RESEARCH UNIVERSITY
EXCELLENCE INITIATIVE

Akademia Górniczo-Hutnicza
im. Stanisława Staszica w Krakowie

AGH University of Krakow

Corporate Credit Risk

Mateusz Mglej, Łukasz Obrzut, Dominik Sobczyk, Faustyna Święch

WMS

Project overview

The goal of this project is to model and evaluate credit default risk using statistical and data-driven methods.

1. Working with data
2. Modelling & performance
3. Logistic regression
4. Probit specification
5. Linear regression
6. Experts model
7. Financial holding type
8. Summary

AGH

Part I

Working with data

www.agh.edu.pl

Dataset overview

- **5804 observations, 12 variables**
- Focus: credit risk assessment of companies
- Each row represents one customer assessment

Identification and context:

- `CUSTOMER_ID` – customer identifier
- `ASSESSMENT_YEAR` – year of evaluation
- `GROUP_FLAG` – belongs to a financial holding (0/1)
- `INDUSTRY` – industry classification
- `TURNOVER` – reported turnover

Key variables: expert scores and target

Expert scores (scale 10–90):

- `PRODUCT_DEMAND` – market position
- `OWNERS_MANAGEMENT` – management quality
- `ACCESS_CREDIT` – access to external funding
- `PROFITABILITY` – profit generation capacity
- `SHORT_TERM_LIQUIDITY` – short-term solvency
- `MEDIUM_TERM_LIQUIDITY` – medium/long-term solvency

Target variable:

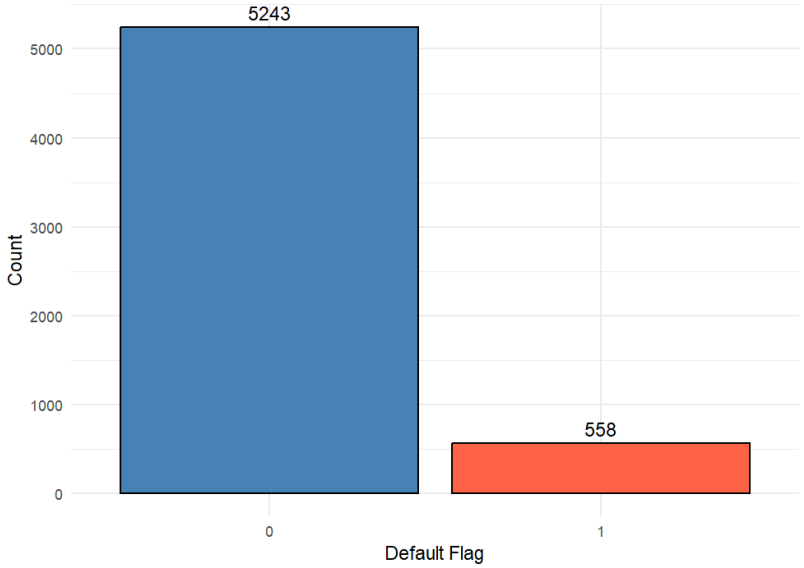
- `DEFAULT_FLAG` – default within 12 months after assessment

Industry category abbreviations

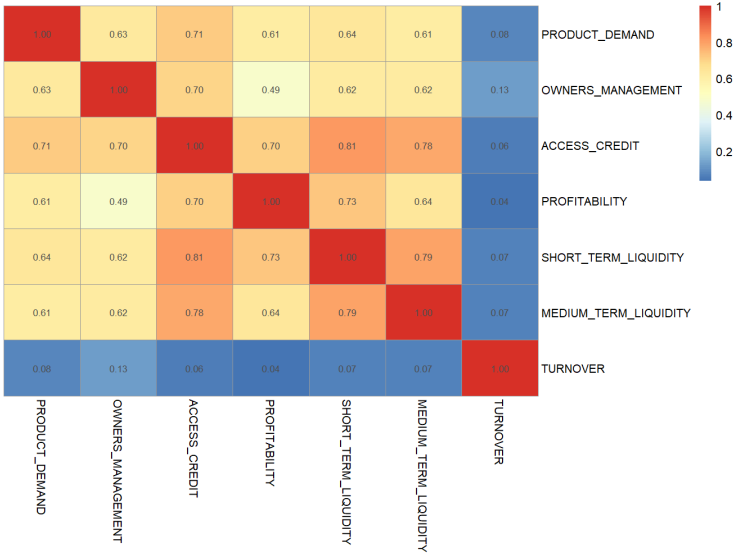
To improve the clarity and readability of plots, long industry names have been replaced with standardized abbreviations.

