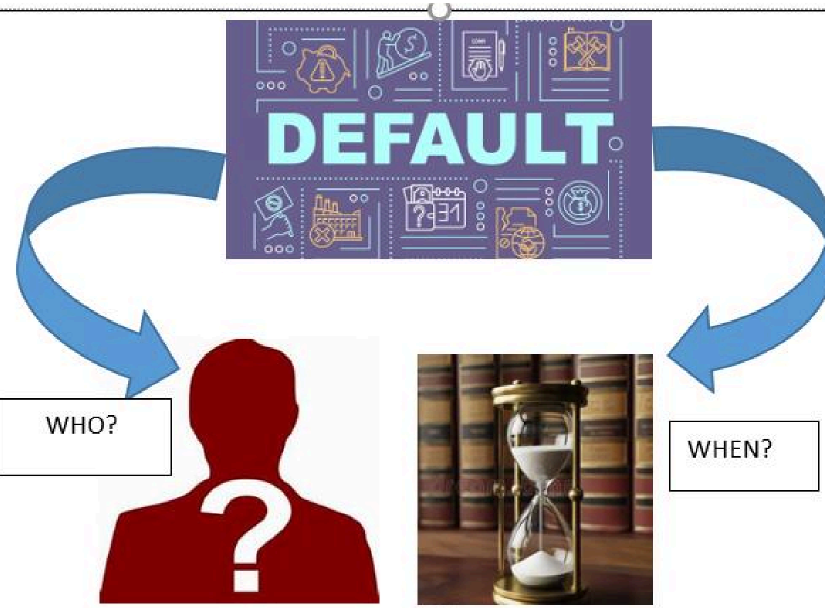
2025 FALL SEMESTER FINAL PROJECT

DEPARTMENT : DATA SCIENCE
COURSE : DATA BUSINESS

SURVIVAL MODELLING - INTRODUCTION



- Consumer credit risk is often framed a classification problem : who will will default?
- However, in practice, when is equally as important.
- Early defaults are costlier than late defaults: they reduce interest income, trigger higher provisioning under IFRS 9 “lifetime expected credit loss”, and may point to origination or fraud issues

SURVIVAL MODELLING - RESEARCH OBJECTIVES

Study Objectives:

1. To model time-to-default (rather than default/no-default only),
2. To compare a classical semi-parametric approach (Cox PH) with a flexible machine-learning approach (Random Survival Forests), and
3. To evaluate predictive utility at business-relevant horizons (e.g., 12/24/36 months) using metrics designed for censored outcomes (e.g., concordance and probability score-based criteria).

This study contributes a practical, end-to-end survival modeling workflow suitable for real credit portfolios:
(i) **formalizing event/censoring definitions from loan performance data**,
(ii) **benchmarking interpretable and nonlinear survival models**, and
(iii) **connecting model outputs (risk rankings and horizon-specific survival probabilities) to lending actions such as monitoring, pricing, and early intervention**.

SURVIVAL MODELLING - DATA + STUDY DESIGN

Dataset: Lending Club loans (Kaggle)

Train/Test split: time-based split by issue date

The **time-to-event outcome** is defined as the number of months from origination to default. Loans that do not default during the observed window are considered as 'censored'

Models compared:

- Cox Proportional Hazards
- Random Survival Forest (nonlinear benchmark)

Metrics reported:

- **C-index** (ranking; higher is better)
- **Time-dependent AUC** at **12/24 months** (discrimination at a horizon)
- **IBS** (Integrated Brier Score)
- **KM-based observed risk + Top-10% lift**

https://www.kaggle.com/datasets/adarshsng/lending-club-loan-data-csv?select=loan.csv

