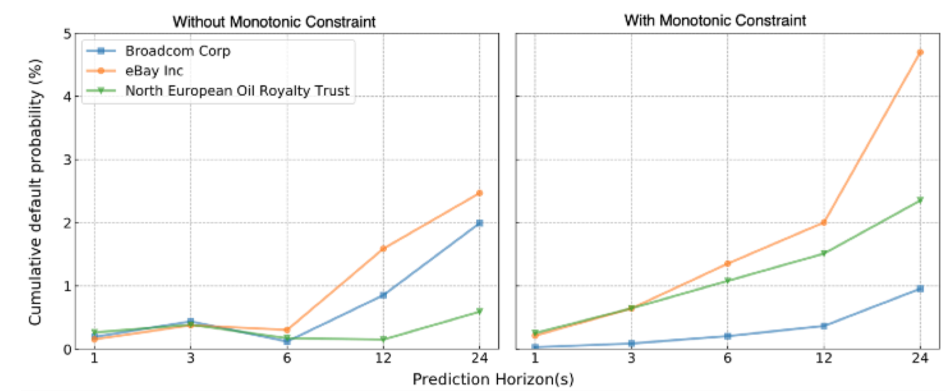
Multiperiod Corporate Default Prediction — A Domain Knowledge-tailored Neural Network Approach

Dr. Chuan-Ju Wang
CITI, Academia Sinica, Taiwan

Joint work with Wei-Lun Luo, Prof. Ming-Feng Tsai, and Prof. Jin-Chuan Duan

August 29, 2023
Cardiff University, UK

Default Analysis



Deep Learning Methods

Risk classification
Risk rankings

Single-period modeling
(Unreasonable term structures)

Linear composites of covariates
(Lack of flexible functional form)

Gap

Flexible functional form
(Overfitting)

Statistical Methods

Term structures of default probabilities
Number of default occurrences

Forward Intensity Model (FIM) [1]

[1] Duan, J. C., Sun, J., & Wang, T. (2012). Multiperiod corporate default prediction—A forward intensity approach. *Journal of Econometrics*, 170(1), 191-209.

Main Contributions

1. **Proposition of a domain-knowledge-tailored neural network**: The paper introduces a novel deep neural network (DNN) model that incorporates economic domain knowledge, specifically designed for multi-period default prediction.
    - Flexible functional forms with DNNs : Enhance the performance
    - Follow FIM structure to model default intensities: Provide consistent term structures of default probabilities
    - Use economic domain knowledge to regulate the networks: Mitigate overfitting
2. **Validation through extensive experiments**: The paper verifies the efficacy of the proposed model through tests conducted on a large US corporate default dataset spanning from 1994 to 2021.
3. **Applicability and insights for machine learning research in finance**: The proposed method can be applied to most neural networks, and it provides valuable insights for ongoing machine learning research, especially in financial applications.

Framework: A Forward-Intensity Approach

- Forward intensities of the two independent doubly stochastic Poisson processes for the time interval between m to $m + \Delta t$
  - Default: $f_m(X_{i,t})$
  - Other exit: $q_m(X_{i,t})$
  - $X_{i,t}$ denotes the set of covariates of the i-th company at prediction time t
- Forward probability for one period, length=Δt, $m = 0,1,2,3,\ldots,\ldots$
  - Survival: $p_s(X_{i,t}; m) = e^{-(f_m(X_{i,t})+q_m(X_{i,t}))\Delta t}$
  - Default: $p_d(X_{i,t}; m) = 1 - e^{-f_m(X_{i,t})\Delta t}$
  - Other exit: $p_o(X_{i,t}; m) = 1 - p_s(X_{i,t}; m) - p_d(X_{i,t}; m) = e^{-f_m(X_{i,t})}\left(1 - e^{-q_m(X_{i,t})\Delta t}\right)$
- Cumulative default probability (for applications not estimation)

$$\text{Prob}[X_{i,t}, n; \Delta t] = \sum_{m=0}^{n-1}\left[p_d(X_{i,t}; m)\prod_{j=0}^{m-1} p_s(X_{i,t}; m)\right]$$

Forward-Intensity Model (FIM)

- Duan et al. (2012) applied a linear composite to obtain the forward intensities.

$$
\begin{aligned}
f_m^{\mathrm{FIM}}(X_{i,t}) &= \exp\left(\beta_0(m) + \beta_1(m)x_{i,t,1} + \ldots + \beta_k(m)x_{i,t,k}\right) \\
&= \exp\left(\beta(m) \cdot X_{i,t}\right) \\
q_m^{\mathrm{FIM}}(X_{i,t}) &= \exp\left(\bar{\beta}_0(m) + \bar{\beta}_1(m)x_{i,t,1} + \ldots + \bar{\beta}_k(m)x_{i,t,k}\right) \\
&= \exp\left(\bar{\beta}(m) \cdot X_{i,t}\right)
\end{aligned}
$$