- **A,L&F** – Agriculture, Livestock and Fisheries
- **EI** – Extractive Industries
- **M** – Manufacturing
- **O** – Other
- **T** – Trade
- **EG&W** – Electricity, Gas and Water
- **H&L** – Hotels and Leisure
- **OM&CI** – Office Machinery and Computer Industries
- **P&CS** – Property and Construction Sectors
- **TS&CI** – Transport, Storage and Communications Infrastructure

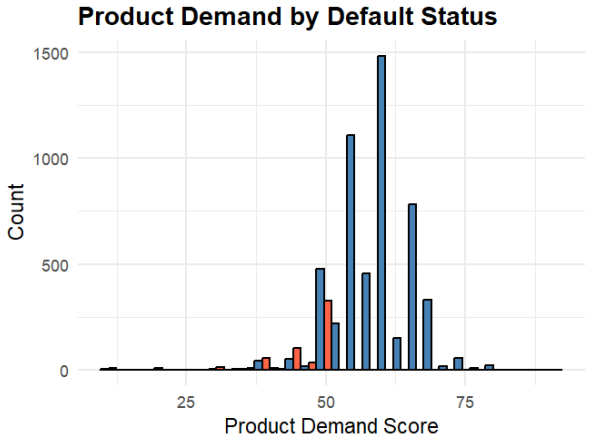
Default status distribution

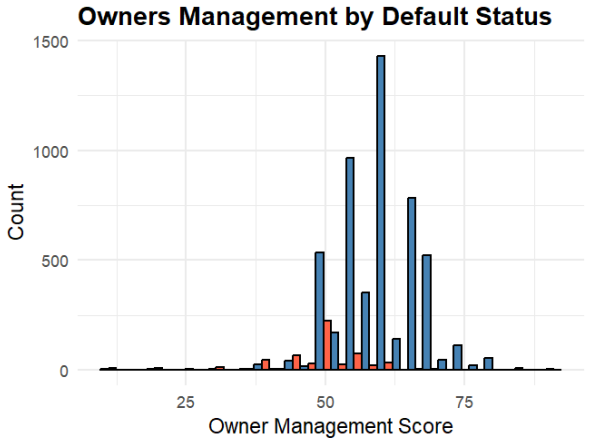


Correlation matrix of numerical variables

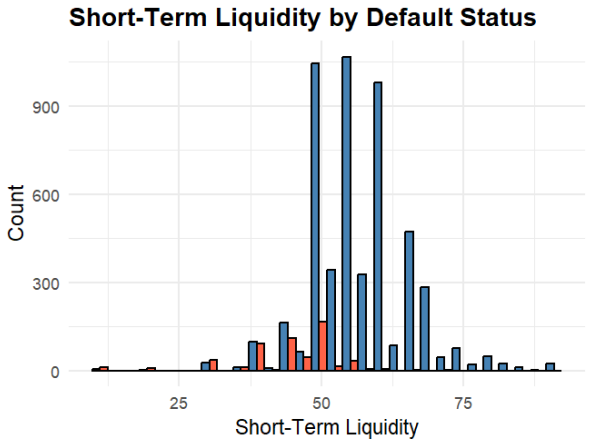


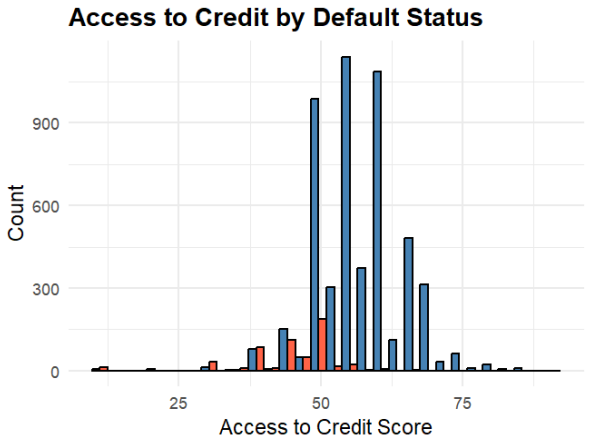
Correlation matrix insights



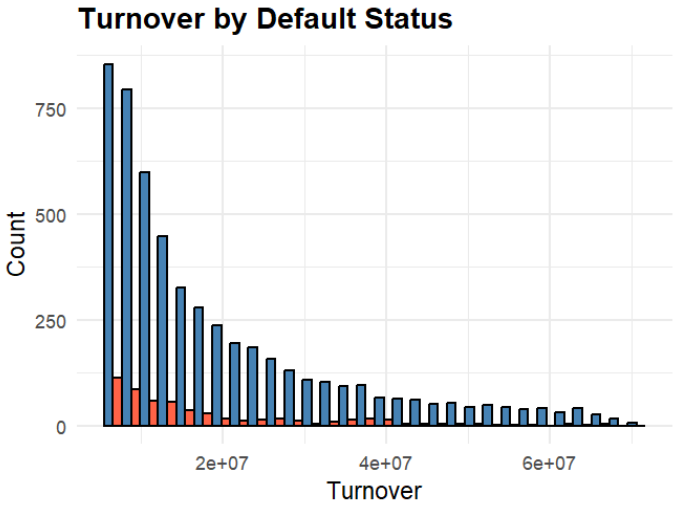


Correlation matrix insights (2)





Correlation matrix insights (3)



Information value

Information Value (IV) measures how well a variable can separate the binary outcome classes, such as default vs. non-default.

Formula:

$$IV = \sum_{i=1}^{n} (\text{DistributionPositive}_i - \text{DistributionNegative}_i) \cdot \ln\left(\frac{\text{DistributionPositive}_i}{\text{DistributionNegative}_i}\right)$$