LENDING CLUB - PORTFOLIO OVERVIEW

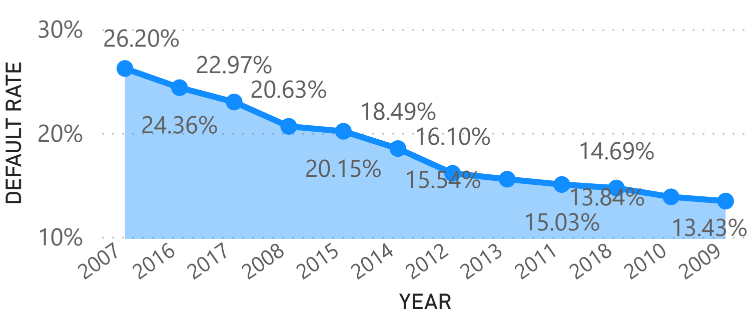
June 2007
EARLIEST ISSUE DATE

December 2018
LATEST ISSUE DATE

LOANS COUNT	TOTAL LOAN DISBURSED VOLUME
1.3M	19bn

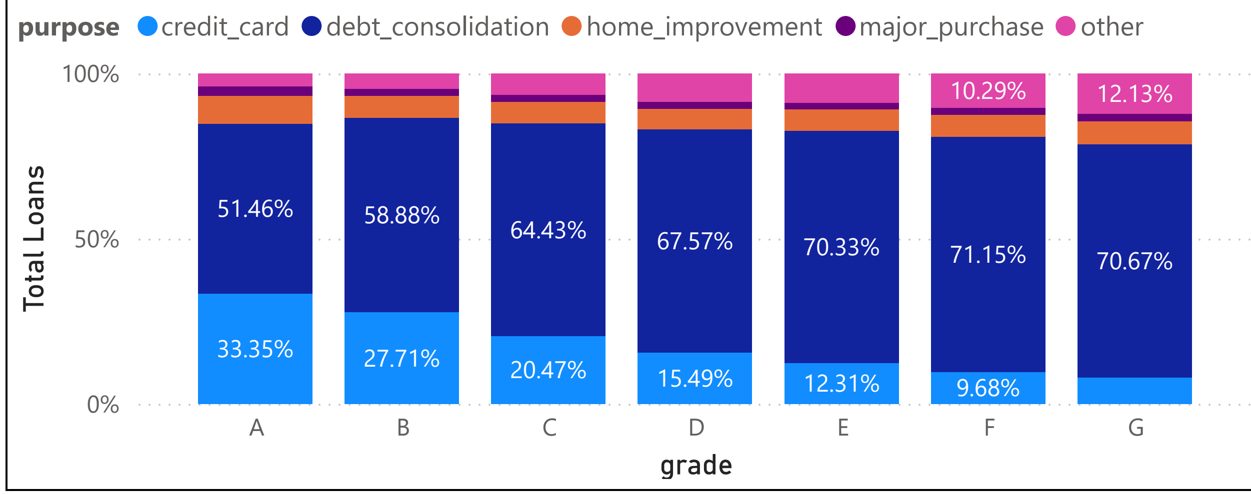
OBSERVED DEFAULT RATE (%)	MEDIAN TIME TO DEFAULT (MONTHS)
20.09	14.03

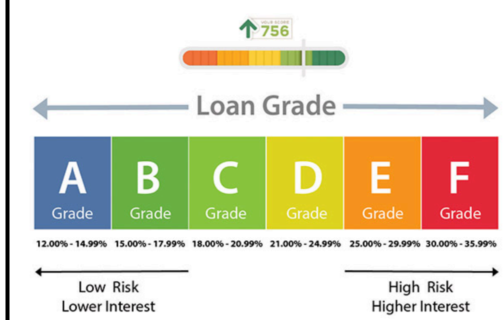
DEFAULT RATE BY ORIGINATION YEAR



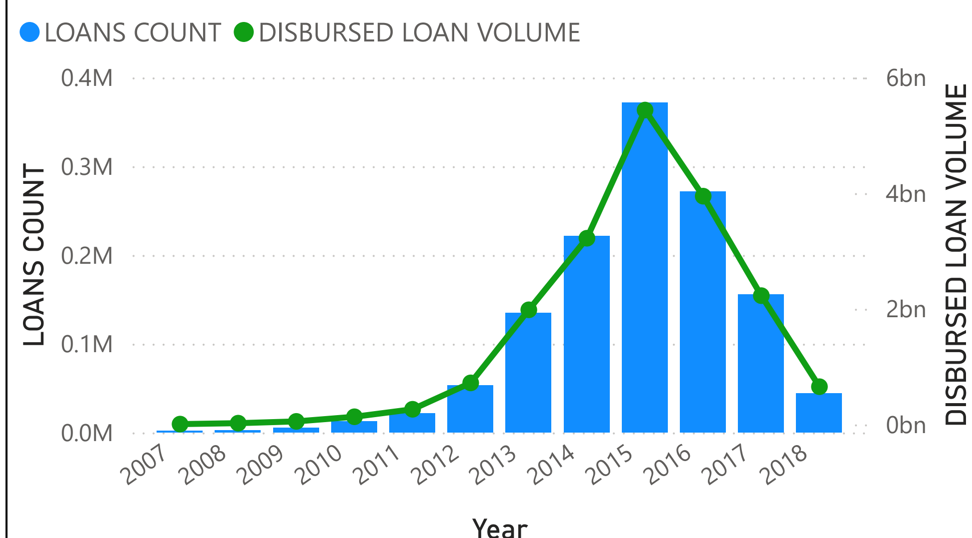
PORTFOLIO MIX : LOAN GRADE BY LOAN PURPOSE

