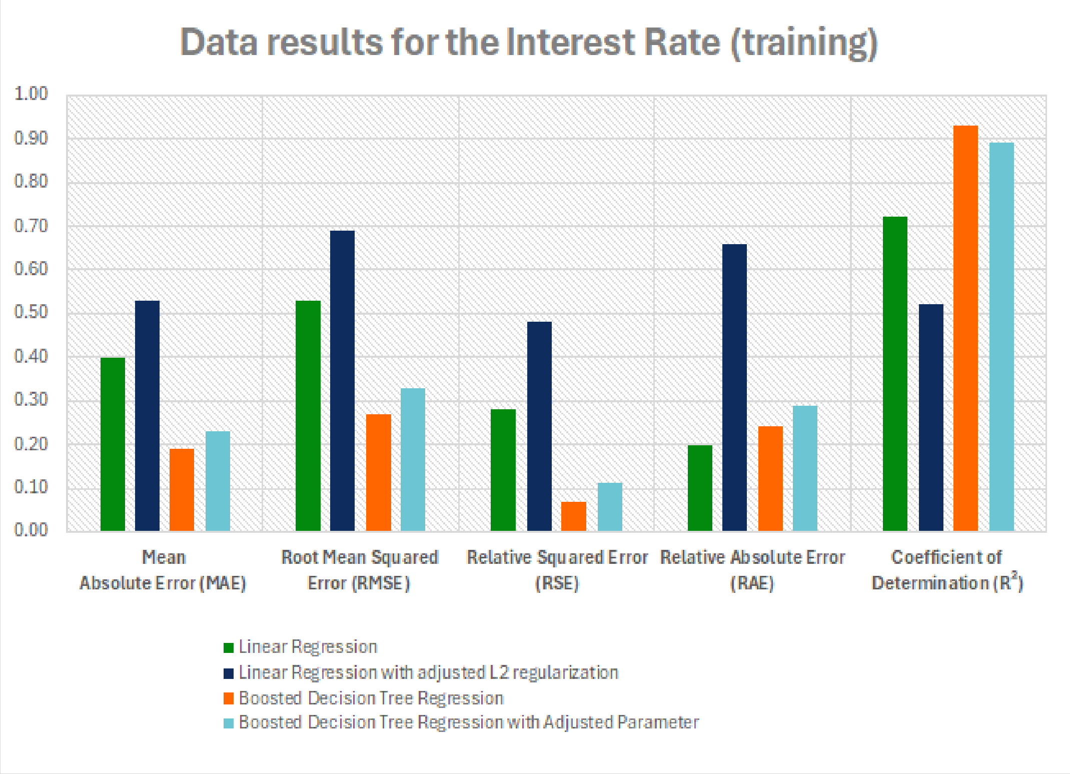
R^2 (Coefficient of Determination):
Highest R^2 of 0.91 ± 0.00, indicating 91% accuracy in predicting interest rates, meaning the model is highly capable of explaining and predicting variations in interest rates based on the analyzed data.

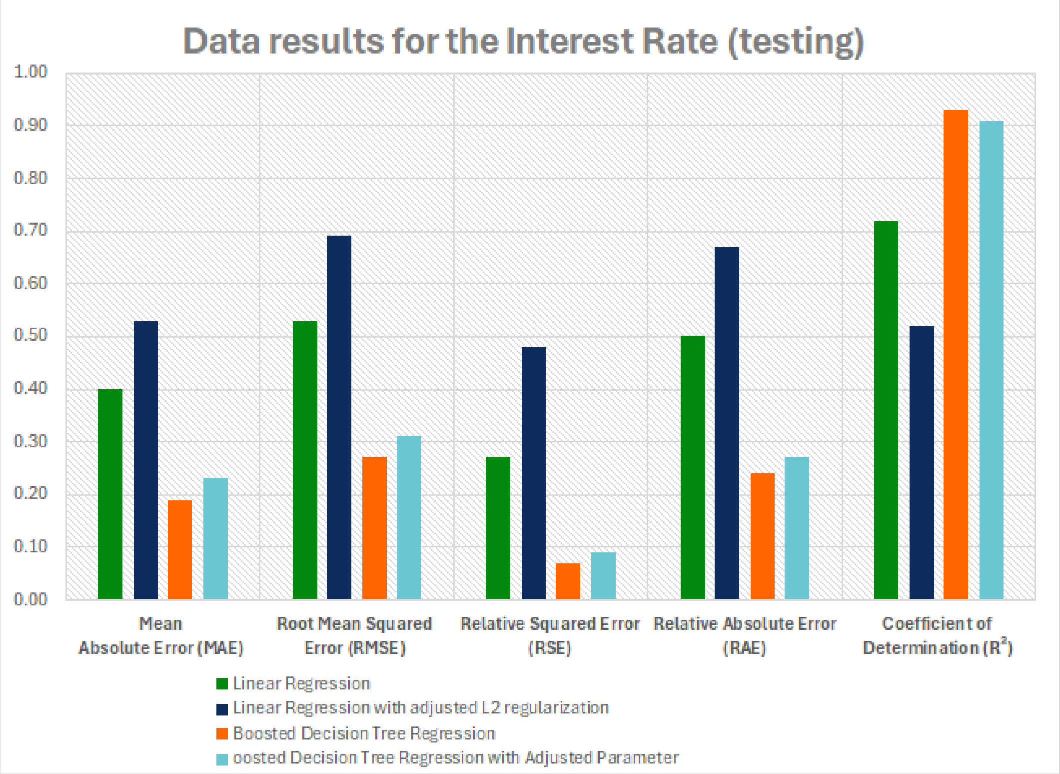
Table 1 – Analysis of Cross validation Results for the Interested Rate (mean accross 10 folds)

Model Version (Cross validation for the Int_rate column)	Mean Absolute Error (MAE) with σ	Root Mean Squared Error (RMSE) with σ	Relative Squared Error (RSE) with σ	Relative Absolute Error (RAE) with σ	Coefficient of Determination (R^2) with σ
Linear Regression	1.74 ± 0.02	2.26 ± 0.03	0.28 ± 0.01	0.51 ± 0.00	0.72 ± 0.01
Linear Regression with adjusted L2 regularization	2.36 ± 0.02	3.06 ± 0.03	0.51 ± 0.01	0.69 ± 0.01	0.49 ± 0.01
Boosted Decision Tree Regression	0.90 ± 0.01	1.29 ± 0.03	0.09 ± 0.00	0.26 ± 0.00	0.91 ± 0.00
Boosted Decision Tree Regression with Adjusted Parameters	0.99 ± 0.01	1.40 ± 0.02	0.11 ± 0.00	0.29 ± 0.00	0.89 ± 0.00

4. Modelo de Machine Learning implementado (ML Studio)

LendingClub





The similarity between the training model result metrics suggests that the model is well-fitted, generalizes well, and is robust.

5. Conclusões

•The company faces challenges with customer profiles and a high default rate of 19%, highlighting the need to improve risk management.

•The analysis underscores the importance of rigorous evaluation criteria, given the negative correlation between FICO scores and defaults.

•Implementing the ML techniques proposed in this report for credit approval analysis can reduce loan risks and optimize credit management, critically managing the outstanding amount of $189 million.

LendingClub

THANK YOU