Interpretation of IV values:

Information Value	Predictive Power
< 0.02	Useless for prediction
0.02 to 0.1	Weak predictor
0.1 to 0.3	Medium predictor
> 0.3	Strong predictor

Information value results

Variable	Identical Rate	Information Value
PRODUCT_DEMAND	0.2105	5.0283
ACCESS_CREDIT	0.1791	2.7138
OWNERS_MANAGEMENT	0.2155	2.4537
SHORT_TERM_LIQUIDITY	0.1778	2.2917
MEDIUM_TERM_LIQUIDITY	0.1968	2.0888
PROFITABILITY	0.2333	1.9309
ASSESSMENT_YEAR	0.3568	1.1853
INDUSTRY	0.3794	0.2178
TURNOVER	0.0013	0.0641
GROUP_FLAG	0.6290	0.0001

Table: Identical rate and information value for selected variables (sorted by IV)

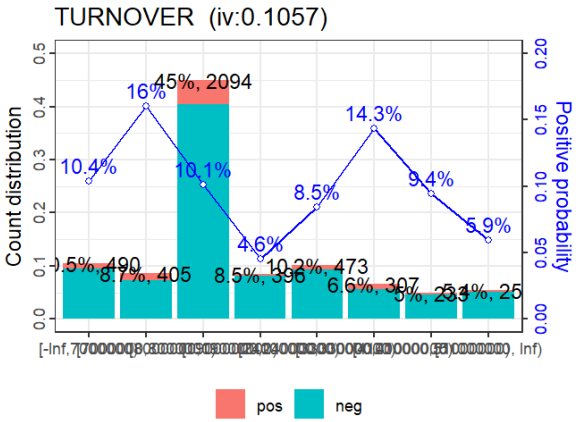
WOE binning

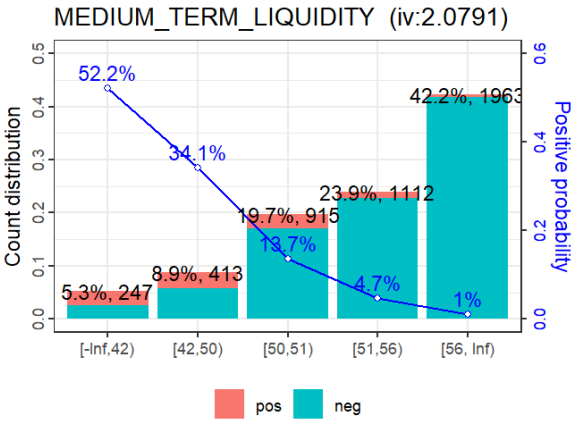
- **WOE binning** transforms variables into bins with calculated Weight of Evidence (WOE) values.
- Helps to distinguish "good" vs. "bad" outcomes based on distribution in each bin.
- The `woebin()` function (from the `scorecard` package) automates binning.
- It maximizes Information Value and enforces WOE monotonicity where possible.

