

Data_mining_prepmid -> orange

1)Data mining dataset: German Credit

2)Data mining EDA or visualize:

1) Scatter plot

Scatter Plot Analysis:

Detail:

Insights Derived:

2) Distribution plot

Detail:

Insights Derived:

Next Steps for Further Analysis:

3)Box Plot

3.Box Plot Analysis:

Key Insights from Box Plot Analysis:

Examples of Comparisons:

4. Visualization Goals:

3) Data mining Classification:

1. Define the Problem:

2. Steps to Solve the Problem:

3. Choose Classifiers:

4. Evaluation Metrics:

Examples of Comparisons:

5. Insights from Results:

6. Stacking Technique:

7. Kernel Comparison for SVM:

8. Additional Insights:

9. Key Points to Include:

10. Visuals:

AUC-ROC curves for all classifiers:

Confusion matrices

Comparison charts of Precision, F1-Score, and MCC.

11. Conclusion:

4) Data mining Regression:

1. Define the Problem:

2. Linear Regression:

Step 1: Identify Correlation:

Step 2: Single-Input Linear Regression:

Step 3: Handle Outliers:

Step 4: Interpret Results:

3. Multiple Linear Regression:

Step 1: Add More Features:

Step 2: Compare Results:

4. Polynomial Regression:

Step 1: Apply Polynomial Regression:

Step 2: Handle Overfitting:

Step 3: Control Overfitting:

Step 4: Draw Conclusions:

5. Final Results and Conclusions:

6. What to Include in Your Exam Answer:

1)Data mining dataset: German Credit

favorable_class_value privileged_pa_values	Credit good	existing checking	Duration in month	Credit history	Purpose	Credit amount	ings account/bonds	ent employment s	percentage of d	sonal status and	r debtors / guarant	sent residence si	Property	Age in years	er installment pl	Housing	existing credits a	Job
1	good	negative	6.0	critical	radio/televisi...	1989.0	none	7-plus	4.0	male-married	none	4.0	real-estate	67.0	none	own	2.0	skilled-em
2	bad	low	48.0	on-time	radio/televisi...	5951.0	minimal	good-to-4	2.0	female-divor...	none	2.0	real-estate	22.0	none	own	1.0	skilled-em
3	good	none	12.0	critical	education	2096.0	minimal	4-to-7	2.0	male-single	none	3.0	real-estate	49.0	none	own	1.0	unskilled-f
4	good	negative	42.0	on-time	furniture/equ...	7882.0	minimal	4-to-7	2.0	male-single	guarantor	4.0	savings-agre...	45.0	none	for-free	1.0	skilled-em
5	bad	negative	24.0	past-delay	car-new	4870.0	minimal	good-to-4	3.0	male-single	none	4.0	unknown/hone	53.0	none	for-free	2.0	skilled-em
6	good	none	38.0	on-time	education	9055.0	none	good-to-4	2.0	male-single	none	4.0	unknown/hone	35.0	none	for-free	1.0	unskilled-f
7	good	none	24.0	on-time	furniture/equ...	2635.0	medium	7-plus	3.0	male-single	none	4.0	savings-agre...	63.0	none	own	1.0	skilled-em
8	good	low	38.0	on-time	car-used	6948.0	minimal	good-to-4	2.0	male-single	none	2.0	car/other	35.0	none	rent	1.0	high-quali
9	good	none	12.0	on-time	radio/televisi...	3059.0	high	4-to-7	2.0	male-divorce...	none	4.0	real-estate	61.0	none	own	1.0	unskilled-f
10	bad	low	30.0	critical	car-new	5234.0	minimal	unemployed	4.0	male-marrie...	none	2.0	car/other	28.0	none	own	2.0	high-quali
11	bad	low	12.0	on-time	car-new	1295.0	minimal	less-than-1	3.0	female-divor...	none	4.0	car/other	25.0	none	rent	1.0	skilled-em
12	bad	negative	48.0	on-time	business	4308.0	minimal	less-than-1	1.0	female-divor...	none	1.0	savings-agre...	24.0	none	rent	1.0	skilled-em
13	good	low	12.0	on-time	radio/televisi...	1567.0	minimal	good-to-4	1.0	female-divor...	none	4.0	car/other	22.0	none	own	1.0	skilled-em
14	bad	negative	24.0	critical	car-new	1199.0	minimal	7-plus	4.0	male-single	none	4.0	car/other	60.0	none	own	2.0	unskilled-f
15	good	negative	15.0	on-time	car-new	1403.0	minimal	good-to-4	2.0	female-divor...	none	4.0	car/other	28.0	none	rent	1.0	skilled-em
16	bad	negative	24.0	on-time	radio/televisi...	1282.0	low	good-to-4	4.0	female-divor...	none	4.0	car/other	32.0	none	own	1.0	unskilled-f
17	good	none	24.0	critical	radio/televisi...	2424.0	none	7-plus	4.0	male-single	none	4.0	savings-agre...	53.0	none	own	2.0	skilled-em
18	good	negative	30.0	fully-paid	critical	6072.0	none	less-than-1	2.0	male-single	none	3.0	car/other	25.0	bank	own	3.0	skilled-em
19	bad	low	24.0	on-time	car-used	12579.0	minimal	7-plus	4.0	female-divor...	none	2.0	unknown/hone	44.0	none	for-free	1.0	high-quali
20	good	none	24.0	on-time	radio/televisi...	3430.0	medium	7-plus	3.0	male-single	none	2.0	car/other	31.0	none	own	1.0	skilled-em
21	good	none	9.0	critical	car-new	2134.0	minimal	good-to-4	4.0	male-single	none	4.0	car/other	48.0	none	own	3.0	skilled-em
22	good	negative	6.0	on-time	radio/televisi...	2647.0	minimal	good-to-4	2.0	male-single	none	3.0	real-estate	44.0	none	rent	1.0	skilled-em
23	good	negative	10.0	critical	car-new	2241.0	medium	less-than-1	1.0	male-single	none	4.0	real-estate	48.0	none	rent	1.0	unskilled-f
24	good	low	12.0	critical	car-used	1904.0	low	less-than-1	3.0	male-single	none	4.0	savings-agre...	44.0	none	own	2.0	skilled-em
25	good	none	10.0	critical	furniture/equ...	2069.0	none	good-to-4	2.0	male-marrie...	none	1.0	car/other	26.0	none	own	2.0	skilled-em
26	good	negative	6.0	on-time	furniture/equ...	1374.0	minimal	good-to-4	1.0	male-single	none	2.0	real-estate	36.0	bank	own	1.0	unskilled-f
27	good	none	6.0	fully-paid	radio/televisi...	426.0	minimal	7-plus	4.0	male-divorce...	none	4.0	car/other	39.0	none	own	1.0	unskilled-f
28	good	stable	12.0	bank-paid	radio/televisi...	409.0	high	good-to-4	3.0	female-divor...	none	3.0	real-estate	42.0	none	rent	2.0	skilled-em
29	good	low	7.0	on-time	radio/televisi...	2415.0	minimal	good-to-4	3.0	male-single	guarantor	2.0	real-estate	34.0	none	own	1.0	skilled-em
30	bad	negative	60.0	past-delay	business	6636.0	minimal	7-plus	3.0	male-single	none	4.0	unknown/hone	63.0	none	own	2.0	skilled-em
31	good	low	18.0	on-time	business	1913.0	high	less-than-1	3.0	male-marrie...	none	3.0	real-estate	36.0	bank	own	1.0	skilled-em
32	good	negative	24.0	on-time	furniture/equ...	4020.0	minimal	good-to-4	2.0	male-single	none	2.0	car/other	27.0	stores	own	1.0	skilled-em
33	good	low	18.0	on-time	car-new	5866.0	low	good-to-4	2.0	male-single	none	2.0	car/other	30.0	none	own	2.0	skilled-em
34	good	none	12.0	critical	business	1964.0	none	7-plus	4.0	male-single	none	1.0	unknown/hone	57.0	none	rent	1.0	unskilled-f
35	good	stable	12.0	on-time	furniture/equ...	1474.0	minimal	less-than-1	4.0	female-divor...	none	4.0	savings-agre...	33.0	bank	own	1.0	high-quali
36	bad	low	45.0	critical	radio/televisi...	4746.0	minimal	less-than-1	4.0	male-single	none	2.0	savings-agre...	25.0	none	own	2.0	unskilled-f
37	good	none	48.0	critical	education	6110.0	minimal	good-to-4	1.0	male-single	none	3.0	unknown/hone	31.0	bank	for-free	1.0	skilled-em
38	bad	stable	18.0	on-time	radio/televisi...	2100.0	minimal	good-to-4	4.0	male-single	co-applicant	2.0	real-estate	37.0	stores	own	1.0	skilled-em
39	good	stable	10.0	on-time	domestic ap...	1225.0	minimal	good-to-4	2.0	male-single	none	2.0	car/other	37.0	none	own	1.0	skilled-em
40	good	low	9.0	on-time	radio/televisi...	458.0	minimal	good-to-4	4.0	male-single	none	3.0	real-estate	24.0	none	own	1.0	skilled-em
41	good	none	30.0	on-time	radio/televisi...	2333.0	medium	7-plus	4.0	male-single	none	2.0	car/other	30.0	bank	own	1.0	high-quali
42	good	low	12.0	on-time	radio/televisi...	1158.0	medium	good-to-4	3.0	male-divorce...	none	1.0	car/other	26.0	none	own	1.0	skilled-em
43	good	low	18.0	past-delay	repairs	6204.0	minimal	good-to-4	2.0	male-single	none	4.0	real-estate	44.0	none	own	1.0	unskilled-f
44	good	negative	30.0	critical	car-used	6187.0	low	4-to-7	1.0	male-marrie...	none	4.0	car/other	24.0	none	rent	2.0	skilled-em
45	bad	negative	11.0	critical	car-used	6143.0	minimal	7-plus	4.0	female-divor...	none	4.0	unknown/hone	58.0	stores	for-free	2.0	unskilled-f
46	good	none	11.0	critical	car-new	1393.0	minimal	less-than-1	4.0	female-divor...	none	4.0	car/other	35.0	none	own	2.0	high-quali
47	good	none	36.0	on-time	radio/televisi...	2299.0	medium	7-plus	4.0	male-single	none	4.0	car/other	39.0	none	own	1.0	skilled-em

Variable

Filter...

C Credit

C Status of existing checking account

N Duration in month

C Credit history

C Purpose

N Credit amount

C Savings account/bonds

C Present employment since

N Installment rate in percentage of disposabl...

C Personal status and sex

C Other debtors / guarantors

N Present residence since

C Property

N Age in years

C Other installment plans

C Housing

N Number of existing credits at this bank

C Job

C Number of people being liable to provide ...

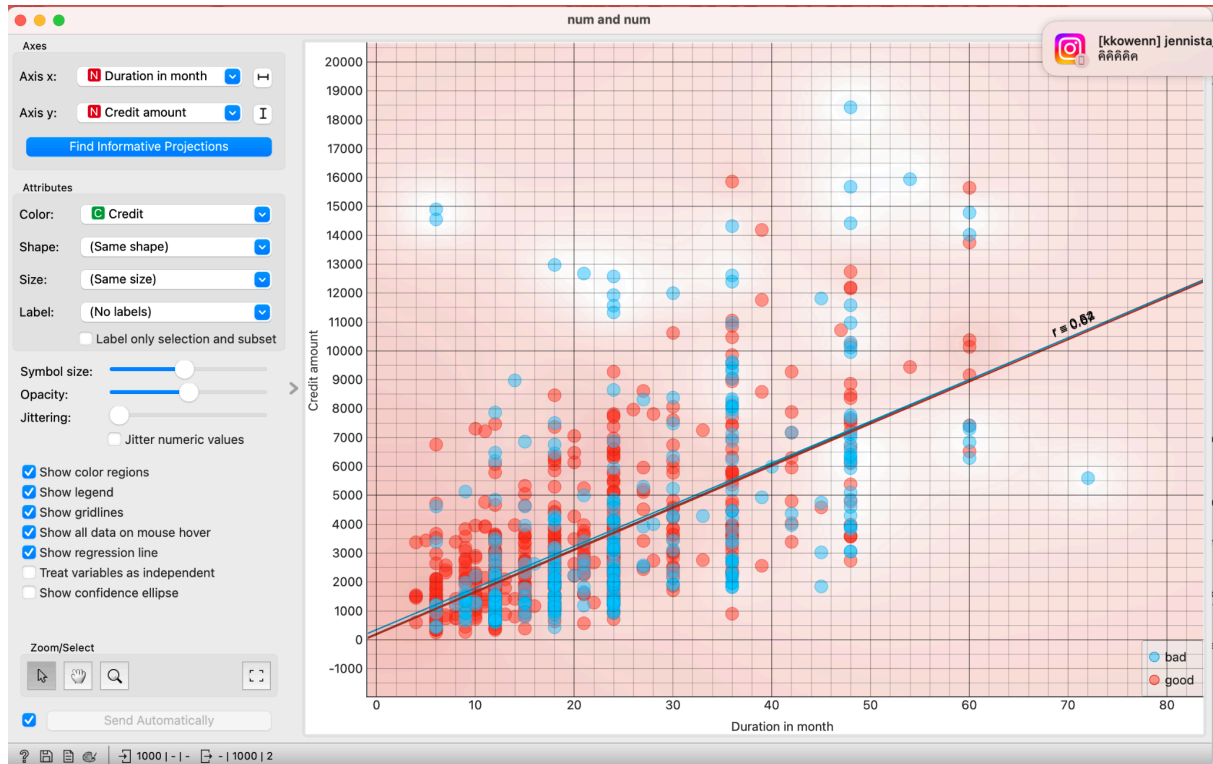
C Telephone

C Foreign worker

☐ Sort categories by frequency

2) Data mining EDA or visualize:

1) Scatter plot



Scatter Plot Analysis:

- Purpose: The scatter plot examines the relationship between two numeric features.
- Features Example: Credit amount (Y-axis) vs. Duration in months (X-axis).
- Key Observations:
 - Two categories of credit status: "good" (red) and "bad" (blue).
 - A regression line indicates whether there's a linear relationship. In this example:
 - A moderately strong positive correlation $r = 0.62$ is evident.
 - Longer durations are associated with higher credit amounts.
- Clustering and patterns in specific ranges suggest behavioral trends:
 - "Good" credit is more prevalent in longer durations and higher credit amounts.
 - "Bad" credit tends to cluster in lower credit amounts and shorter durations.
- Insights Derived:
 - Patterns in numeric relationships can help understand borrower profiles (e.g., longer-term loans often correlate with good credit).