A higher credit grade (grade A) typically means a lower interest rate

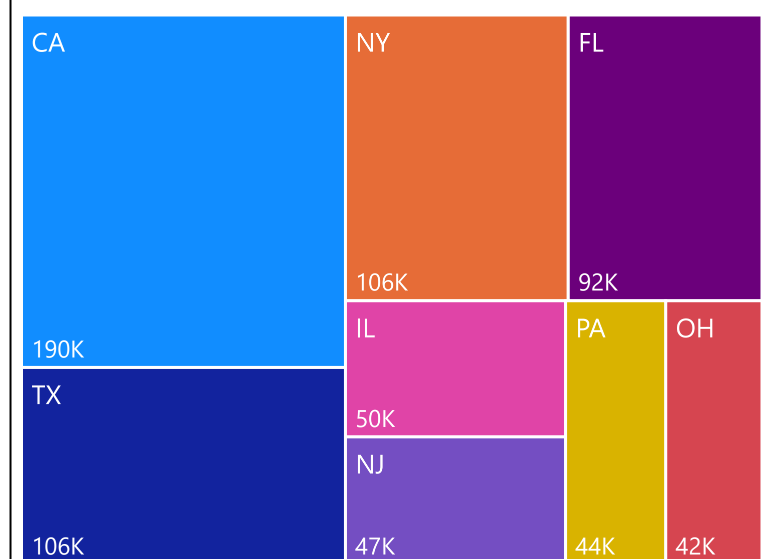




LOAN ORIGINATION OVER TIME



LOAN ORIGINATION BY STATE



LENDING CLUB DATA - RISK SEGMENTATION

DEFAULT SHARE (FUNDED AMT)

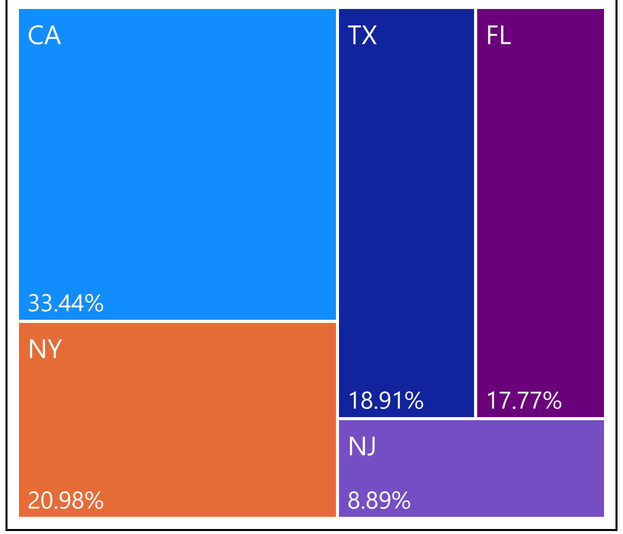
21.67%

AVERAGE FUNDED AMT

14.39K

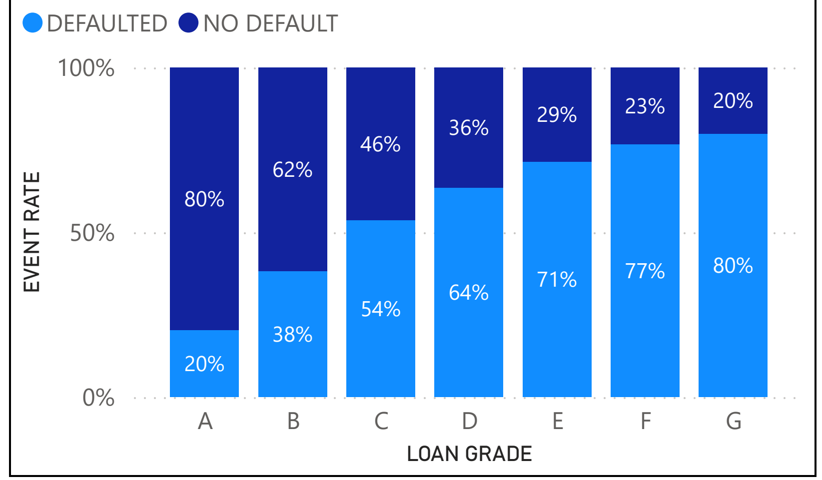
Months	Early Default % (Funded)	Eligible Exposure ($)	Early Default Exposure ($)
12	11.50%	$14,651,574,850	$1,685,564,900
24	30.04%	$10,551,163,575	$3,169,772,250
36	53.03%	$7,282,173,350	$3,861,548,825
60	98.27%	$4,110,087,525	$4,038,954,725