WOE Formula:

$$WOE_i = \ln\left(\frac{\text{DistributionPositive}_i}{\text{DistributionNegative}_i}\right)$$

WOE binning example





Variable binning results

Variable	Binning intervals
ASSESSMENT_YEAR	2000–2006, 2007, 2008
PRODUCT_DEMAND	[-Inf,50), [50,51), [51,Inf)
OWNERS_MANAGEMENT	[-Inf,50), [50,52), [52,62), [62,Inf)
ACCESS_CREDIT	[-Inf,50), [50,52), [52,58), [58,Inf)
PROFITABILITY	[-Inf,50), [50,52), [52,58), [58,Inf)
SHORT_TERM_LIQUIDITY	[-Inf,50), [50,52), [52,60), [60,Inf)
MEDIUM_TERM_LIQUIDITY	[-Inf,42), [42,50), [50,51), [51,56), [56,Inf)
TURNOVER	<7M, 7–8M, 8–19M, 19–24M, 24–33M, 33–41M, 41–51M, >51M
INDUSTRY	(A,L&F, EG&W, EI), (H&L, M), (OM&CI, O), (P&CS), (T, TS&CI)

Figure: Variable binning results after `woebin()`

Summary of data exploration

- Removed `CUSTOMER_ID` as it is not informative for modelling.
- Handled missing data by dropping 3 incomplete records.
- Split the dataset into training (80%) and testing (20%) subsets.
- Applied `var_filter()` to remove low-informative variables based on Information Value:
  - `GROUP_FLAG` was excluded (IV < 0.02).
- Performed WOE binning on predictor variables using `woebin()`:
  - Ensured monotonicity and maximized IV.

AGH

Part II

Modeling & Performance

www.agh.edu.pl

Model Building Plan

We consider the following types of models:

- Logistic regression
- Probit regression
- Linear regression
- Expert model
- Segmentation by `GROUP_FLAG`

After selecting the best variant within each model type, we will compare these models against each other.

Variable selection

Models are built using:

- Full models (with all available variables)
- Variable selection based on Information Value (IV)
- Forward selection
- Backward selection

In our case **Forward and backward selection** aim to minimize the Akaike Information Criterion (AIC), which balances model fit and complexity:

$$\text{AIC} = 2k - 2\ln(\hat{L})$$

- k: number of estimated parameters (variables)
- $\hat{L}$: maximum value of the likelihood function

Forward selection:

- Starts with an empty model
- At each step, adds the variable that most reduces AIC
- Continues until no remaining variable improves the AIC

Backward selection:

- Starts with a full model
- At each step, removes the variable that least increases AIC
- Stops when all remaining variables are significant by AIC

Confusion Matrix

	Predicted	
	0 (Non-default)	**1 (Default)**
Actual 0 (Non-default)	True Negative (TN)	False Positive (FP)
Actual 1 (Default)	False Negative (FN)	True Positive (TP)

- **False Positive (Type I error)**: predicted default but actually repaid
- **False Negative (Type II error)**: predicted repayment but actually defaulted

Evaluation Metrics

The outcome variable is **imbalanced**, so we use specific metrics to evaluate model performance.

- **F1 Score** – compromise between precision and recall:

$$F1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2\,\text{TP}}{2\,\text{TP} + \text{FP} + \text{FN}}$$

  - **Precision**: $\frac{TP}{TP+FP}$
  - **True Positive Rate (Recall)**: $\frac{TP}{TP+FN}$

Evaluation Metrics

- **AUC**: area under the ROC curve (**TPR vs FPR**)
  - **True Positive Rate (Recall)**: $\frac{TP}{TP+FN}$
  - **False Positive Rate**: $\frac{FP}{FP+TN}$
- **Kolmogorov–Smirnov statistic** measures the maximum separation between the distributions of defaulted and non-defaulted clients

$$KS = \max_x |F_{\text{default}}(x) - F_{\text{non-default}}(x)|$$

Optimal Cut-off

- Cut-off determined using **Youden's J statistic**:

$$J = \text{Sensitivity} + \text{Specificity} - 1 = \text{TPR} + \text{TNR} - 1$$

Selects the threshold that best balances sensitivity and specificity

Best Model Selection

- All models are evaluated using the same metrics and cutoff method for consistency.
- For each model type (logit, probit, linear, expert):
  - Multiple variants are built and tested using 10-fold cross-validation.
  - The best-performing variant is selected.
- Final comparison is done on a separate test set to choose the overall best model.