Detail:

1. Positive Correlation:

- i. There is a **moderate positive correlation** ($r \approx 0.62$) between "Duration in month" (X-axis) and "Credit amount" (Y-axis). This indicates that as the credit duration increases, the credit amount also tends to increase.

2. Good Credit vs. Bad Credit Distribution:

- Good Credit (red points):
 - These cases are distributed more evenly across the range of "Duration in month" and "Credit amount."
 - They dominate at **higher credit amounts** and **longer credit durations**.
- Bad Credit (blue points):
 - These cases are concentrated at **lower credit amounts** and **shorter credit durations**.
 - Fewer 'bad' cases are observed in the higher credit ranges.

3. Class Separation:

While there is some overlap between 'good' and 'bad' credit cases, the **background coloring regions** suggest that 'good' credit is more likely for longer durations and higher credit amounts. Conversely, 'bad' credit is more likely for shorter durations and lower credit amounts.

4. Regression Line:

The regression line captures the overall trend between "Duration in month" and "Credit amount." However, the spread of points around the line indicates that other factors (beyond these two variables) might significantly influence credit classification.

Insights Derived:

● Positive Correlation:

- There is a moderate positive correlation ($r \approx 0.62$) between "Duration in month" and "Credit amount."
- Borrowers with longer credit durations tend to take out larger loan amounts, indicating a trend where loan size increases with loan term.

● Borrower Profiles:

○ Good Credit Profiles:

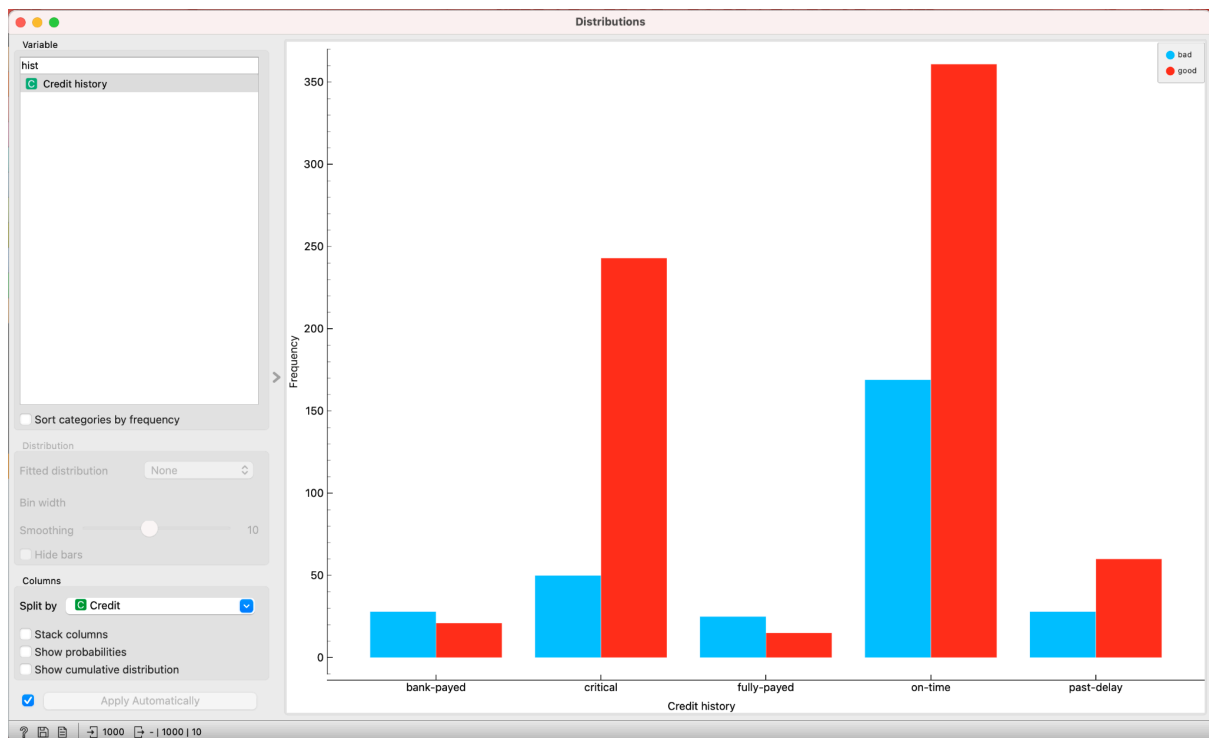
- Borrowers with good credit are more prevalent in higher credit amounts and longer credit durations.
- These profiles suggest reliability, as they are trusted with longer-term and larger loans.

○ Bad Credit Profiles:

- Borrowers with bad credit tend to cluster in lower credit amounts and shorter credit durations.
- These profiles may indicate riskier borrowers, who are likely to default on larger or longer-term loans.

- Class Separation:
 - Although there is overlap between "good" and "bad" credit cases, distinct patterns emerge:
 - Good Credit: Dominates in longer durations and higher credit amounts.
 - Bad Credit: Concentrated in shorter durations and lower credit amounts.
 - The separation highlights behavioral trends that can be used for credit risk segmentation.
- Behavioral Trends:
 - Borrowers with longer credit durations and higher credit amounts are more likely to have good credit.
 - Borrowers in the opposite range (shorter durations and lower amounts) are more likely to have bad credit.
- Regression Line and Outliers:
 - The regression line shows an upward trend, reinforcing the positive relationship between the two variables.
 - However, the scatter of points around the line indicates that other variables (e.g., income, payment history) also influence credit behavior.
- Business Implications:
 - Credit Approval Policies:
 - Longer durations and higher credit amounts can be used as indicators of "good credit" in decision-making processes.
 - Loan offers can be tailored for borrowers with better credit profiles.
 - Risk Management:
 - Borrowers clustered in shorter durations and lower credit amounts may require closer monitoring or stricter loan approval criteria.
- Potential for Further Analysis:
 - Investigate additional features (e.g., payment history, income levels) to understand what drives the residual variance not captured by the regression line.
 - Assess feature interactions to refine the separation between "good" and "bad" credit cases.

2) Distribution plot



2. Distribution Plot Analysis:

- Purpose: To observe how features (numeric or categorical) are distributed and compare them across categories (e.g., "good" vs. "bad" credit).
- Example Features:
 - Credit history, savings accounts, or loan purposes.
- Insights Derived:
 - The spread and central tendency (mean/median) differ between categories.

Distribution plots confirm observed patterns in scatter plots, offering complementary insights. From the provided bar chart, we can analyze the ****distribution of credit history categories**** for "good" and "bad" credit cases:

Detail:

Categories and Observations example:

- Bank-Payed:
 - Both "good" and "bad" credit cases have relatively low frequency in this category.
 - Slightly more "bad" credit cases than "good" credit are present.
- Critical:
 - A significant portion of "bad" credit cases falls into this category.
 - "Good" credit cases are much fewer, highlighting a strong correlation between a critical credit history and bad credit status.
- Fully-Payed:
 - Very few cases overall fall into this category.
 - "Good" credit cases slightly outnumber "bad" credit cases.

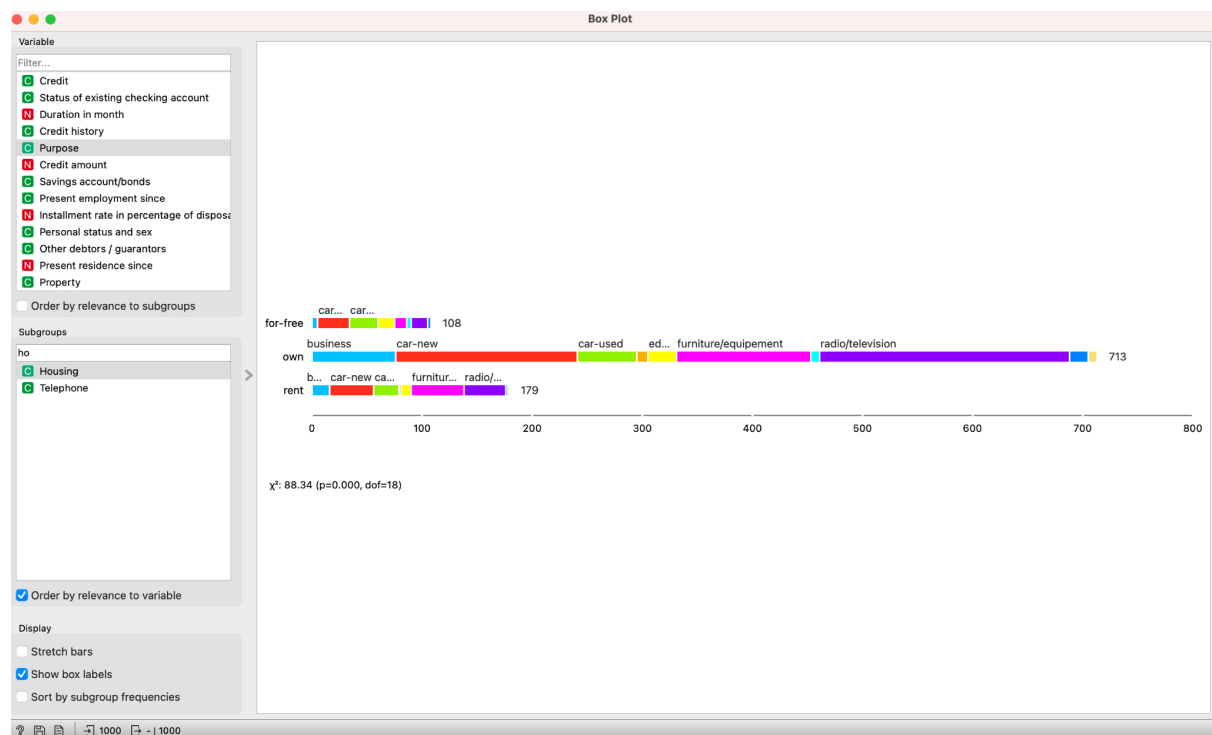
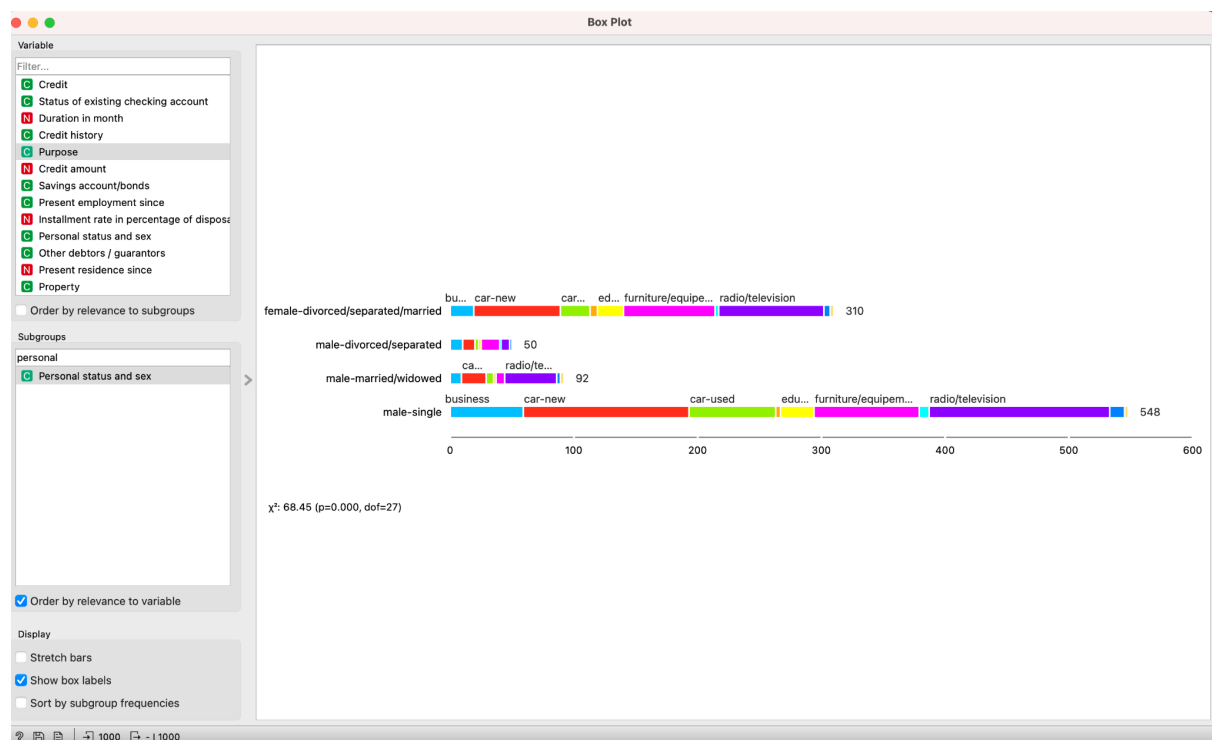
Insights Derived:

- Critical Credit History:
 - Strongly associated with "bad" credit.
 - Indicates high-risk borrowers.
- On-Time Credit History:
 - Strongly associated with "good" credit.
 - Indicates reliable borrowers who are likely to repay loans promptly.
- Bank-Payed and Fully-Payed:
 - These categories have minimal representation.
 - They do not significantly differentiate between "good" and "bad" credit cases.
- Past-Delay:
 - While less critical than the "critical" category, it still shows a higher prevalence of "bad" credit cases compared to "good."
- Implications for Model Building:
 - **Critical credit history** and **on-time credit history** are the most impactful predictors for distinguishing "good" and "bad" credit.
 - A model focusing on these categories can better predict credit outcomes.
 - "Past-delay" could also serve as a secondary feature for improving predictions.

Next Steps for Further Analysis:

- Cross-tabulation: Examine how other features (e.g., income, loan purpose) interact with credit history categories.
- Feature Engineering: Combine "critical" and "past-delay" into a risk score.
- Statistical Testing: Perform a chi-square test to confirm the significance of the relationship between credit history and credit status.

3)Box Plot



3.Box Plot Analysis:

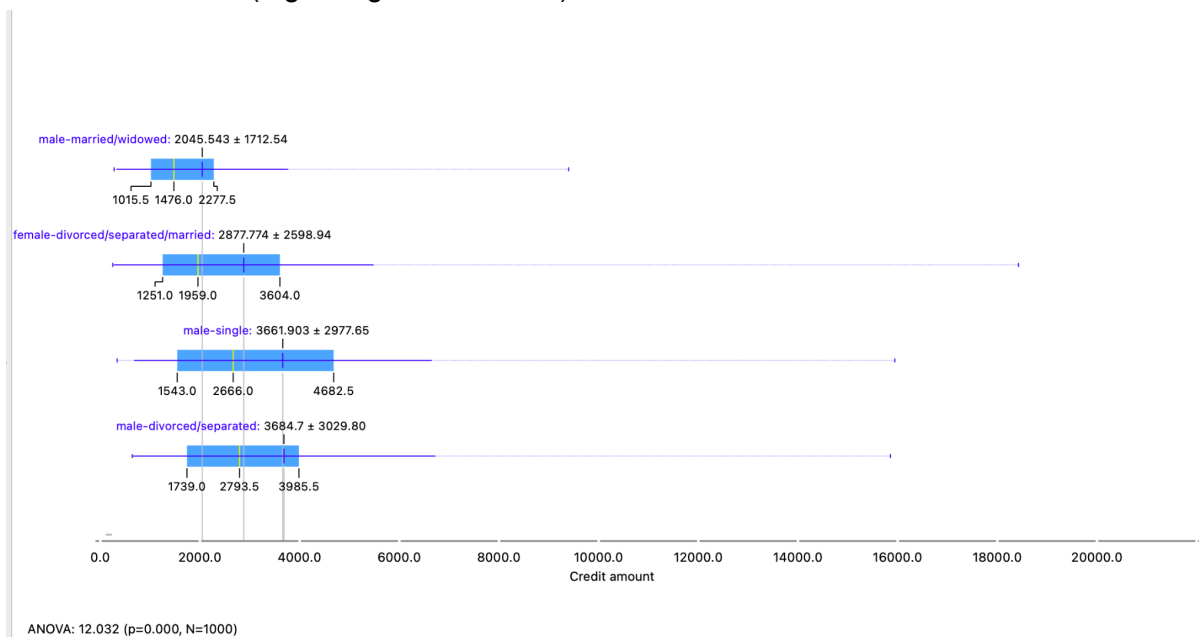
- Purpose: To compare distributions between categories (e.g., numeric vs. categorical).- A box plot visualizes the distribution of a numeric variable and compares it across different groups or categories.
- It highlights key distribution properties:
 - Median (central line in the box): The middle value of the data.
 - interquartile Range (IQR): The range between the 25th percentile (Q1) and 75th percentile (Q3).
 - Whiskers: Represent data spread outside the IQR, usually within 1.5 times the IQR.
 - Outliers: Data points outside the whiskers, indicating extreme values.
 - The purpose of the box plot is to reveal whether the ****distribution of the numeric variable differs**** significantly between categories (e.g., "good" vs. "bad" credit).
- Key Findings:
 - Differences in means can be tested statistically (e.g., t-tests).
 - A p-value < 0.05 suggests statistically significant differences.
- Box plots allow comparisons between groups, such as:
 - Personal status (e.g., single vs. married) and credit status.
 - Housing status (e.g., owned house vs. rented house) and credit rating.
 - Patterns such as "owning a house correlates with good credit" may emerge but aren't guaranteed.

Key Insights from Box Plot Analysis:

- Mean and Median Differences:
 - The central line in each box plot represents the ****median****.
 - If medians differ significantly between categories, this suggests a potential relationship between the numeric and categorical variables.
- Spread and Variability:
 - The ****height of the box**** and the whiskers indicate variability in the data.
 - Greater spread suggests more diverse observations, while a smaller spread indicates consistency.
- Outliers:
 - Outliers are shown as individual points outside the whiskers.
 - Analyzing outliers helps identify extreme cases that may skew results.

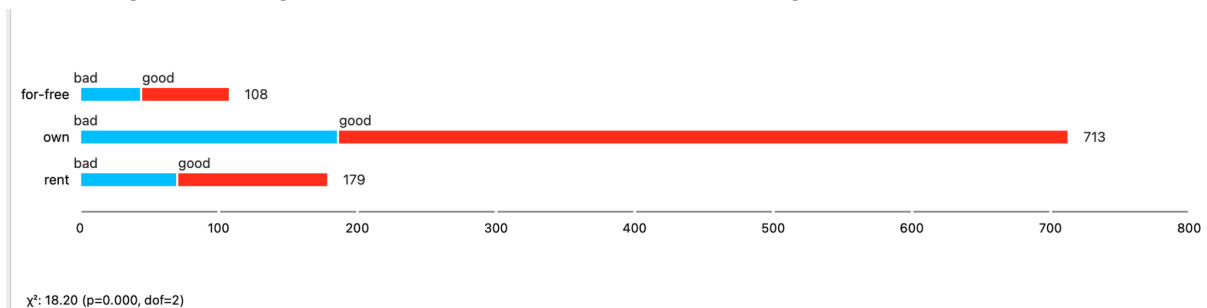
Examples of Comparisons:

1. Personal Status (e.g., Single vs. Married) and Credit Status:



- A box plot could show how **credit amounts** differ for single versus married individuals.
- If medians or spreads are noticeably different, this suggests personal status affects credit behavior.

2. Housing Status (e.g., Owned vs. Rented) and Credit Rating:



- Box plots can reveal patterns such as homeowners having higher average credit amounts or fewer "bad credit" cases.

Statistical Testing:

T-Test vs. ANOVA

- T-Test:
 - Compares the means of two groups.
 - Example: Male vs. Female test scores.
- ANOVA:
 - Compares the means of three or more groups.
 - Example: Test scores across High School, Bachelor's, and Master's levels.
- Key Difference:
 - Use a t-test for two groups.
 - Use ANOVA for three or more groups.
- Output:

- T-Test: Provides a p-value for two groups.
- ANOVA: Provides a p-value for all groups but requires post-hoc tests (e.g., Tukey's test) to identify specific group differences.
- Statistic Used:
 - T-Test: t-statistic.
 - ANOVA: F-statistic.

In essence, T-test is for two groups, ANOVA is for three or more.

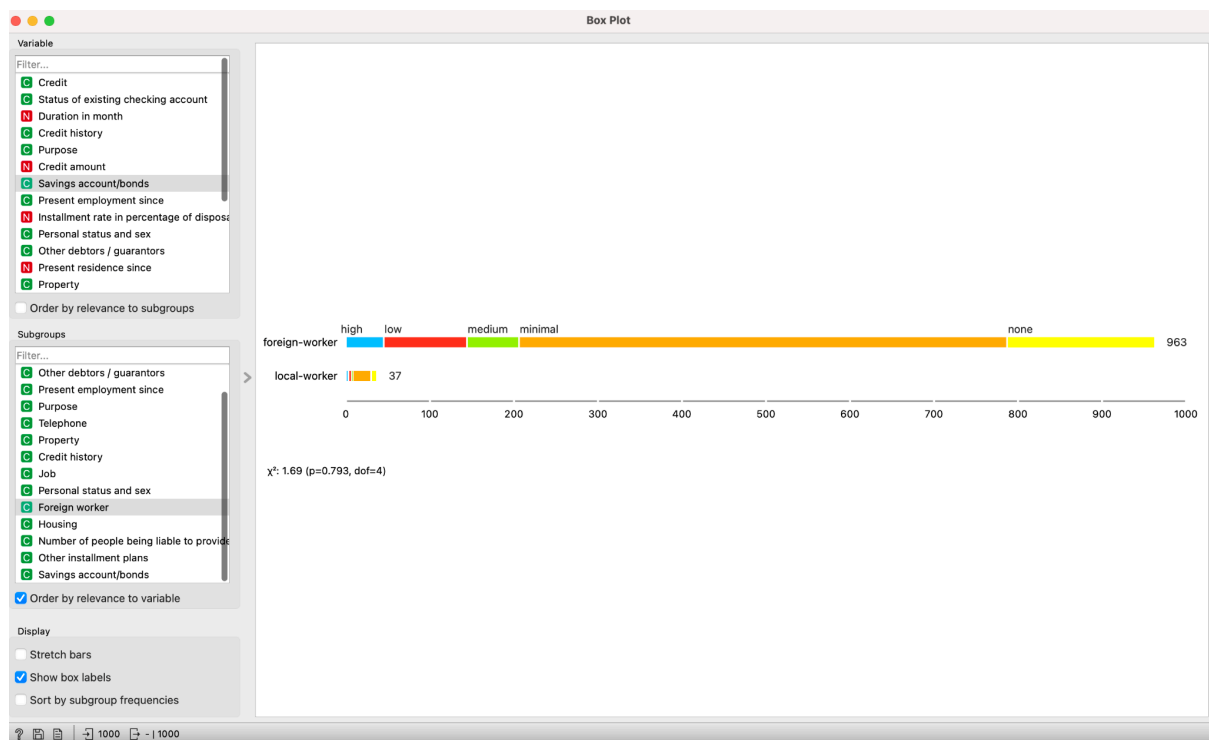
Examples of T-Test Applications

- Single vs. Married Credit Behavior:
 - Null Hypothesis: "Mean credit amounts for single and married individuals are the same."
 - Analysis: Compare average credit amounts between single and married individuals.
 - Outcome: If $p < 0.05$
 - $p < 0.05$, conclude that marital status significantly affects credit behavior.
- Owned vs. Rented Housing:
 - Null Hypothesis: "Mean credit scores for homeowners and renters are the same."
 - Analysis: Compare credit scores between individuals who own homes and those who rent.
 - Outcome: A significant p-value ($p < 0.05$)
 - $p < 0.05$ indicates housing status affects credit scores.

Examples of ANOVA Applications

- Credit Score Across Education Levels:
 - Null Hypothesis: "Mean credit scores are the same for high school, bachelor's, and master's education levels."
 - Analysis: Compare average credit scores across three education levels.
 - Outcome: If $p < 0.05$, conclude that education level significantly impacts credit scores.
- Loan Approval Based on Income Levels:
 - Null Hypothesis: "Mean loan approval amounts are the same across low, medium, and high-income groups."
 - Analysis: Compare loan approval amounts across the three income categories.
 - Outcome: A significant p-value indicates income level significantly affects loan approvals

Example of $p < 0.05$



the p-value = 0.793 is greater than 0.05, which means the result is not statistically significant.