RISK EXPOSURE BY STATE

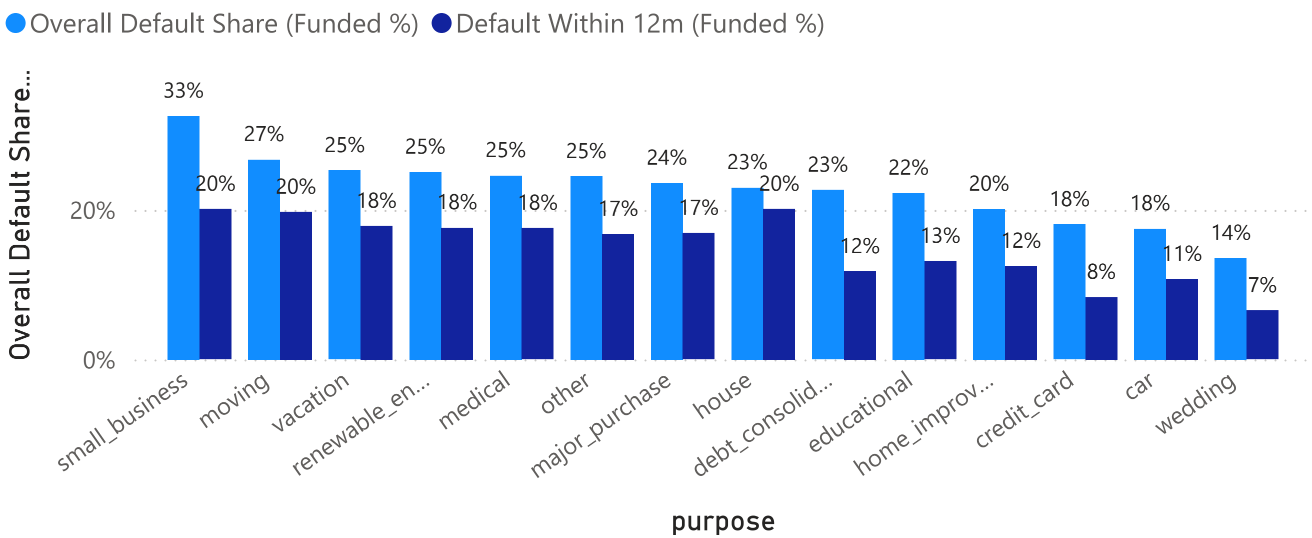


DEFAULT RATE BY LOAN GRADE

The higher the loan grade the lower the default rate



Default Risk by Loan Purpose: Overall vs First 12 Months



- **Overall default exposure is high (21.67% of funded amount)**, suggesting defaults are concentrated in relatively larger loans, not just many small loans.
- **Early-default risk spikes over time**: ~**11.5% within 12 months** and rises steeply by **36–60 months**,
- **Risk is concentrated by segment**: lower grades (E–G) show much higher default share than A–C, and **Small Business** purpose stands out as the highest-risk purpose (overall and early).