  - $\beta(m), \bar{\beta}(m)$: Coefficient vectors of the forward period m

View FIM as a Special Case of Deep Neural Networks

Default

$$f_m^{\mathrm{FIM}}(X_{i,t}) = \exp\left(\beta(m) \cdot X_{i,t}\right)$$

Other exit

$$q_m^{\mathrm{FIM}}(X_{i,t}) = \exp\left(\bar{\beta}(m) \cdot X_{i,t}\right)$$

$$\left(f_m^{\mathrm{MLP}}(X_{i,t}), q_m^{\mathrm{MLP}}(X_{i,t})\right)_{m=0,1,\ldots,n-1} \rightarrow \Theta^{\mathrm{MLP}}(X_{i,t}; m = 0,1,\cdots, n-1)$$

- MLP stands for a simple architecture of neural networks: multi-layer perceptron (MLP).
- The MLP model generates the two types of forward intensities for all prediction horizons at once [2].
- Θ^{MLP} is the parameters of the MLP, and n is a parameter deciding how many prediction horizons for each forward intensity that the MLP can generate.

[2] Divernois, M. A. (2020). A Deep Learning Approach to Estimate Forward Default Intensities. *Swiss Finance Institute Research Paper*, (20-79).

Capture Time Dynamics of Covariates

$$\left(f_m^{\mathrm{MLP}}(X_{i,t}), q_m^{\mathrm{MLP}}(X_{i,t})\right)_{m=0,1,\ldots,n-1} \rightarrow \Theta^{\mathrm{MLP}}(X_{i,t}; m = 0,1,\cdots,n-1)$$

$$\Downarrow$$

$$\left(f_m^{\mathrm{RNN}}(X_{i,t}), q_m^{\mathrm{RNN}}(X_{i,t})\right)_{m=0,1,\ldots,n-1} \rightarrow \Theta^{\mathrm{RNN}}(X_{i,t-h}, \ldots, X_{i,t}; m = 0,1,\ldots,n-1)$$

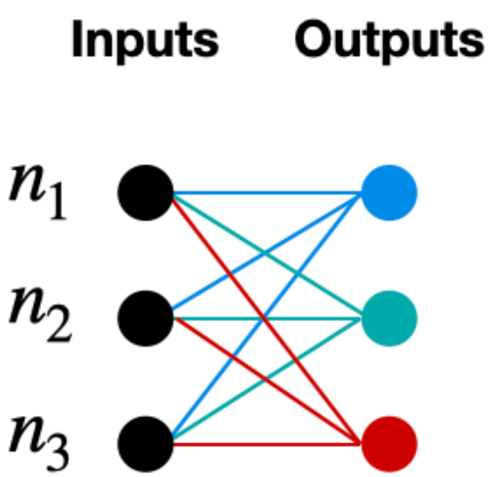
- Recurrent Neural Network, often abbreviated as RNN, is a type of artificial neural network designed to recognize patterns in sequences of data.
  - Long short-term memory (LSTM)
  - Gated recurrent unit (GRU) [fewer parameters than LSTM]
- For MLP, it only takes the covariates of a given company at the current time t.
- However, for RNN, it takes the covariates of each given company in the past h months of the current time t.

Our Domain Knowledge Tailored (DKT) Approach

- **Complex machine learning models**: Machine learning models with complex functional forms often achieve superior performance.
- **Risk of overfitting**: Despite their improved performance, these complex models are prone to overfitting.
- **Incorporation of domain knowledge**: We incorporate economic domain knowledge to simplify the model, effectively reducing the overfitting issue.
- **Tailoring fully connected layers**: The paper leverages economic insights specifically to revise the fully connected layers, which are a fundamental component of deep learning models.

Fully Connected Layer
— A Fundamental Component of DNNs

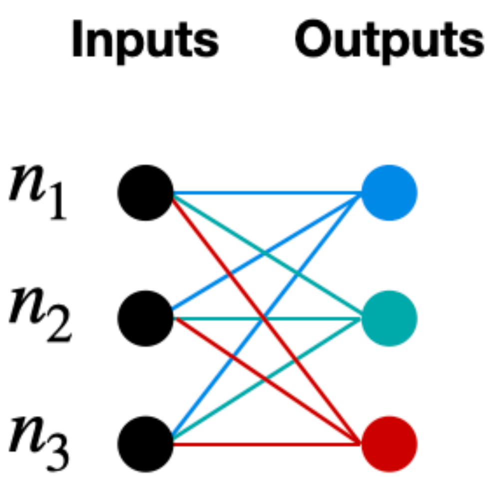
An example of a fully connected layer with 3 input variables and 3 output nodes



- **Fully connected layer interpretation**: Beyond being viewed as a matrix multiplication operation, a fully connected layer can also be seen as a multiple grouping mechanism.
- **Example of node calculation**: Each output node is calculated by a unique linear composite of each input variable.
  - For instance, the blue node is calculated as $w_1n_1 + w_2n_2 + w_3n_3$, where w_1, w_2, w_3 are model parameters.
- **Distinct groupings**: Different linear composites can be interpreted as distinct methods for grouping the input variables, as illustrated in the figure (see different colors of the edges).