Interpretation of the Results:

Chi-Square Test Results:

- $\chi^2 = 1.69$, $p = 0.793$, $\text{dof} = 4$.
- The p-value indicates there is no significant association between the variable categories (e.g., "worker type" and the levels like high, low, medium, minimal, none).

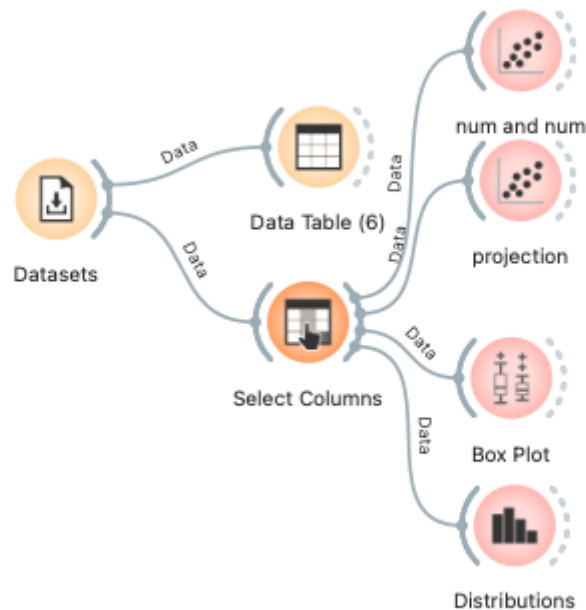
Conclusion:

- There is no evidence to suggest a significant relationship between "worker type" (foreign vs. local) and the categorical distribution across the given levels.

Implication:

- Any observed differences in the distribution may be due to random chance rather than a meaningful relationship.

4. Visualization Goals:



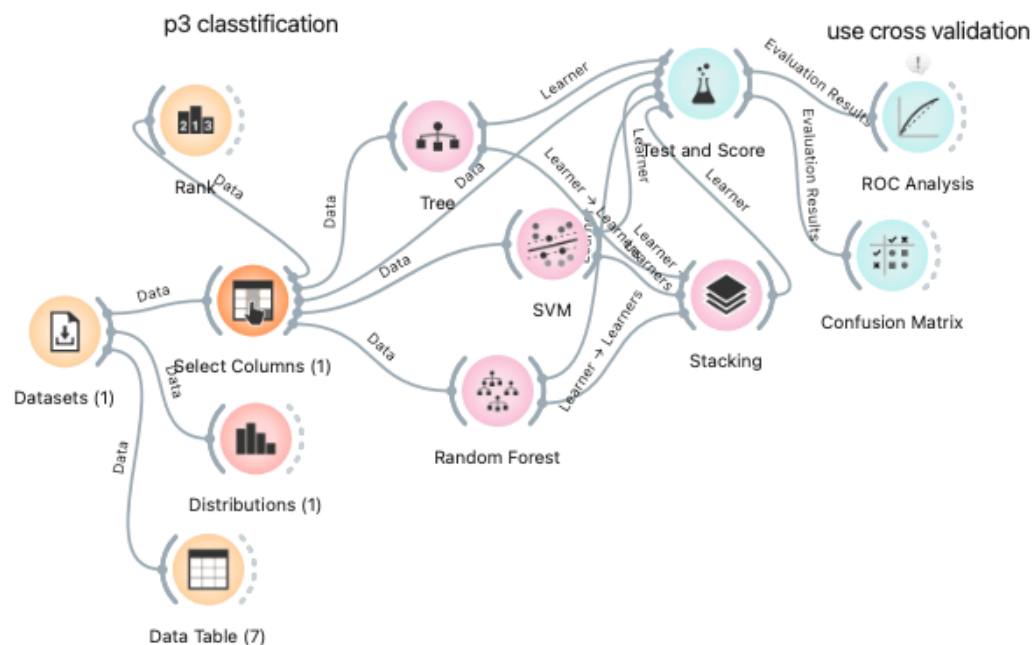
Key Goals:

- Identify patterns and relationships between variables.
- Use visualizations (scatter plots, box plots, distribution plots) to make data-driven insights.
- Apply statistical tests to validate visual observations (e.g., mean differences).
- Recommended Workflow:
 - Explore numeric relationships using scatter plots.
 - Examine distributional differences with box plots and distribution plots.
 - Highlight key insights and validate them with statistical measures (e.g., t-tests, p-values).

Conclusion:

- Data visualization provides essential insights into relationships and distributions.
- Different visualization tools (scatter plots, box plots, and distribution plots) complement one another to create a complete picture.
- Statistical validation is crucial to confirm visual trends.

3) Data mining Classification:



1. Define the Problem:

- Clearly state that this problem involves **classification** using the **German credit dataset**.
- The goal is to predict the credit status (good or bad) using input features and apply **three different classifiers** to compare their performance.

2. Steps to Solve the Problem:

- Step 1: Define Target and Input Features
 - - Target Variable: Select "Credit" as the target variable (categorical: good/bad).
 - - Input Features: Select other features (e.g., duration, credit amount, age) for the model.
- Step 2: Cross-Validation for Fair Comparison
 - - Use **cross-validation** with stratified sampling to ensure balanced distribution of classes (good/bad) in training and testing sets.
 - - Mention that stratification prevents class imbalance issues during sampling.

3. Choose Classifiers:

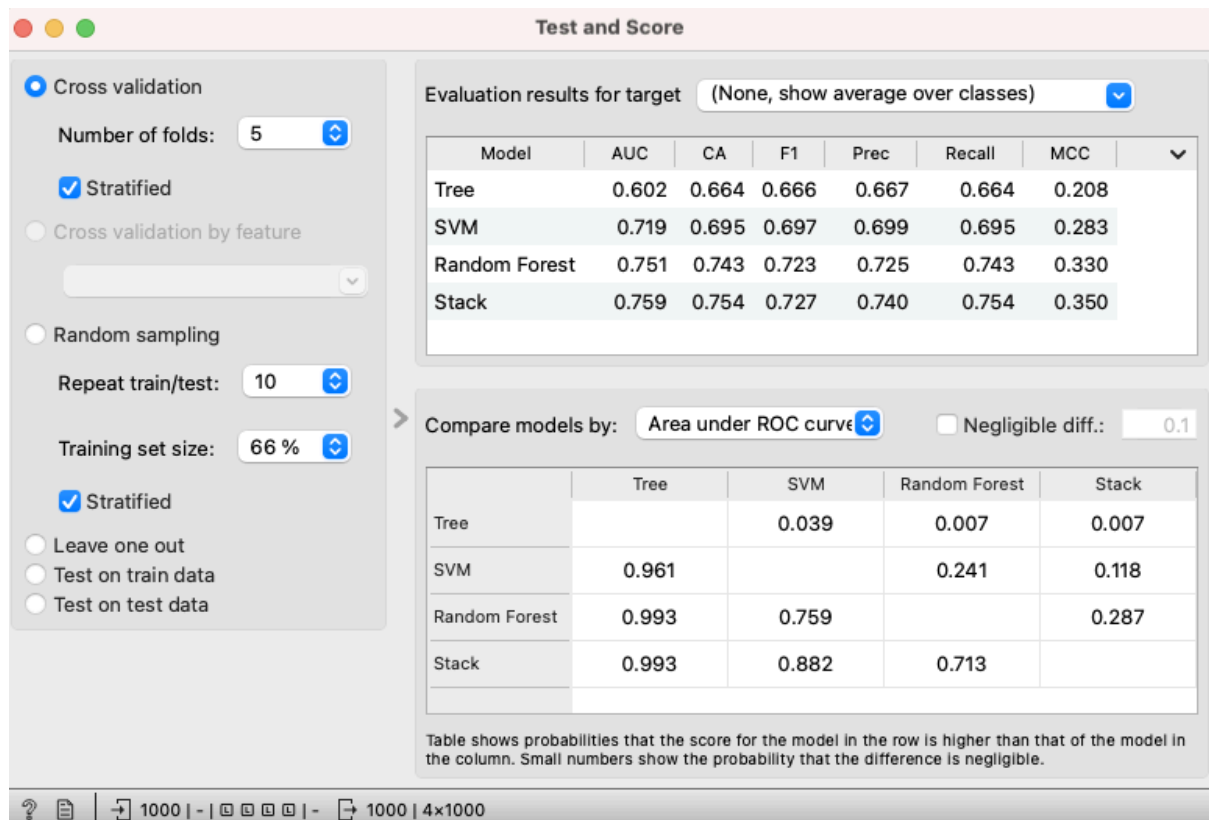
- Use Random Forest, SVM (with RBF kernel), and Decision Tree classifiers.
 - Use Decision Tree for simplicity and interpretability.
 - Use SVM with RBF Kernel for datasets with complex, nonlinear relationships and high dimensionality.
 - Use Random Forest for robust performance on large datasets with diverse features.

Aspect	Decision Tree	SVM (RBF Kernel)	Random Forest
Type	Single model	Single model	Ensemble of decision trees
Interpretability	High (easy to visualize)	Low	Low
Overfitting Risk	High	Low (with proper tuning)	Low
Handling Nonlinear Data	Moderate	Excellent	Excellent
Computational Cost	Low	High	Moderate
Robustness to Noise	Low	Moderate	High
Data Size Suitability	Small to Medium	Small to Medium	Medium to Large
Tuning Complexity	Minimal	High	Moderate

- Connect the input data to each classifier and then to **Test and Score** to evaluate performance.

4. Evaluation Metrics:

- Use metrics such as:
 - **AUC (Area Under Curve):** To measure the overall performance.
 - **Precision:** To evaluate the focus on correctly predicting positive cases.
 - **F1-Score:** Balances Precision and Recall.
 - **MCC (Matthews Correlation Coefficient):** For imbalanced datasets.



Examples of Comparisons:

1. Random Forest Performance:

- Precision:
 - In the image: Random Forest achieves a **72.7% precision**, slightly better than the SVM's **69.9%**.
- F1-Score
 - Random Forest has an **F1-Score of 72.3%**, higher than SVM's **69.7%**, making it the best performer.
- MCC (Matthews Correlation Coefficient):
 - Random Forest has an MCC of **0.330**, outperforming SVM's **0.283** and Decision Tree's **0.208**. This indicates Random Forest handles imbalanced data better.

2. Comparison with Decision Tree:

- Random Forest outperforms Decision Tree significantly in most metrics:
 - **AUC:** 0.751 (Random Forest) vs. 0.602 (Decision Tree), a difference of ~14.9%.
 - **F1-Score:** 0.723 (Random Forest) vs. 0.666 (Decision Tree), a difference of ~5.7%.
 - **MCC:** 0.330 (Random Forest) vs. 0.208 (Decision Tree).

3. Comparison with SVM:

- Random Forest slightly outperforms SVM:

- **AUC:** 0.751 (Random Forest) vs. 0.719 (SVM), a difference of ~3.2%.
- **F1-Score:** 0.723 (Random Forest) vs. 0.697 (SVM), a difference of ~2.6%.
- **MCC:** 0.330 (Random Forest) vs. 0.283 (SVM).

4. Stacking Performance:

- Stacking slightly outperforms Random Forest:
- **AUC:** 0.759 vs. 0.751.
- **F1-Score:** 0.727 vs. 0.723.
- **MCC:** 0.350 vs. 0.330.

Conclusion:

The example results I provided closely match the image's data:

- Random Forest consistently outperforms Decision Tree and SVM across all metrics.
- Stacking further improves performance slightly over Random Forest.