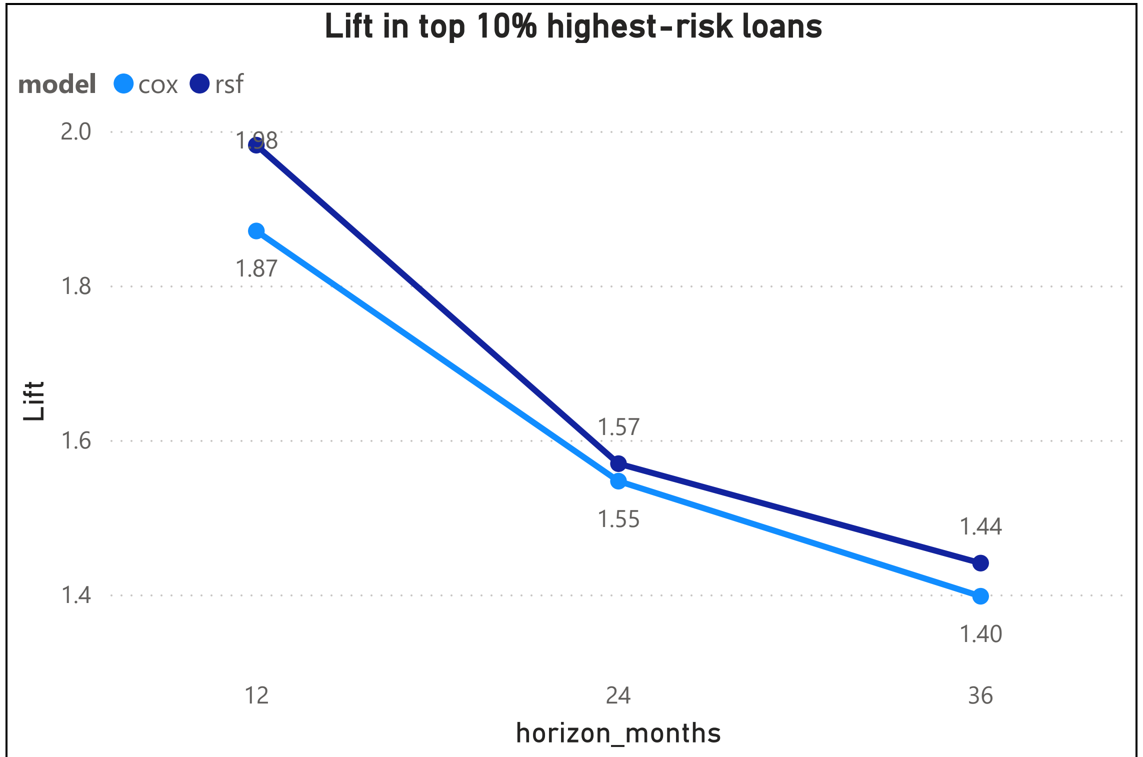
SURVIVAL ANALYSIS MODELLING

Performance Measures Evaluation

C-INDEX (COX MODEL)	C-INDEX (RSF MODEL)	IBS SCORE (COX MODEL)	IBS SCORE (RSF MODEL)
0.67	0.66	0.121	0.119

Time Dependent AUC

horizon_months	cox	rsf
12	0.68	0.67
24	0.69	0.68



- The models can reasonably rank who defaults earlier.
- Time-dependent AUC is **~0.68–0.69** at **12–24 months**,
- IBS is **0.121 (Cox)** vs **0.119 (RSF)** *(lower is better)* — very similar, so Cox provides near-equal performance
- Top-10% lift is **~1.9× at 12m**, dropping to **~1.55× at 24m** and **~1.4× at 36m**. (Cox Model)

*Targeting the **top 10% predicted risk** captures a disproportionately high share of near-term defaults*

*****longer horizons are influenced by post-origination shocks (job loss, macro conditions), so early-horizon signals from origination data fade.*****

SURVIVAL ANALYSIS MODELLING

Horizon-Specific Risk Segmentation and Targeting Lift (Kaplan–Meier–Adjusted)

Risk by horizon (KM-observed vs predicted + top-10% targeting lift) - 12 Months

horizon_months	12			
model	Observed default prob (KM)	Avg predicted default prob	Observed default prob in top 10% (KM)	Top 10% lift (vs overall)
cox	0.23	0.07	0.43	1.87
rsf	0.23	0.07	0.45	1.98

Risk by horizon (KM-observed vs predicted + top-10% targeting lift) - 24 Months

horizon_months	24			
model	Sum of Observed default prob (KM)	Avg predicted default prob	Observed default prob in top 10% (KM)	Top 10% lift (vs overall)
cox	0.50	0.19	0.78	1.55
rsf	0.50	0.19	0.79	1.57