Fully Connected Layer — A Fundamental Component of DNNs

An example of a fully connected layer with 3 input variables and 3 output nodes



- **Grouping methods determination**: The grouping methods within a fully connected layer are determined by the trained weights.
- **Potential redundancy and negative Impact**: Some of these trained weights may be redundant or have a negative impact on the model's performance.
- **Selective weight removal**: It can be beneficial to selectively remove weights in the fully connected layer.
- **Replacement with economically relevant grouping**: The removed weights can be replaced with grouping methods that have more relevance to economics.

The DKT Framework

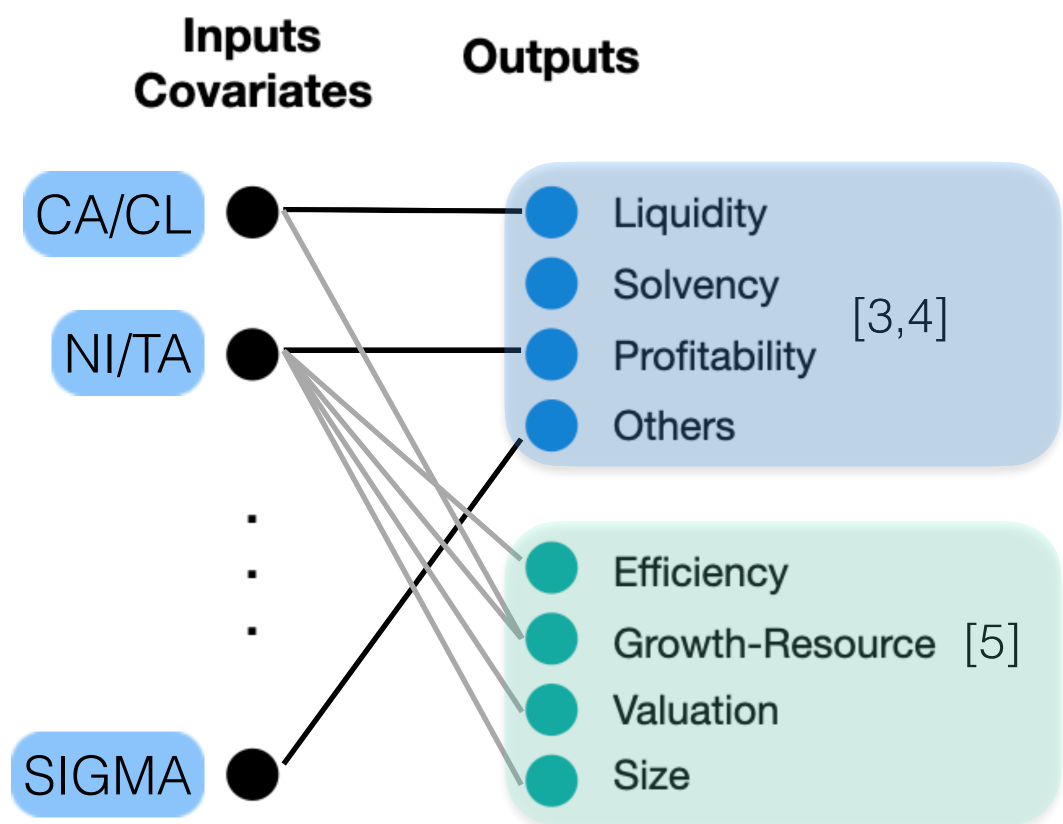
- Recall that $X_{i,t} = (x_{i,t,1}, x_{i,t,2}, \ldots, x_{i,t,k})$ is the set of the state variables (input) that affect the forward intensities for the i-th firm at the current time t.
- These variables may include two types of variables: macroeconomic factors and firm-specific attributes.
- The CRI database includes 16 variables for each firm-month observation, consisting of 4 common variables and 12 firm-specific variables.

1. **Common Variables:**
   - **Interest Rate**: This is measured using the 3-month short-term US Treasury bill rate.
   - **Stock Index Return**: This refers to the trailing one-year return on the S&P500 index.
   - **Financial Aggregate DTD**: This is the median distance to default (DTD) of financial firms in the US.
   - **Non-Financial Aggregate DTD**: This represents the median DTD of non-financial firms in the US.

Main ideas: We explicitly group the variables and prune the networks (i.e., remove some edges of the fully connected layers) to simplify the networks (less parameters).