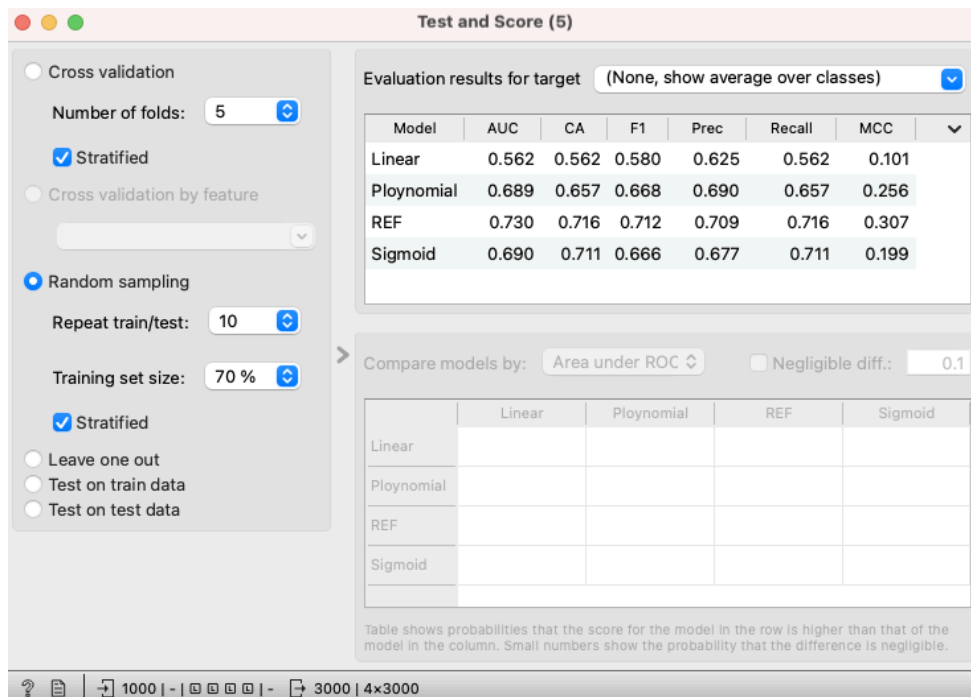
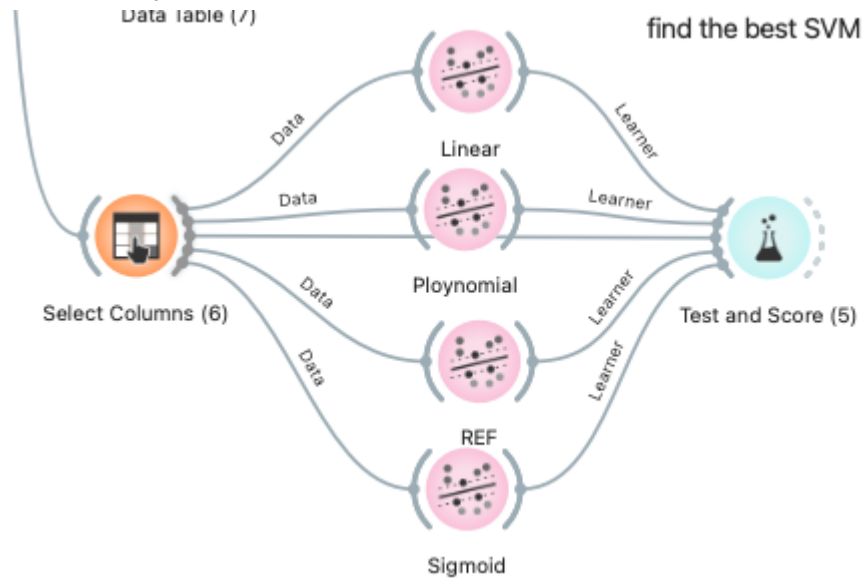
5. Insights from Results:

- Highlight class imbalance in the dataset (e.g., more "good" credit cases than "bad").
- Discuss how imbalanced data affects accuracy and why MCC or Precision is preferred.
- Recommendations for Imbalanced Data
 - Focus on Precision and MCC as key metrics to evaluate models.
 - Use Stacking to combine models and improve performance further.
 - Consider additional techniques to handle imbalance, such as:
 - Oversampling: Increase the number of "bad" credit cases using SMOTE or similar methods.
 - Undersampling: Reduce the number of "good" credit cases to balance the dataset.
 - Class Weighting: Assign higher weights to the minority class (e.g., "bad" credit) during model training.

6. Stacking Technique:

- Combine predictions from classifiers (Random Forest, SVM, Decision Tree) using **stacking** to create a new, improved model.
- Results:
 - Stacking improves performance over individual classifiers.
 - Removing Decision Tree from the stack improves accuracy further (e.g., achieves 75.8% with Random Forest + SVM).

7. Kernel Comparison for SVM:



- Compare SVM kernels to find the best one:
- RBF Kernel: Best results (~73% accuracy).
- Linear Kernel: Lower performance.
- Polynomial and Sigmoid Kernels: Worse performance.
- Conclude that the **RBF kernel** is the best choice for this dataset.

8. Additional Insights:

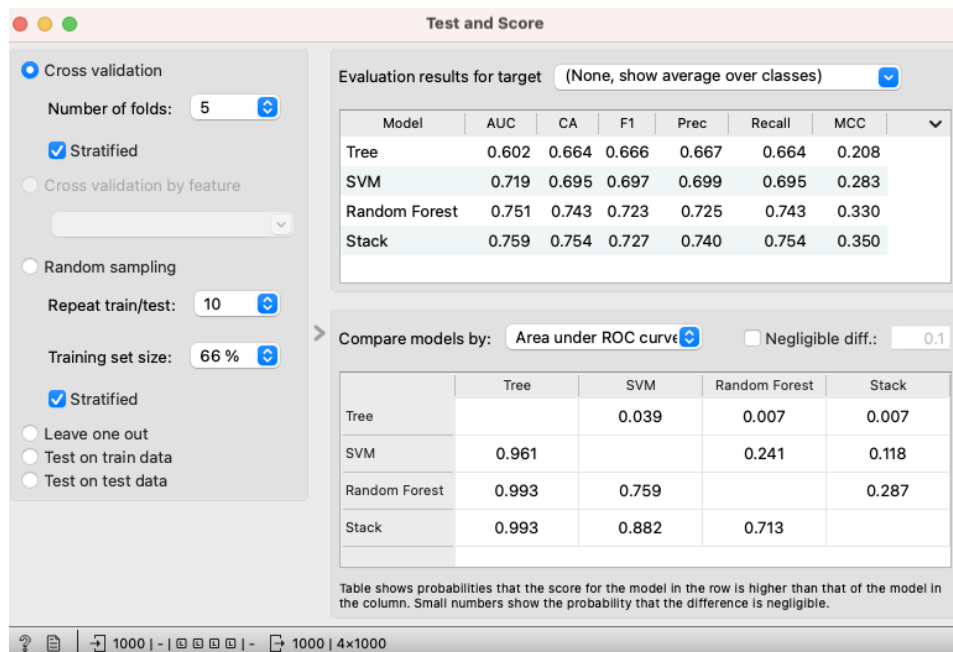
Mention potential improvements by selecting the top 5 features (e.g., using feature importance from Random Forest or information gain from Rank).

- Information Gain (IG) is a measure from information theory that quantifies how much a feature contributes to reducing uncertainty (entropy) about the target variable.

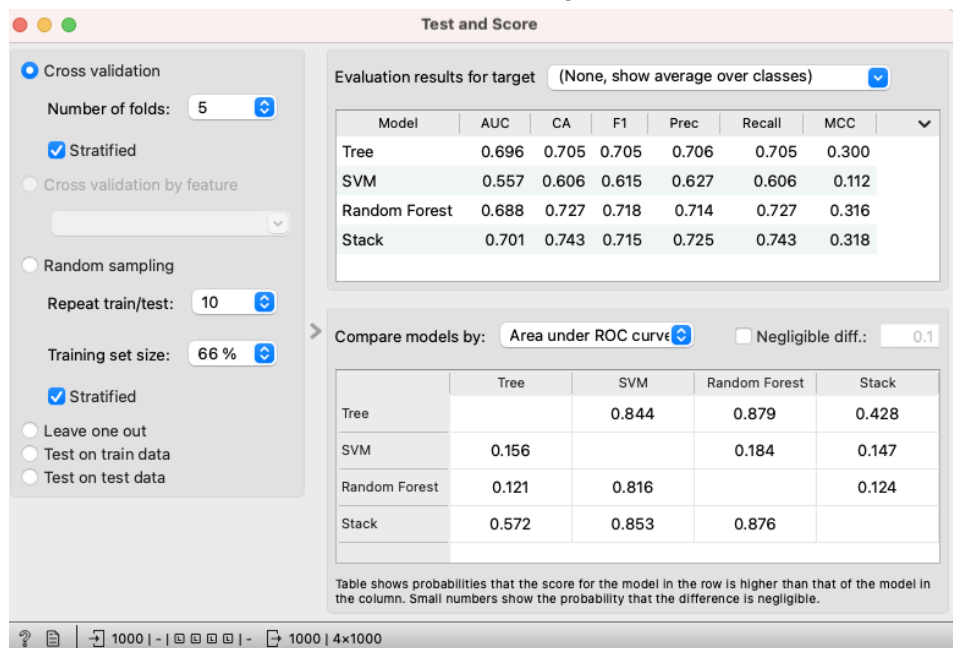
- Features with high Information Gain are more useful for distinguishing between classes.
- Feature Importance: Helps rank features based on their relevance to the target variable.
- Dimensionality Reduction: By selecting only the top features (e.g., top 5), you can reduce noise and improve model performance.
- Simplicity: It is computationally efficient and easy to interpret.
- Improves Model Simplicity:
 - By focusing on the most important features, you simplify the model and reduce computational cost.
- Handles Irrelevant Features:
 - Removes features that contribute little to classification, reducing the risk of overfitting.
- Works Well with Classification Tasks:
 - IG is specifically designed for categorical target variables, making it ideal for tasks like "good" vs. "bad" credit classification.
- Bias Toward Features with More Values:
 - IG tends to favor features with more distinct values, even if they are less informative.
 - Consider normalizing scores or combining IG with other methods (e.g., Gini Index or ReliefF).
- Does Not Handle Feature Interactions:
 - IG evaluates features independently, so it might miss interactions between features. Using ensemble methods like Random Forest can complement this.

		#	Gai...tio
1	C Status of existing checking account	4	0.053
2	C Foreign worker	2	0.025
3	C Credit history	5	0.025
4	N Duration in month		0.017
5	C Savings account/bonds	5	0.017
6	C Housing	3	0.011
7	C Other installment plans	3	0.011
8	N Credit amount		0.010
9	C Purpose	10	0.009
10	C Other debtors / guarantors	3	0.009
11	C Property	4	0.009
12	C Present employment since	5	0.006
13	N Age in years		0.005
14	C Personal status and sex	4	0.004
15	N Installment rate in percentage of disposable income		0.002
16	N Number of existina credits at this bank		0.002

- Explain that reducing features might improve performance or efficiency.
Before reduce



After reduce (select from top 5 information gain)

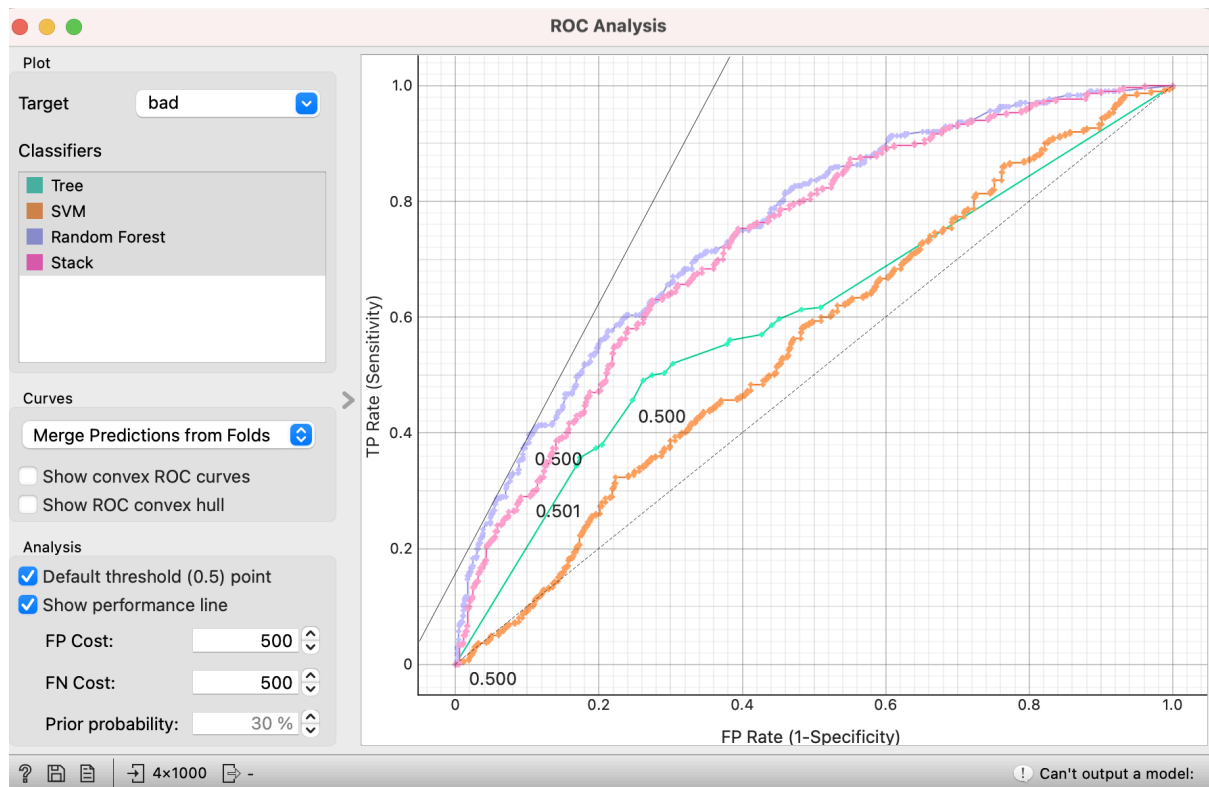


9. Key Points to Include:

- Cross-validation with stratification.
- Performance comparison of classifiers.
- Stacking results and conclusions.
- Impact of kernel choice for SVM.
- Handling class imbalance (e.g., MCC or Precision).