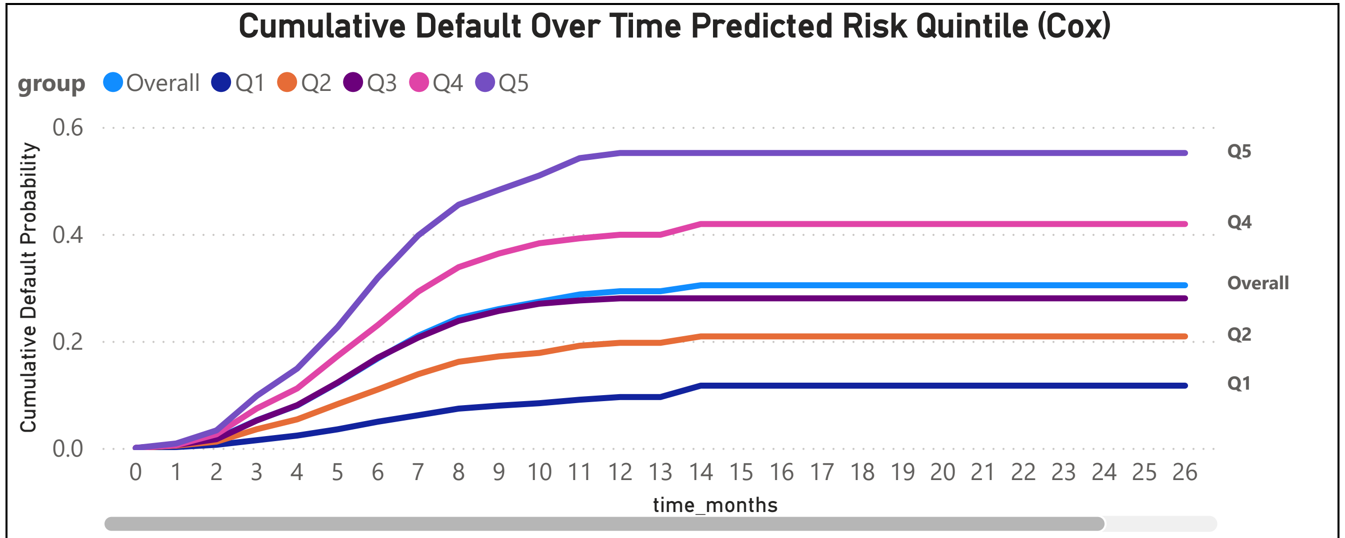
Risk by horizon (KM-observed vs predicted + top-10% targeting lift) - 36 Months

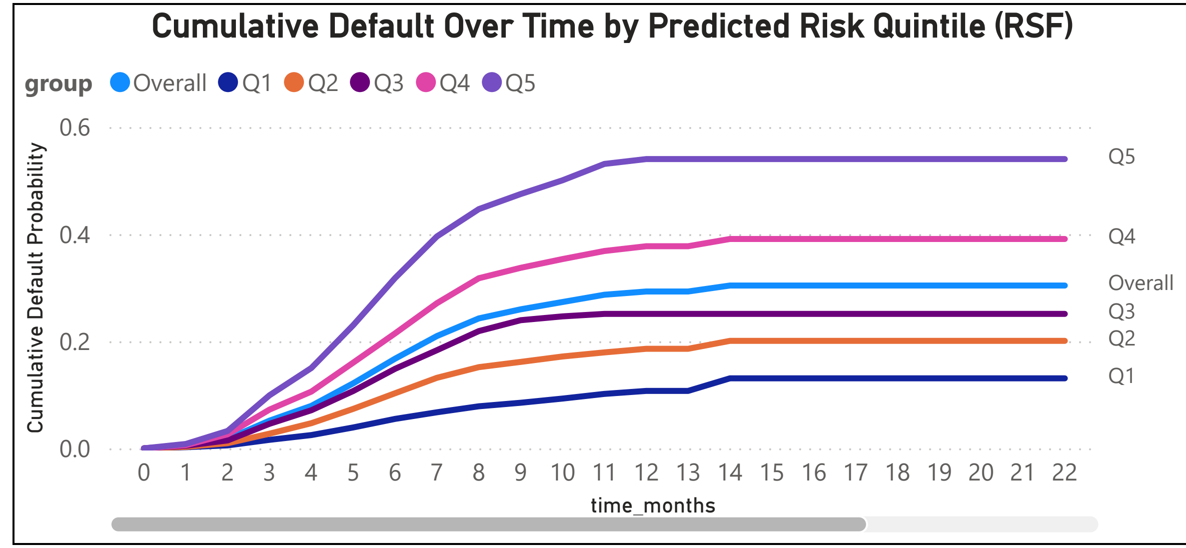
horizon_months	36			
model	Sum of Observed default prob (KM)	Avg predicted default prob	Observed default prob in top 10% (KM)	Top 10% lift (vs overall)
cox	0.63	0.28	0.88	1.40
rsf	0.63	0.28	0.91	1.44