2. **Firm-Specific Variables:**
   - **DTD**: The distance to default (DTD) of individual firms is used to measure volatility-adjusted leverage, following the methodology by Metron. For financial firms, DTD calculation is based on the setting in FIM.
   - **NI/TA**: This ratio of net income to total assets is used as a measure of a company's profitability.
   - **CASH/TA**: The logarithm of the ratio of the sum of cash and short-term investments to the total assets is used as a measure of a financial firm's liquidity.
   - **CA/CL**: The logarithm of the ratio of current assets to current liabilities serves as a measure of a non-financial firm's liquidity.
   - **Size**: This is measured by the logarithm of the ratio of a firm's market capitalization to the median market capitalization of the firms in the US over the past year.
   - **M/B**: The ratio of a firm's market-to-book asset ratio divided by the median market-to-book ratio of the firms in the US.
   - **SIGMA**: This is the 1-year idiosyncratic volatility, calculated following the method by Shumway. It's computed by regressing the daily return of a firm's market capitalization against the daily return of the S&P500 index, and defined as the standard deviation of the residuals from this regression.

Grouping the Covariates



- **Categorization of default and other-exit events**
- **Examples of covariates**:
  - The covariate "CA/CL" (logarithm of the ratio of current assets to current liabilities) is classified under "Liquidity."
  - The covariate "NI/TA" (ratio of net income to total assets) falls under the "Profitability" category.
  - The specifics of these grouping methods are further described in Appendix B.

[3] Zhang, L., Chen, S., & Zhang, X. (2005). Financial distress early warning based on MDA and ANN technique. *Systems Engineering*, 11, 50-58.
[4] Xie, C., Luo, C., & Yu, X. (2011). Financial distress prediction based on SVM and MDA methods: the case of Chinese listed companies. *Quality & Quantity*, 45, 671-686.
[5] Rodrigues, B. D., & Stevenson, M. J. (2013). Takeover prediction using forecast combinations. *International Journal of Forecasting,* 29(4), 628-641.

Dataset

- Experiments were conducted using the Credit Research Initiative (CRI) database from the Asian Institute of Digital Finance (AIDF) of the National University of Singapore.
  - Include data from 17,560 public firms in the US and contains a total of 1,833,106 firm-month observations from 1994 to 2021.
  - The annual default rate varies from 0.21% to 2.51%, while the rate of other exits ranges from 3.22% to 11.57%.
  - Variables:
    - The CRI database includes 16 variables for each firm-month observation, comprising 4 common variables and 12 firm-specific variables.
    - These variables were chosen for their predictive power in corporate defaults in the US [6].

Year	Active Firms	Default/bankruptcies	(%)	Other exit	(%)
1994	6915	17	0.25	223	3.22
1995	7395	16	0.22	362	4.90
1996	7947	17	0.21	401	5.05
1997	8305	48	0.58	568	6.84
1998	8270	75	0.91	891	10.77
1999	7961	85	1.07	921	11.57
2000	7624	106	1.39	782	10.26
2001	6930	174	2.51	757	10.92
2002	6229	118	1.89	533	8.56
2003	5825	80	1.37	472	8.10
2004	5664	37	0.65	371	6.55
2005	5649	35	0.62	384	6.80
2006	5591	21	0.38	382	6.83
2007	5611	23	0.41	463	8.25
2008	5275	58	1.10	382	7.24
2009	4983	105	2.11	322	6.46
2010	4855	29	0.60	313	6.45
2011	4704	32	0.68	304	6.46
2012	4591	39	0.85	262	5.71
2013	4621	28	0.61	239	5.17
2014	4772	27	0.57	212	4.44
2015	4858	40	0.82	275	5.66
2016	4802	65	1.35	362	7.54
2017	4710	42	0.89	311	6.60
2018	4737	20	0.42	262	5.53
2019	4772	33	0.69	292	6.12
2020	4967	70	1.41	238	4.79
2021	5785	17	0.29	242	4.18

[6] Credit Research Initiative. (2020). NUS Credit Research Initiative Technical Report. https://d.rmicri.org/static/pdf/Technical%20report_2020.pdf.

Experimental Setup

- **Cross-sectional experiments**
  - 1.8 million monthly samples were mixed and divided into training and testing sets at a 9:1 ratio.
  - The training set was further divided into a 9:1 ratio for sub-training and validation subsets.
    - ✓ The optimal number of training epochs was determined using this setup.
  - Notably, data samples from different periods were combined, a common practice in the machine learning literature.

The training and testing datasets have similar distributions.

Objective: Test the capability of the DNN models

- **Overtime experiments**
  - This setting uses an expanding window approach over time, useful for modeling time-dependent scenarios.
  - Initially, a 10-year training sample (from January 1994 to December 2003) is used.
  - Every month for the next year, predictions for 1 month to 5 years are made.
  - The model is retrained each December using the expanded dataset until the end of the dataset.
  - This results in out-of-sample predictions spanning 18 years (from 2004 to 2021).

The training and testing datasets may have dissimilar distributions.

Objective: Evaluate the model's ability to adapt to new incoming data, mirroring real-world applications.

Experimental Setup

We re-estimate the model at each year-end starting from the first month of 2004 and use only the data available at the time for estimation.