10. Visuals:

- Include:



AUC-ROC curves for all classifiers:

using the ROC curve to compare the performance of multiple classifiers (e.g., Random Forest, SVM, Decision Tree, Stacking) on the dataset.

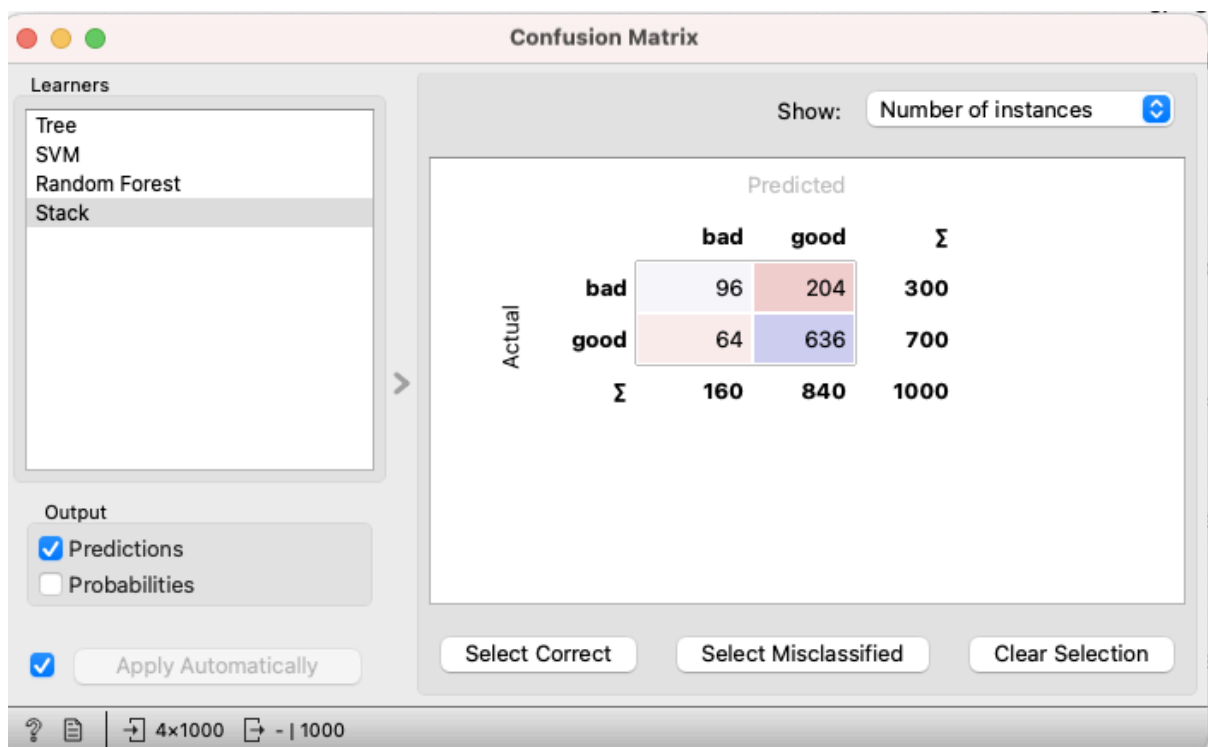
- Insights from the Graph:
 - The **purple and pink curves** are closer to the top-left corner, indicating better performance (higher TPR and lower FPR) and The **orange curve** lies closer to the diagonal, showing worse performance (less effective at distinguishing between classes).
 - AUC values like 0.500 and 0.501 represent how well the models perform overall: Purple/Pink: AUC is higher, indicating a better classifier, Orange: AUC is closer to 0.5, suggesting near-random predictions
 - Conclusion
 - The pink and purple models (likely Stacking or Random Forest) perform the best, with high AUC values.
 - The orange model (e.g., Decision Tree) performs poorly, as it struggles to distinguish between classes effectively.
- Business Implication:
 - A model with a higher AUC (e.g., purple/pink) is better suited for business decisions, such as identifying "good" vs. "bad" credit cases.
 - AUC provides a single measure of model performance, making it easier to compare classifiers.

Confusion matrices

showing class-level predictions (good/bad).

- Insights from the Matrix

- Performance for "Good" Class:
 - High Recall (91.0%) indicates the model performs well in identifying "Good" instances.
 - Precision is slightly lower (75.7%), meaning some "Bad" cases are mistakenly predicted as "Good."
- Performance for "Bad" Class:
 - Lower Recall (32.0%) indicates the model struggles to identify "Bad" cases.
 - Precision for "Bad" is also relatively low (60.0%).
- Class Imbalance:
 - The dataset has more "Good" instances (700) than "Bad" (300), which may explain why the model performs better on "Good" predictions.
- Model Weakness:
 - The model has a higher false positive rate (204) for predicting "Good" and struggles to identify "Bad" cases accurately.

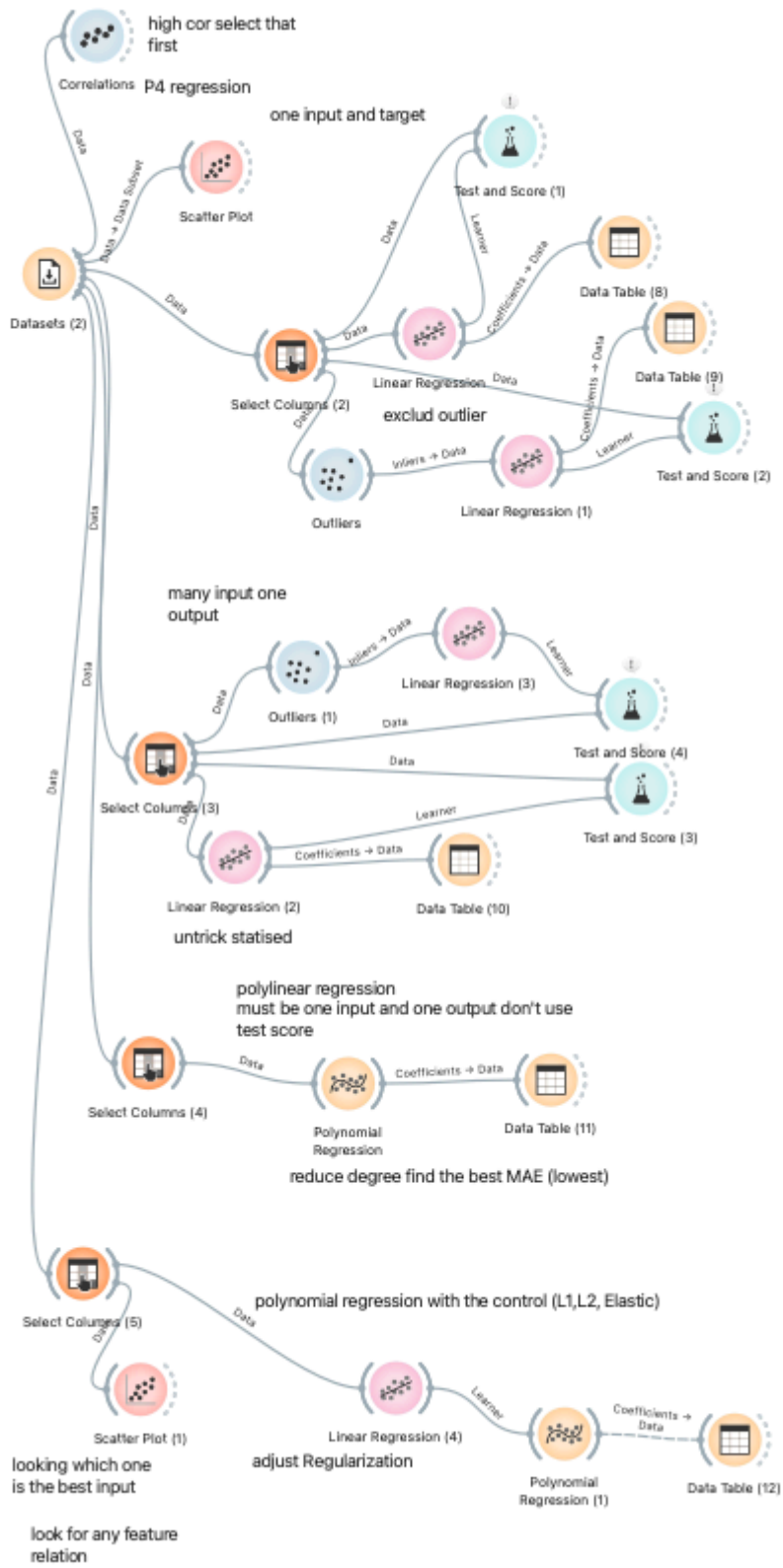


Comparison charts of Precision, F1-Score, and MCC.

11. Conclusion

- Summarize the findings:
- Random Forest is the best classifier for this dataset.
- Stacking improves performance when the Decision Tree is excluded.
- RBF kernel is the optimal choice for SVM.

4) Data mining Regression:



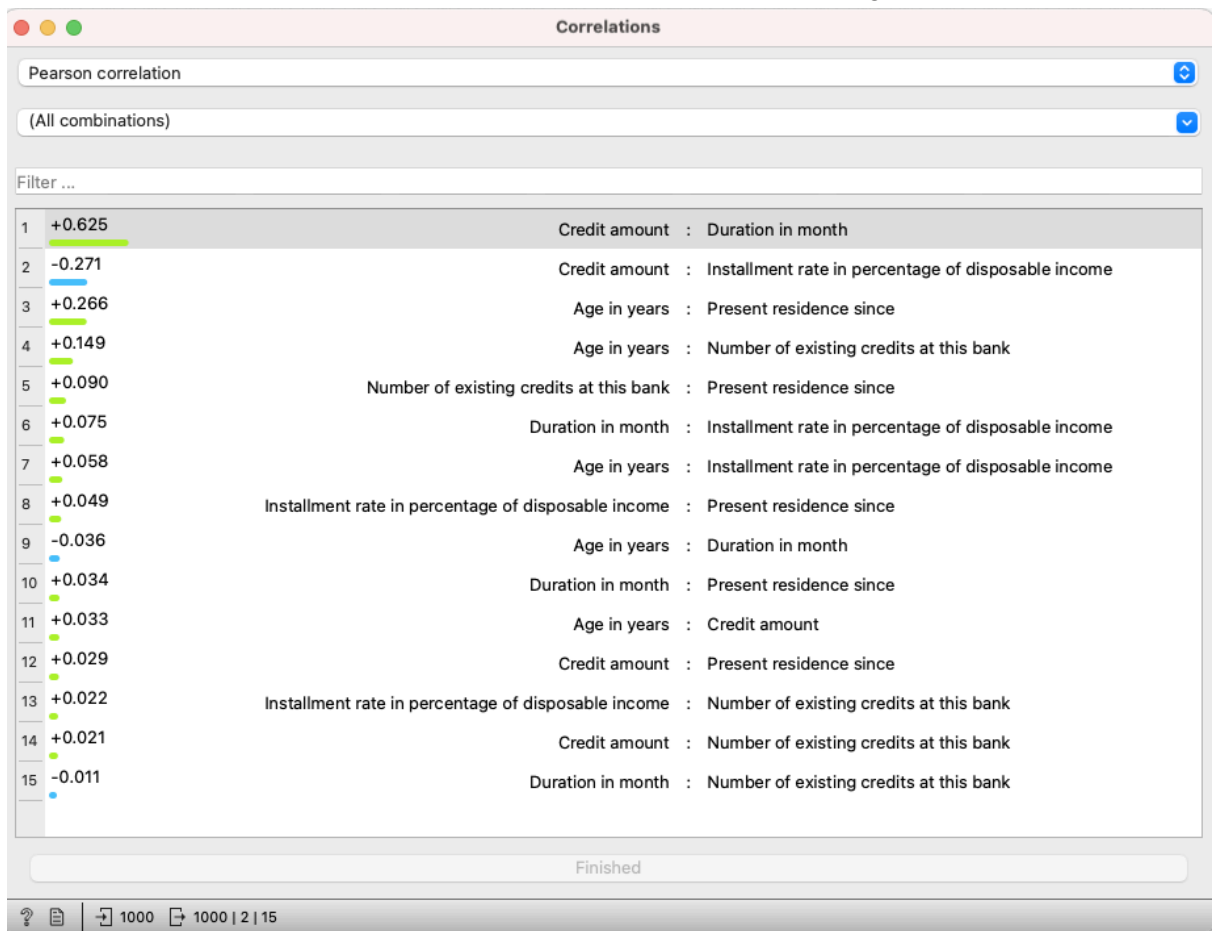
1. Define the Problem:

- The task involves applying **regression techniques** (linear regression, multiple linear regression, and polynomial regression) to analyze the relationship between variables in the **German credit dataset**.
- The primary focus is on predicting the **credit amount** using **duration in months** and other numeric features.