AGH

Part III

Logistic regression

www.agh.edu.pl

Models

	Without WOE Binning	WOE Binning
Model	**Description**	**Description**
Model 1	Full model	Full model
Model 2	Full model without – *TURNOVER*	Full model without – *TURNOVER*
Model 3 (forward)	Full model without – *MEDIUM TERM LIQUIDITY* – *INDUSTRY*	Full model without – *MEDIUM TERM LIQUIDITY*

Table 4: Descriptions of logistic models used in the analysis

Variable	Identical Rate	Information Value
PRODUCT_DEMAND	0.2105	5.0283
ACCESS_CREDIT	0.1791	2.7138
OWNERS_MANAGEMENT	0.2155	2.4537
SHORT_TERM_LIQUIDITY	0.1778	2.2917
MEDIUM_TERM_LIQUIDITY	0.1968	2.0888
PROFITABILITY	0.2333	1.9309
ASSESSMENT_YEAR	0.3568	1.1853
INDUSTRY	0.3794	0.2178
TURNOVER	0.0013	0.0641
GROUP_FLAG	0.6290	0.0001

Table: Identical rate and information value for selected variables (sorted by IV)

Metrics

	Without WOE Binning			WOE Binning		
Metric	**Model 1**	**Model 2**	**Model 3**	**Model 1**	**Model 2**	**Model 3**
AUC	0.9644	0.9647	0.9656	0.9706	0.9699	0.9707
KS	0.8622	0.8628	0.8627	0.8779	0.8762	0.8784
F1	0.9437	0.9440	0.9412	0.9455	0.9519	0.9468

Table 5: Comparison of model performance metrics with and without WOE Binning

We selected Model 2 with and without WOE based primarily on the **F1 score**.

Part IV

Probit specification

Description of probit models

	Without WOE Binning	WOE Binning
Model	**Description**	**Description**
Model 1	Full model	Full model
Model 2	Full model without – *TURNOVER*	Full model without – *TURNOVER*
Model 3 (forward)	Full model without – *PROFITABILITY* – *MEDIUM TERM LIQUIDITY* – *INDUSTRY – TURNOVER*	Full model without – *MEDIUM TERM LIQUIDITY*
Model 4 (backward)	Full model without – *MEDIUM TERM LIQUIDITY* – *TURNOVER*	

Table 6: Descriptions of probit models used in the analysis

Metrics

	Without WOE Binning				WOE Binning		
Metric	**Model 1**	**Model 2**	**Model 3**	**Model 4**	**Model 1**	**Model 2**	**Model 3**
AUC	0.9394	0.9370	0.9395	0.9205	0.9707	0.9699	0.9707
KS	0.8214	0.8131	0.8213	0.7813	0.8784	0.8768	0.8796
F1	0.9474	0.9491	0.9477	0.9486	0.9458	0.9502	0.9464

Table 7: Comparison of model performance metrics with and without WOE Binning

Performance of Logit vs Probit models - confusion matrices

Model	Prediction	Actual 0	Actual 1
Logit (Original Variables)	0	965 (TN)	5 (FN)
	1	95 (FP)	86 (TP)
Logit (WOE)	0	982 (TN)	5 (FN)
	1	78 (FP)	86 (TP)

Model	Prediction	Actual 0	Actual 1
Probit (Original Variables)	0	973 (TN)	7 (FN)
	1	87 (FP)	84 (TP)
Probit (WOE)	0	992 (TN)	6 (FN)
	1	68 (FP)	85 (TP)

Comparison of classification performance

The table below compares the classification performance of the four evaluated models: logit and probit, each estimated on both raw and WOE-transformed variables. The comparison focuses on AUC, KS, and especially F1-score, which is the main evaluation criterion in this analysis.

Model	AUC	KS	F1 Score
Logit (Original Variables)	0.9622	0.8554	0.9507
Logit (WOE)	0.9648	0.8715	0.9595
Probit (Original Variables)	0.9585	0.8410	0.9539
Probit (WOE)	0.9644	0.8699	0.9640

Table: Performance metrics for logit and probit models (test set)

AGH

Part V

Linear regression

- We have to convert our target variable `DEFAULT_FLAG` to numeric type.
- The linear regression model can be written as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \epsilon,$$

where ϵ is the error term and $\epsilon \sim \mathcal{N}(0, \sigma^2)$.