Data arrival time	Prediction time	
Train	**Test**	
1994/1/1 - 2003/12/31	2004/1/1 - 2004/12/31	
1994/1/1 - 2004/12/31	2005/1/1 - 2005/12/31	
...	...	
1994/1/1 - 2019/12/31	2020/1/1 - 2020/12/30	(Only 23 horizons can be predicted)
1994/1/1 - 2020/12/31	2021/1/1 - 2021/11/30	(Only 11 horizons can be predicted)

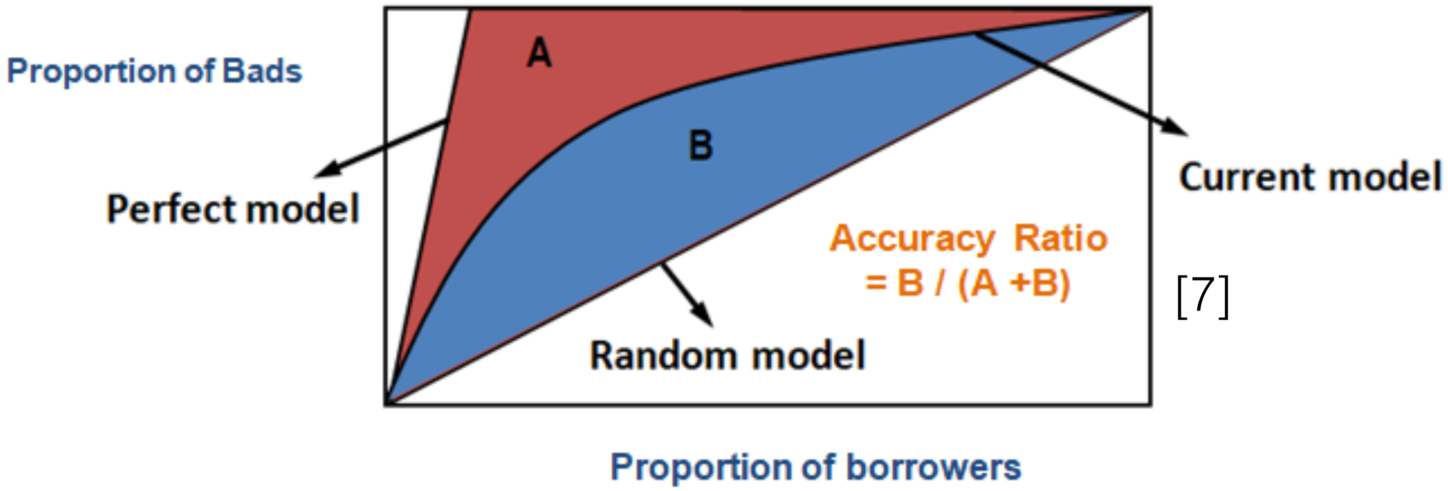
Validation

Cross-sectional: 9:1

Evaluation Metrics

Order

Accuracy Ratio
(AR, %)



[7]

Value

R-square
(Compared with FIM)

$$SS_{\mathrm{DKT}} = \sum_{i=1} (y_i - f_{i,\mathrm{DKT}})^2 = \sum_{i=1} e_i^2$$

$$SS_{\mathrm{FIM}} = \sum_{i=1} (y_i - f_{i,\mathrm{FIM}})^2 = \sum_{i=1} e_i^2$$

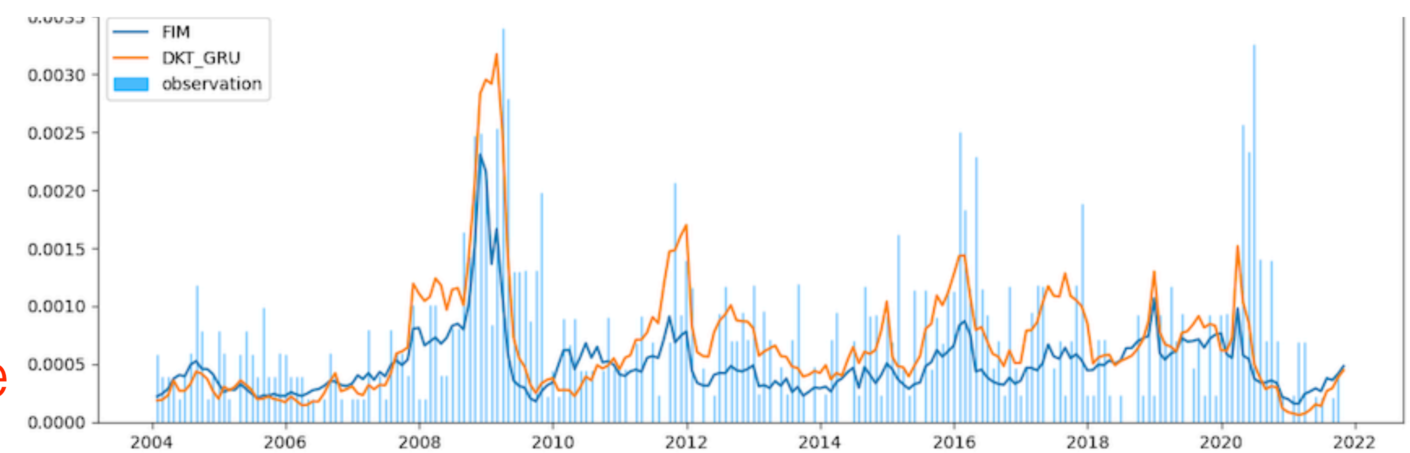
$$R^2 = 1 - \frac{SS_{\mathrm{DKT}}}{SS_{\mathrm{FIM}}}$$