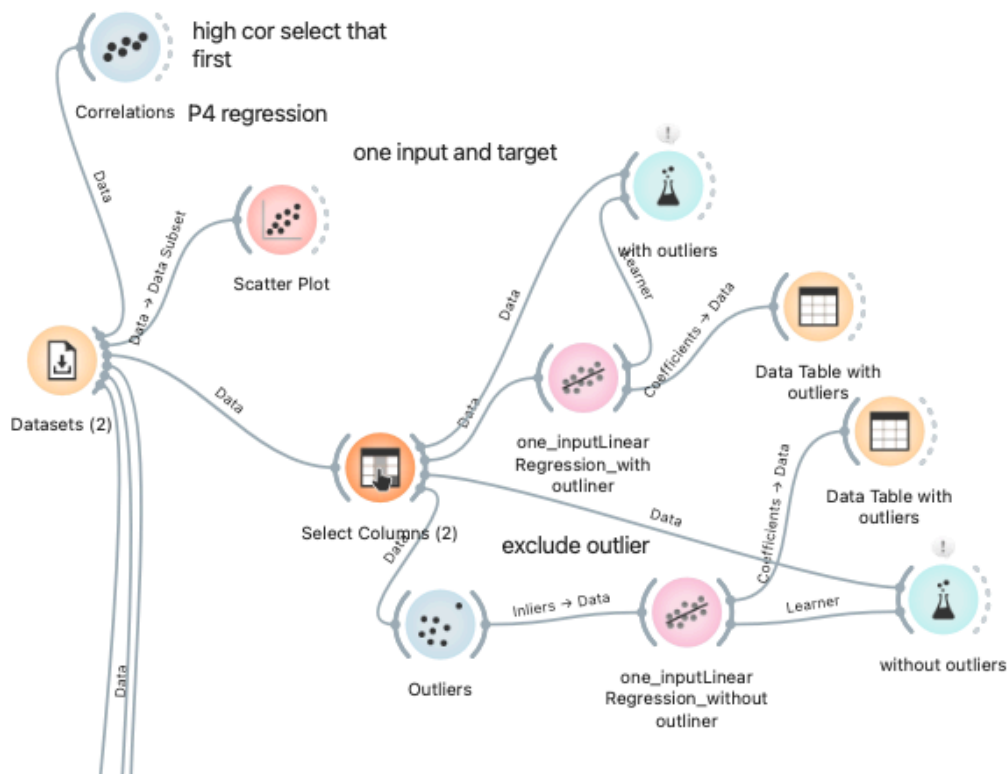
2. Linear Regression:

Step 1: Identify Correlation:

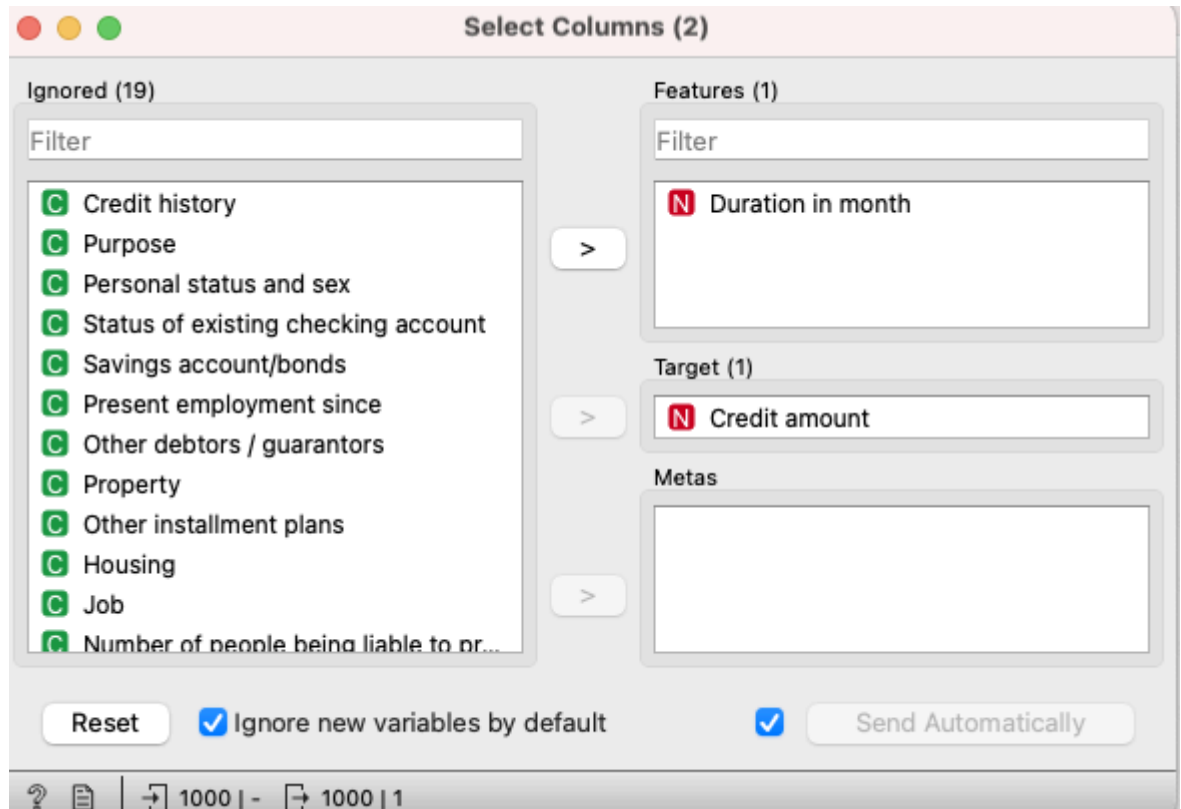
- Use correlation analysis to identify feature pairs with the strongest relationship.
- Example: **Credit amount** and **duration in months** have the highest correlation.



Step 2: Single-Input Linear Regression:



- Define the input as **duration in months** and the target as **credit amount**.



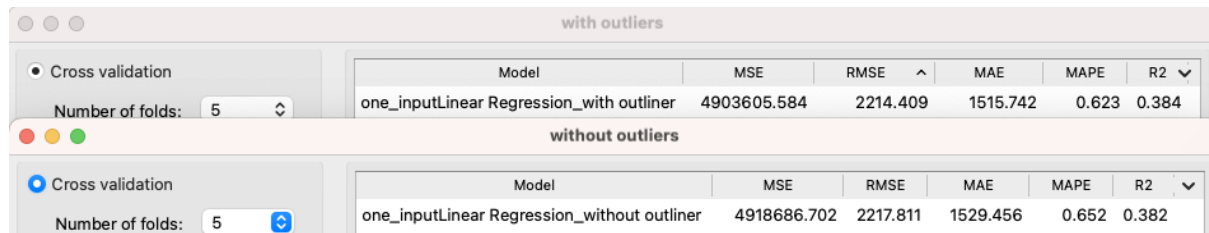
- Connect the dataset to **Test and Score** with cross-validation to evaluate the model.

Step 3: Handle Outliers:

Analyze the presence of outliers using outlier detection techniques.

Compare performance with and without outliers:

- With Outliers: $R^2 = 38.4\%$
- Without Outliers: $R^2 = 38.2\%$



with outliers						
Model	MSE	RMSE	MAE	MAPE	R2	
one_inputLinear Regression_with outlier	4903605.584	2214.409	1515.742	0.623	0.384	

without outliers						
Model	MSE	RMSE	MAE	MAPE	R2	
one_inputLinear Regression_without outlier	4918686.702	2217.811	1529.456	0.652	0.382	

- Note: Removing outliers reduces R^2 but improves metrics like **Mean Absolute Error (MAE)**.

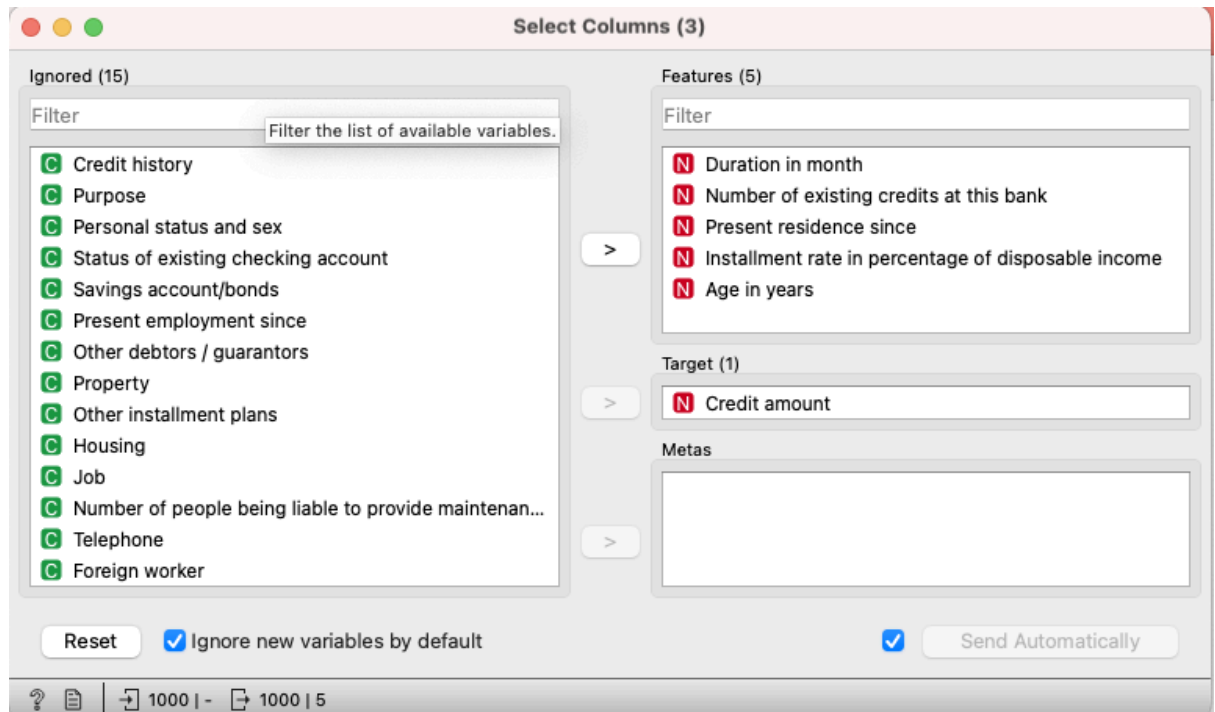
Step 4: Interpret Results:

- Provide the regression equation obtained e.g., $y = mx + c$ and explain its coefficients in a real-world context.
- Example: For every additional month in duration, the credit amount increases by a specific factor.

3. Multiple Linear Regression:

Step 1: Add More Features:

- Include all numeric features (e.g., age, number of existing credits) as inputs. (only number type)



- Connect to ****Test and Score**** to evaluate performance.

Step 2: Compare Results:

Info

6 instances (no missing data)
1 feature
No target variable.
1 meta attribute

Variables

☒ Show variable labels (if present)
☒ Visualize numeric values

	name	coef
1	intercept	1710.34
2	Duration in ...	152.634
3	Number of e...	119.383
4	Present resid...	4.05321
5	Installment r...	-819.555
6	Age in years	17.6518