- The model is estimated by minimizing the sum of squared residuals

$$\min_{\beta} \sum_{i=1}^{N} (Y_i - \hat{Y}_i)^2,$$

where $\hat{Y}_i$ denotes the predicted value for the i-th observation.

Description of linear models

Model	Description
Model 1	Full model
Model 2	Full model without – *TURNOVER*
Model 3 (forward)	Full model without – *PROFITABILITY* – *MEDIUM TERM LIQUIDITY* – *TURNOVER*

Table 8: Descriptions of linear models used in the analysis

Evaluation of the models

Metric	Model 1	Model 2	Model 3
AUC	0.9602	0.9603	0.9603
KS	0.8367	0.8356	0.8387
F1	0.9309	0.9313	0.9378

Table: Comparison of model performance metrics

Prediction	Actual 0	Actual 1
0	965 (TN)	6 (FN)
1	95 (FP)	85 (TP)

Table: Confusion matrix for linear regression model - Model 3 (Test Set)

Comparison

The table below summarizes the test set metrics for the linear model versus the top WOE-based probit classifier:

Model	AUC	KS	F1 Score
Linear Regression	0.9615	0.8444	0.9503
Probit (WOE-based)	0.9644	0.8699	0.9640

Table: Performance comparison: linear regression vs. WOE-based probit

The problem with linear regression

- logistic or probit regression - naturally output values interpretable as probabilities in the $[0, 1]$ range
- linear regression - can yield predictions outside this interval
  For example, some predicted values on the test set include:

  `-0.024, 0.269, -0.091, -0.039, -0.080, 0.050`
- the lack of probabilistic meaning weakens its practical applicability in credit risk modeling

Linear regression is not recommended as a primary model for default prediction in this context.

AGH

Part VI

Experts model

www.agh.edu.pl

Description of expert model

The expert model can be written as:

$$p := \mathbb{P}(Y = 1 \mid X) = \frac{1}{1 + e^{-0.1 \times Score}},$$

where *Score* is the weighted average of the following variables:

Variable	WEIGHT
PRODUCT DEMAND	20%
OWNERS MANAGEMENT	10%
ACCESS CREDIT	10%
PROFITABILITY	15%
SHORT TERM LIQUIDITY	25%
MEDIUM TERM LIQUIDITY	20%

Prediction	Actual 0	Actual 1
0	906 (TN)	10 (FN)
1	154 (FP)	81 (TP)

Table: Confusion matrix for expert model (test set)

Model	AUC	KS	F1 Score
Expert Model	0.9385	0.7448	0.9170

Table: Performance metrics for expert model (test set)

Part VII

Financial Holding Type

Metrics and test set evaluation

Independent Entity

Prediction	Actual 0	Actual 1
0	345 (TN)	1 (FN)
1	22 (FP)	28 (TP)

Subsidiary

Prediction	Actual 0	Actual 1
0	607 (TN)	4 (FN)
1	86 (FP)	58 (TP)

Model	AUC	KS	F1 Score
Independent Entity	0.9683	0.9056	0.9677
Subsidiary	0.9404	0.8114	0.9310

AGH

Part VIII

Summary

Metrics

Model	AUC	KS	Accuracy	F1
Logit (Original Variables)	0.9622	0.8554	0.9131	0.9507
Logit (WOE)	0.9648	0.8715	0.9279	0.9595
Probit (Original Variables)	0.9585	0.8419	0.9192	0.9544
Probit (WOE)	0.9644	0.8699	0.9357	0.9640
Linear	0.9615	0.8444	0.9123	0.9503
Experts model	0.9385	0.7448	0.8575	0.9170
Independent Entity	0.9684	0.9056	0.9419	0.9677
Subsidiary	0.9404	0.8114	0.8808	0.9310

Table: Comparison of models performance metrics

Important observations

- performance differences between the logit and probit models are minimal
- linear regression is not good for binary classification problem
- model with limitations (e.g expert model) has worse performance
- WOE binning transformation is noteworthy improvement which helps to better detect dependencies in data
- entity type segmentation is important for credit risk modelling.

Thank you for your attention!