[7] https://www.listendata.com/2019/09/gini-cumulative-accuracy-profile-auc.html

Evaluation Metrics

- Every month end, we calculate the predicted number of defaults amongst the active firms for a given prediction horizon.
- We then compare this with the observed number of defaults during the specified prediction period.

$m = 1$ (one month prediction)



Value

R-square
(Compared with FIM)

$$SS_{\text{DKT}} = \sum_{i=1} (y_i - f_{i,\text{DKT}})^2 = \sum_{i=1} e_i^2$$

$$SS_{\text{FIM}} = \sum_{i=1} (y_i - f_{i,\text{FIM}})^2 = \sum_{i=1} e_i^2$$

$$R^2 = 1 - \frac{SS_{\text{DKT}}}{SS_{\text{FIM}}}$$

The higher the better.

Results — Cross-sectional Experiments

Horizons (months)	1	3	6	9	12	24	36	48	60
Panel A	Accuracy ratio (AR) (%)								
FIM	95.443	93.337	91.178	86.746	86.192	76.925	69.649	64.687	60.070
MLP	96.144	94.317	92.538	89.174	88.693	81.771	75.783	70.794	66.418
GRU	**97.346**	**95.025**	**93.787**	**91.591**	**91.302**	**86.342**	**81.375**	**76.863**	**73.079**
DKT_GRU	97.330	94.912	93.364	90.645	90.311	84.844	79.678	74.807	70.666
Improvement (%)	1.994	1.809	2.861	5.585	5.928	12.241	16.837	18.822	21.655
Panel B	R-square (compared with FIM)								
MLP	0.037	0.059	0.096	0.176	0.193	0.280	0.354	0.320	0.273
GRU	0.025	**0.205**	**0.231**	**0.360**	**0.402**	**0.578**	0.579	0.479	0.431
DKT_GRU	**0.040**	0.177	0.223	0.332	0.379	0.553	**0.583**	**0.496**	**0.446**

- All neural models notably outperformed FIM across all prediction horizons.
  - Significant improvement highlights the potential of neural networks in cross-sectional default prediction.
- GRU-based models excelled, underscoring the importance of incorporating economic dynamics.
- GRU showed superior performance, thanks to its complex structure adeptly encapsulating the relationship between firms' variables and default events when training and testing datasets have similar label distributions.

Results — Over-time Experiments

Horizons (months)	1	3	6	9	12	24	36	48	60
Panel A	Accuracy ratio (AR) (%)								
FIM	93.538	92.191	90.040	86.383	85.619	76.410	68.086	60.356	53.915
MLP	93.445	92.195	89.856	85.830	85.000	74.169	65.814	58.851	52.765
GRU	94.268	93.143	91.515	88.667	88.018	78.472	70.856	64.483	59.294
DKT_GRU	**94.767**	**93.559**	**92.000**	**89.301**	**88.693**	**80.379**	**73.681**	**67.330**	**61.914**
Improvement (%)	1.314	1.483	2.177	3.378	3.591	5.193	8.218	11.556	14.837
Panel B	R-square (compared with FIM)								
MLP	0.110	0.123	−0.001	−0.046	−0.036	−0.101	−0.144	−0.092	0.053
GRU	−0.470	−0.486	−0.770	−0.594	−0.557	−0.475	−0.329	−0.243	−0.081
DKT_GRU	**0.156**	**0.315**	**0.279**	**0.160**	**0.155**	**0.098**	**0.370**	**0.554**	**0.757**

- MLP often performed worse than FIM in the overtime experiment, suggesting adding functional flexibility alone might not suffice.
- GRU outperformed MLP and FIM in terms of AR, but not in R-square, indicating the need of model regularization.
- Our proposed DKT (GRU) model outperformed other models in risk ranking and aggregate default distribution prediction for new incoming data, especially for long-term prediction horizons.