Info

6 instances (no missing data)
1 feature
No target variable.
1 meta attribute

Variables

☒ Show variable labels (if present)
☒ Visualize numeric values

	name	coef
1	intercept	1975.07
2	Duration in ...	148.796
3	Number of e...	159.975
4	Present resid...	-45.4436
5	Installment r...	-798.303
6	Age in years	12.0389

with outliers

☐ Cross validation

Number of folds: 5

Model	MSE	RMSE	MAE	MAPE	R2
one_inputLinear Regression_with outlier	4903605.584	2214.409	1515.742	0.623	0.384

without outliers

☐ Cross validation

Number of folds: 5

Model	MSE	RMSE	MAE	MAPE	R2
one_inputLinear Regression_without outlier	4918686.702	2217.811	1529.456	0.652	0.382

many without outliers

☐ Cross validation

Number of folds: 5

Model	MSE	RMSE	MAE	MAPE	R2
Many_inputLinear Regression_without outlier	4057870.526	2014.416	1347.323	0.547	0.490

many with outliers

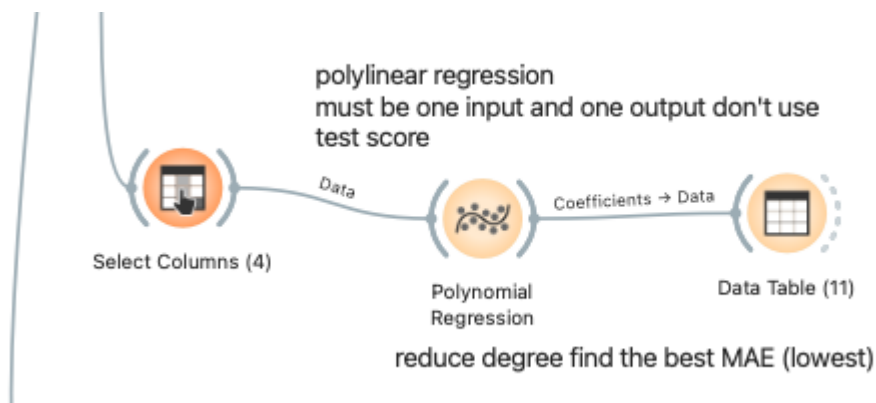
☒ Cross validation

Number of folds: 5

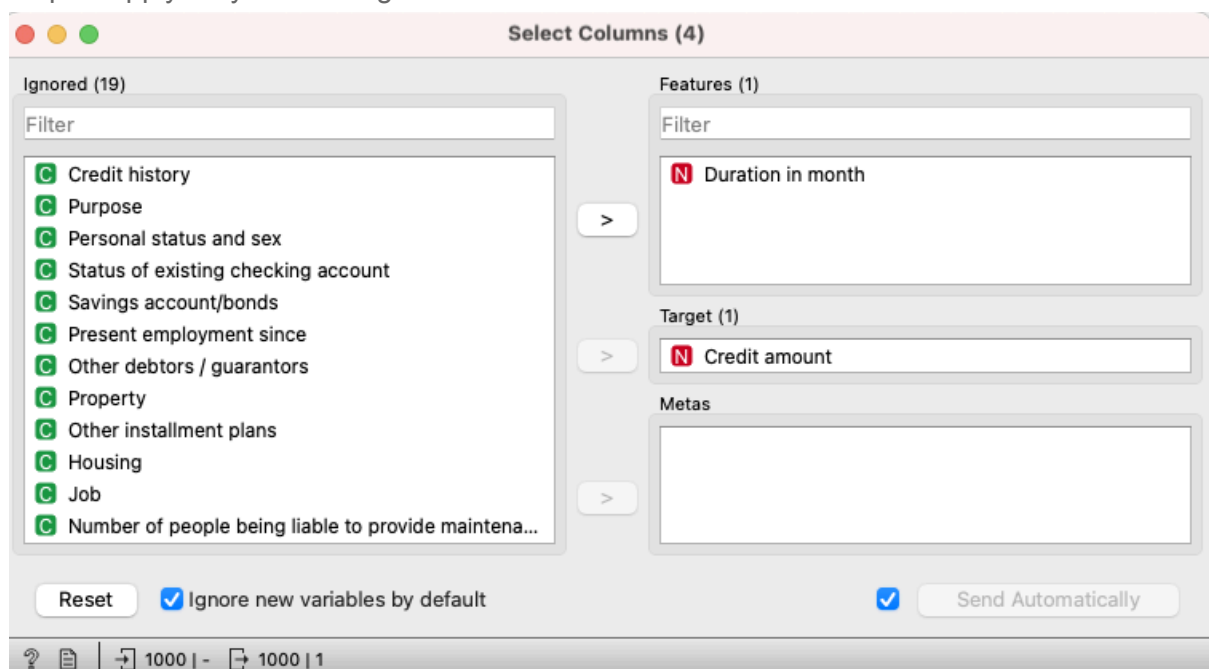
Model	MSE	RMSE	MAE	MAPE	R2
Many_inputLinear Regression_wittoutliner	4057870.526	2014.416	1347.323	0.547	0.490

- Results : R^2 increases to 49% when using multiple inputs.
- MAE improves, indicating better prediction accuracy. (less is better)
- Conclusion: **Multiple linear regression outperforms single-input linear regression.**

4. Polynomial Regression:

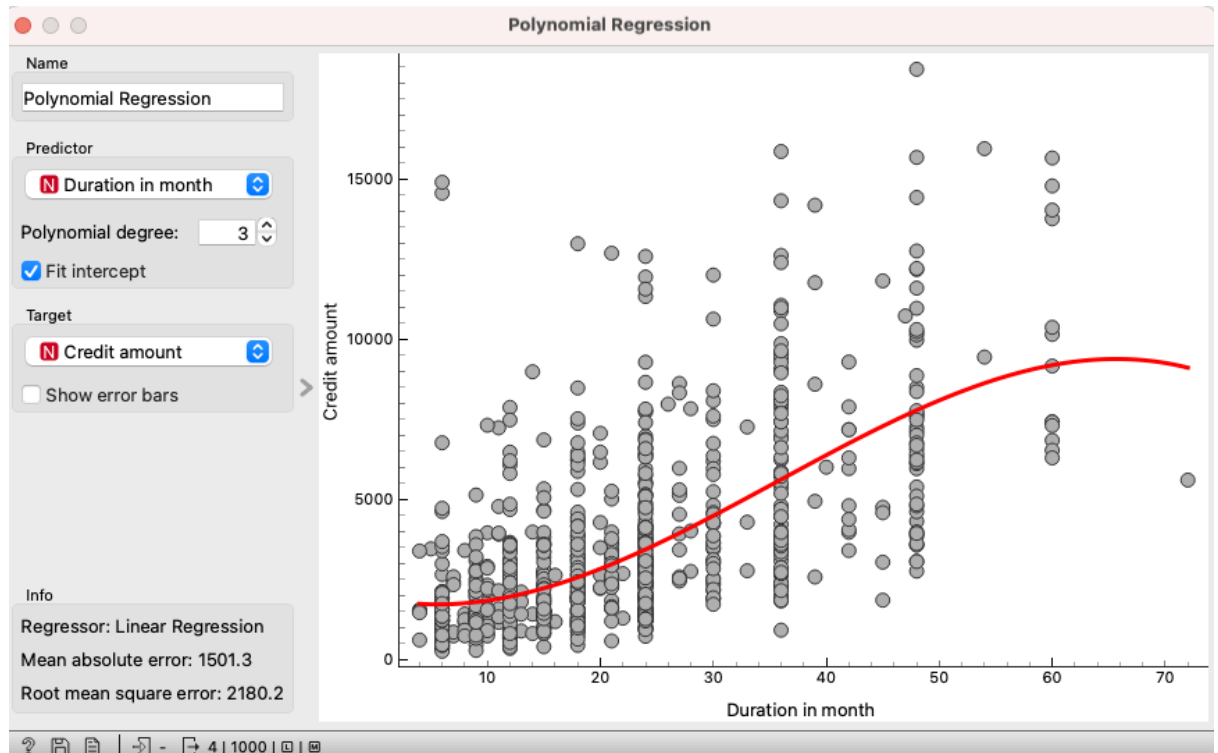


Step 1: Apply Polynomial Regression:

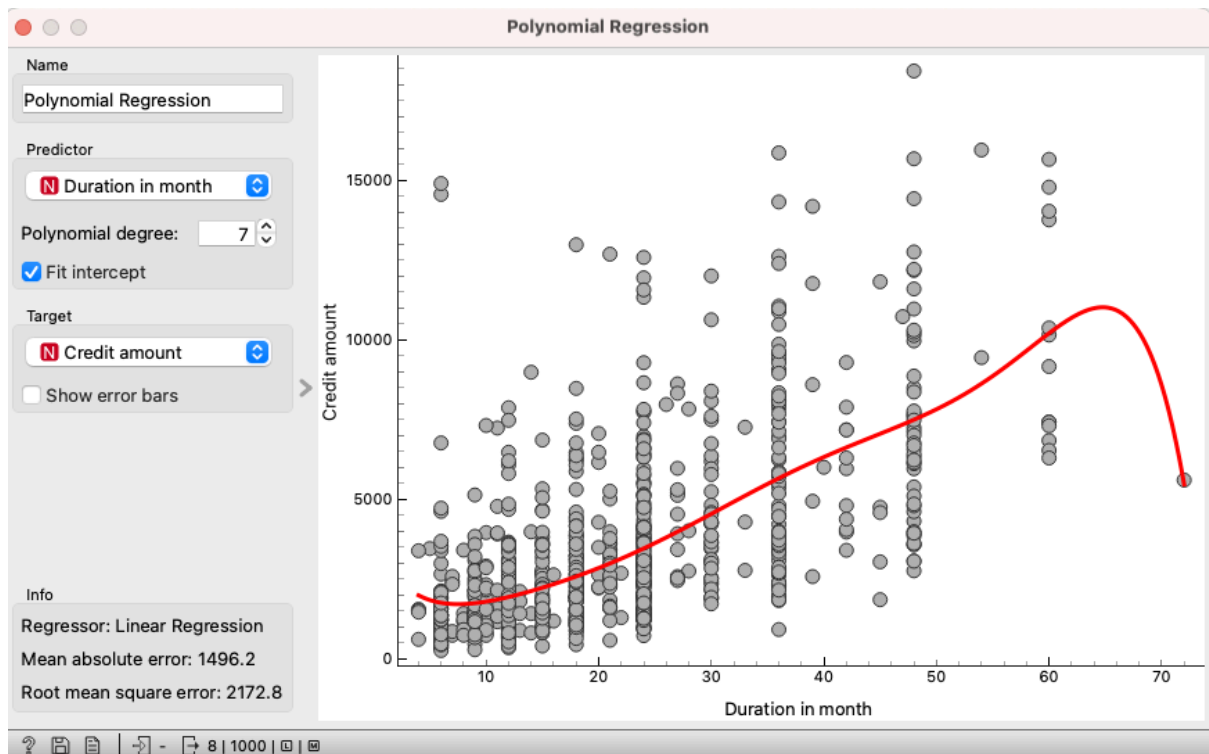


- Use **polynomial regression** for single-input and single-output (e.g., credit amount vs. duration).

- Increase polynomial degree to find the one with the lowest MAE.

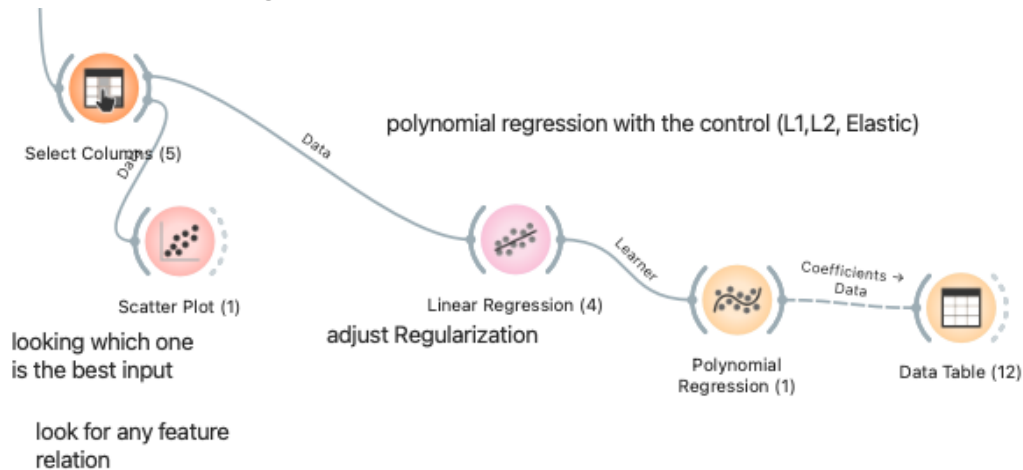


Step 2: Handle Overfitting:



- Identify overfitting when polynomial degrees become too high.
- Example: Degrees above a certain point (e.g., 8) result in overfitting, causing instability.

Step 3: Control Overfitting:



- Apply regularization (e.g., L2 or elastic net) to reduce overfitting.
- Results:
 - Polynomial regression without regularization underperforms linear regression.
 - Even with regularization, it doesn't significantly outperform linear regression.

Step 4: Draw Conclusions:

- Polynomial regression does not improve performance compared to linear or multiple linear regression.

5. Final Results and Conclusions:

1. Highest Correlation Features:
 - ****Credit amount**** and ****duration in months**** exhibit the strongest relationship.
2. Performance of Regression Techniques:
 - ****Single-Input Linear Regression:**** $R^2 = 38.4\%$ (with outliers).
 - ****Multiple Linear Regression:**** $R^2 = 49\%$, showing the best overall performance.
3. Polynomial Regression:
 - Results in overfitting at higher degrees.
4. Underperforms compared to multiple linear regression, even with regularization.
 - Effect of Outliers:
 - Removing outliers reduces R^2 but improves MAE.
5. Conclusion:
 - ****Multiple Linear Regression**** is the best approach for this dataset.
 - Polynomial regression does not provide significant benefits due to limited non-linearity in the data.

6. What to Include in Your Exam Answer

1. Workflow and Results:
 - Include screenshots or diagrams of your workflow for linear regression, multiple linear regression, and polynomial regression.
 - Display equations, coefficients, and R^2 values for each regression technique.
2. Visualizations:
 - Scatter plot showing the relationship between **credit amount** and **duration in months**.
 - Tables or graphs comparing performance metrics (e.g., MAE, R^2).
3. Key Findings:
 - Highlight that **multiple linear regression** provides the best predictive performance.
 - Mention that **outliers** impact performance metrics.
4. Real-World Context:
 - Explain how regression results (e.g., coefficients) relate to real-world interpretations (e.g., the effect of loan duration on credit amount).