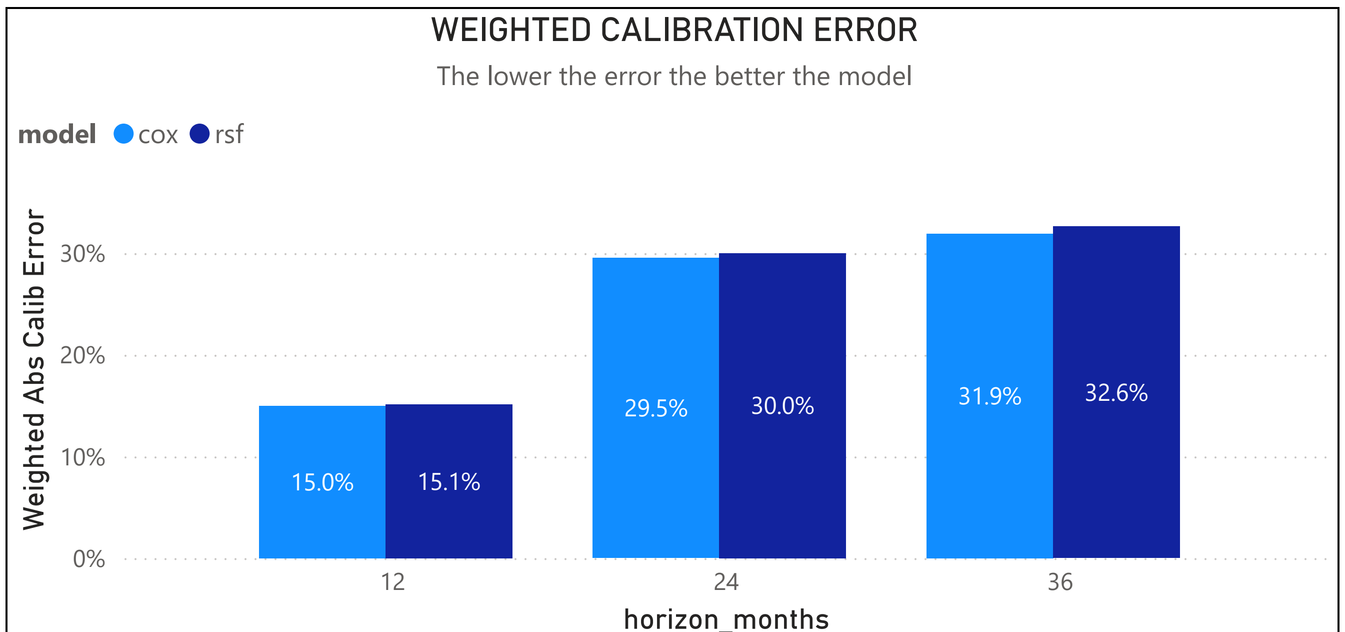
- **Both models deliver similar top decile targeting lift.**
- The **top 10% predicted-risk group** shows substantially higher KM-observed default probability than the full cohort at **12/24/36 months**,
- Lift is strongest at **12 months and declines by 36 months** as cumulative defaults become more common
- Mean predicted PD is lower than KM-observed PD at each horizon, indicating **underestimation of absolute risk** even when ranking is reasonable
- ***RSF is slightly stronger for targeting.***