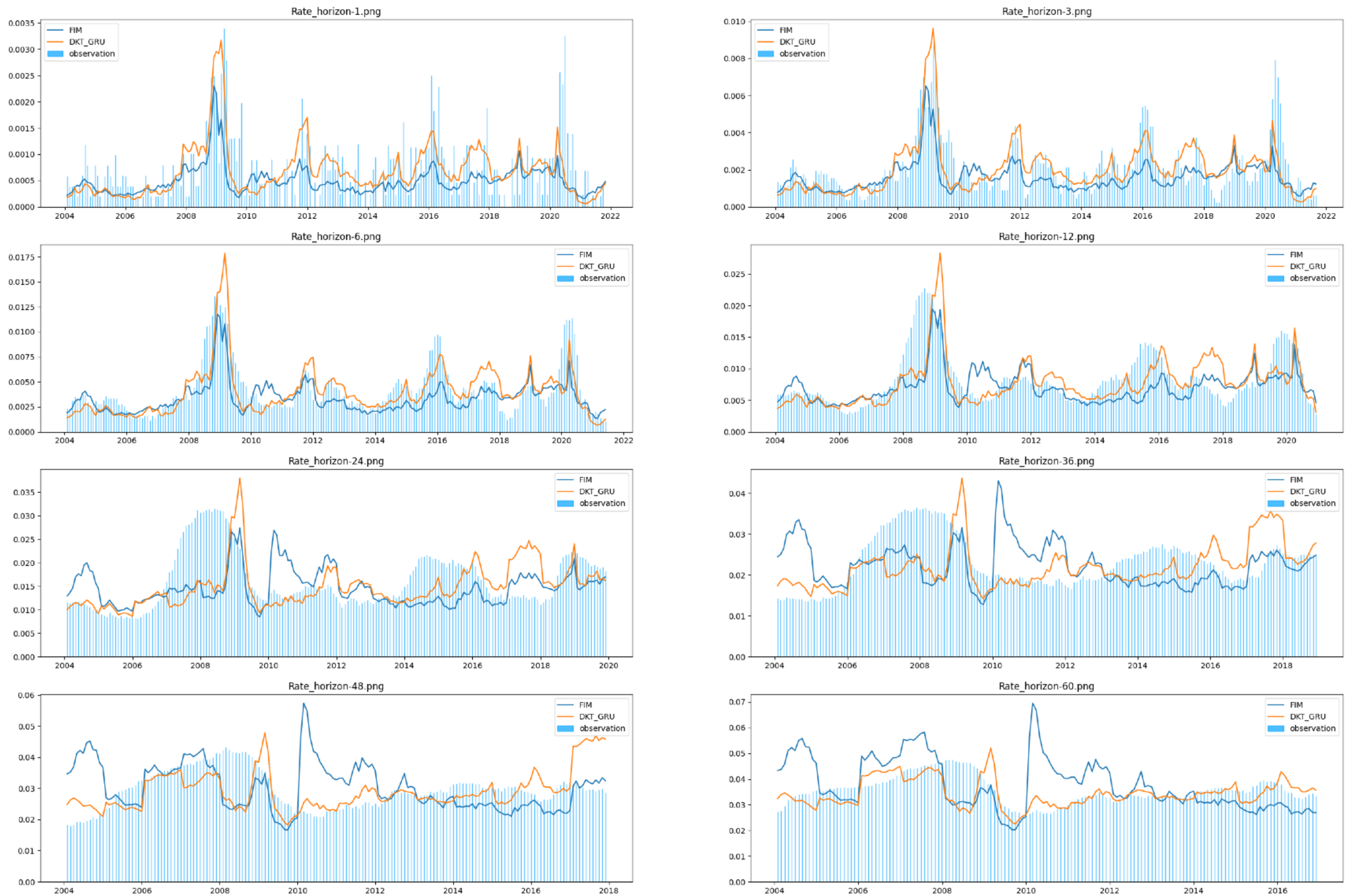
Results — Over-time Experiments

Horizons (months)	1	3	6	9	12	24	36	48	60
Panel A	Accuracy ratio (AR) (%)								
FIM	93.538	92.191	90.040	86.383	85.619	76.410	68.086	60.356	53.915
MLP	93.445	92.195	89.856	85.830	85.000	74.169	65.814	58.851	52.765
GRU	94.268	93.143	91.515	88.667	88.018	78.472	70.856	64.483	59.294
DKT_GRU	**94.767**	**93.559**	**92.000**	**89.301**	**88.693**	**80.379**	**73.681**	**67.330**	**61.914**
Improvement (%)	1.314	1.483	2.177	3.378	3.591	5.193	8.218	11.556	14.837
Panel B	R-square (compared with FIM)								
MLP	0.110	0.123	−0.001	−0.046	−0.036	−0.101	−0.144	−0.092	0.053
GRU	−0.470	−0.486	−0.770	−0.594	−0.557	−0.475	−0.329	−0.243	−0.081
DKT_GRU	**0.156**	**0.315**	**0.279**	**0.160**	**0.155**	**0.098**	**0.370**	**0.554**	**0.757**

- The performance difference between cross-sectional and overtime experiments underscores the impact of training and testing dataset distribution variation on standard neural model performance.
- The unmodified neural-based models may not be suitable for real-world applications due to these variations.

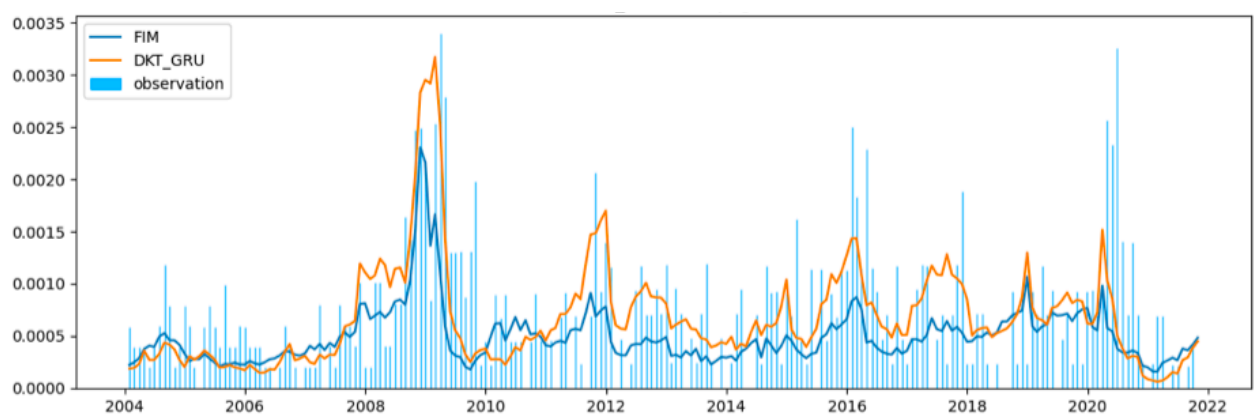
★ The long-term (e.g., 60-month) default prediction showed significant improvements, demonstrating the effectiveness of the DKT in preventing overfitting and improving performance.

Results — Over-time Experiments

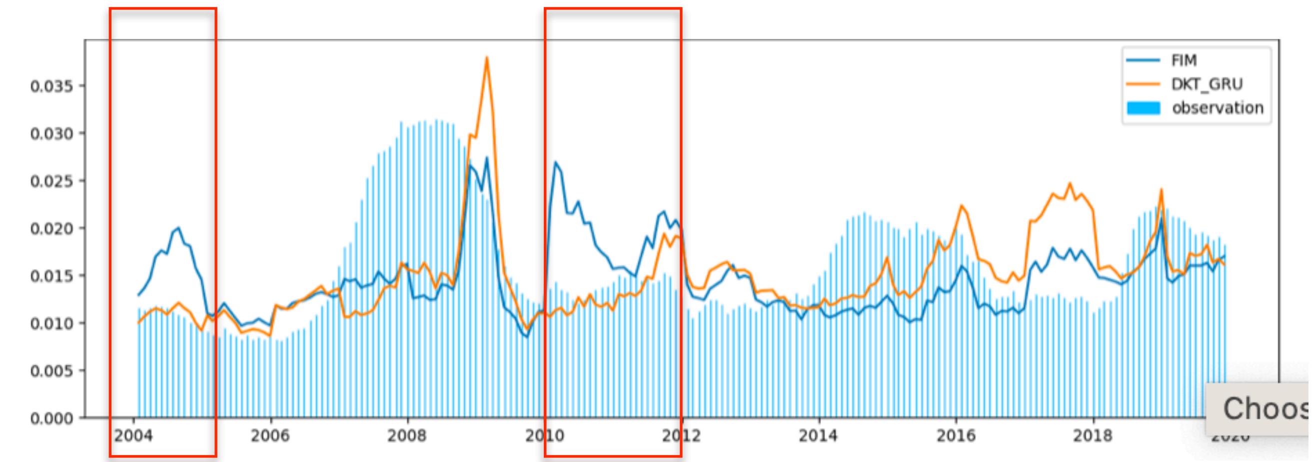


Results — Over-time Experiments

$m = 1$ (one month prediction)



$m = 24$ (two-year prediction)



- The models' predicted default rates closely match observed rates for short prediction horizons.
- As prediction horizons increase, a discrepancy arises between predicted and observed rates, suggesting a decline in model performance.
- Despite this discrepancy, the predictions of our DKT (GRU) are more stable over time, especially during 2004-2005 and 2010-2012 periods, than FIM's predictions.
- These observations suggest DKT (GRU) effectively regulates the model to yield more stable predictions.

Conclusions

Statistical Methods → Deep Learning Methods

Complex functional form
- Nonlinearity
- Capture time dynamics

Design deep neural networks based on FIM
- Generate consistent term structures of default probabilities
- Suitable for real-world scenarios

Domain knowledge tailored approach
- Prevent overfitting
- Good for real-world usage scenarios (overtime experiments)