SURVIVAL ANALYSIS MODELLING

KM Risk Separation and Horizon Calibration

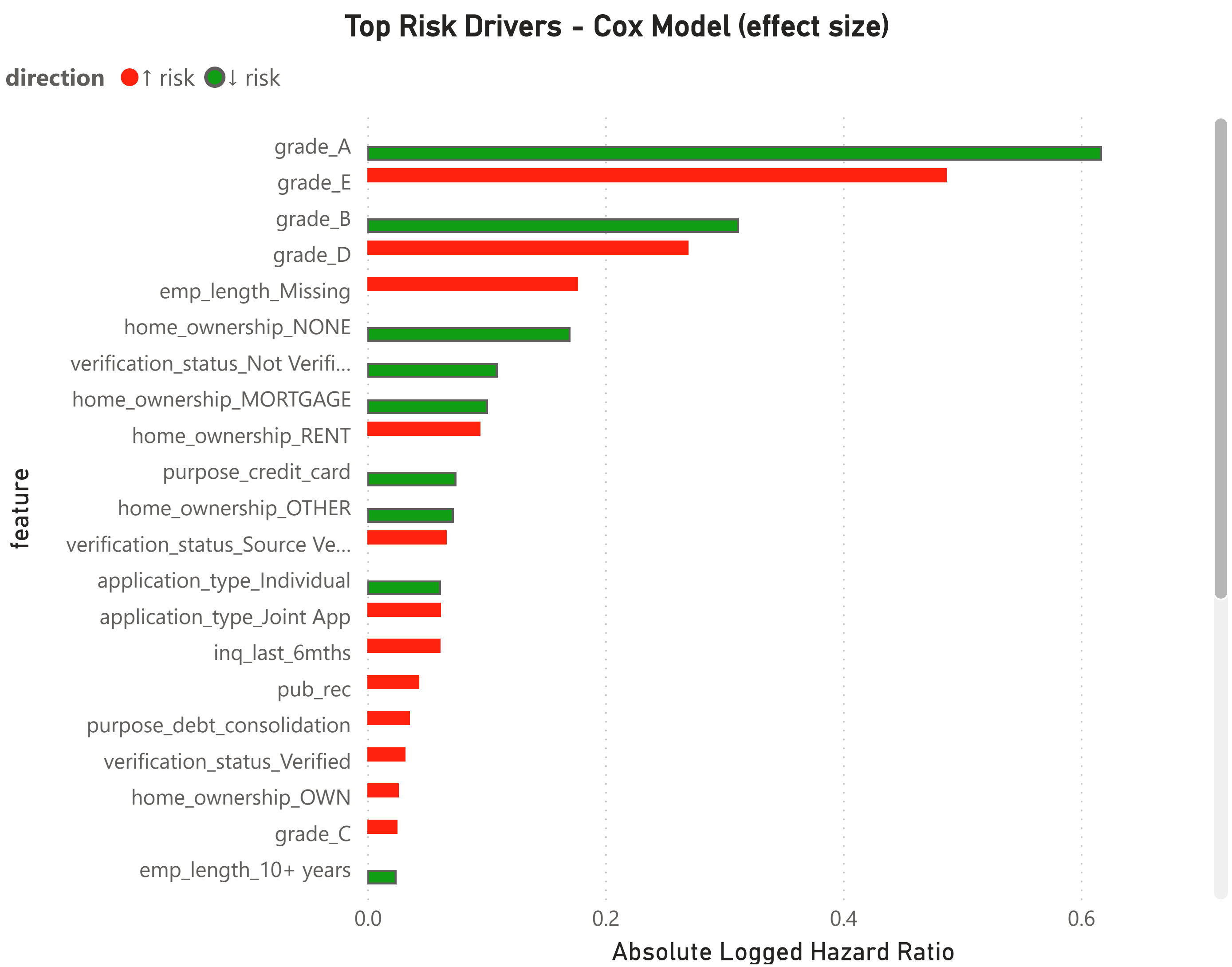






- Q5 consistently shows the **highest cumulative default** and Q1 the lowest.
- Curves diverge quickly in the first months, suggesting the **models capture early-default dynamics**.
- The KM quintile curves are similarly shaped and separated for both models, implying **comparable ranking performance**.
- Weighted calibration error increases from ~15% (12m) to ~30%+ (24–36m), meaning **absolute PDs are less reliable long-term even** if ranking remains stable

SURVIVAL ANALYSIS MODELLING - TOP RISK DRIVERS



- **Credit Grade** variables dominate the strongest effects
- Additional drivers include **borrower stability / verification signals** (e.g., employment length, verification status) and **loan purpose / housing indicators**.

***Red = higher hazard (earlier default), Green = lower hazard (slower default).

CONCLUSION

- **Discrimination (ranking):** Cox **C-index = 0.67**, RSF **0.66** : both models *separate higher-risk borrowers reasonably well*.
- **Horizon discrimination:** time-dependent AUC is about **0.68–0.69 (12–24m)** → strongest predictive separation is in the **early repayment window**.
- **Overall accuracy/calibration:** IBS is **0.121 (Cox)** vs **0.119 (RSF)** (lower is better) → essentially similar; RSF slightly lower.
- **Targeting value:** the **top 10% predicted-risk** group has much higher observed default:
- The **lift declines with horizon** as defaults become more widespread over time.
- **Calibration gap:** mean predicted PD is below KM-observed PD at each horizon → the models **rank well** but likely **underestimate absolute risk**, suggesting a need for **post-hoc calibration**.

LIMITATIONS & FUTURE WORK

ACADEMIC CONTRIBUTIONS :

- Horizon-specific risk segmentation
- Benchmarking linear vs nonlinear survival models
- Explainability of credit risk factors
- Integration of survival analysis and credit risk

LIMITATIONS:

- Static origination-time features only.
- Single platform and market
- Simplified outcome definition
- Limited exploration of nonlinear ML models

FUTURE WORK:

Future work could extend the pipeline to multi-state or competing-risks frameworks, incorporate macroeconomic covariates, explore advanced machine-learning survival models, and quantify economic value in terms of IFRS 9 expected-loss reductions.