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
- Linear composites of covariates
  (Lack of flexible functional form)

Risk classification
- Risk rankings

Statistical Methods
- Forward Intensity Model (FIM) [1]

Risk classification
- Single-period modeling (Unreasonable term structures)
- Flexible functional form (Overfitting)

Term structures of default probabilities
Number of default occurrences

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=\( \Delta t \), \( m = 0,1,2,3,\ldots \)
  - 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: \( p_o(X_{i,t}; m) = 1 - p_s(X_{i,t}; m) - p_d(X_{i,t}; m) = e^{-f_m(X_{i,t})} (1 - e^{-q_m(X_{i,t})\Delta t}) \)
- 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.

\[
f_m^{\text{FIM}}(X_{i,t}) = \exp \left( \beta_0(m) + \beta_1(m)x_{i,t,1} + \ldots + \beta_k(m)x_{i,t,k} \right)
\]

\[
= \exp \left( \beta(m) \cdot X_{i,t} \right)
\]

\[
q_m^{\text{FIM}}(X_{i,t}) = \exp \left( \bar{\beta}_0(m) + \bar{\beta}_1(m)x_{i,t,1} + \ldots + \bar{\beta}_k(m)x_{i,t,k} \right)
\]

\[
= \exp \left( \bar{\beta}(m) \cdot X_{i,t} \right)
\]

- \( \beta(m), \bar{\beta}(m) \): Coefficient vectors of the forward period \( m \)
View FIM as a Special Case of Deep Neural Networks

Default

\[ f_m^{\text{FIM}}(X_{i,t}) = \exp(\beta(m) \cdot X_{i,t}) \]

Other exit

\[ q_m^{\text{FIM}}(X_{i,t}) = \exp(\bar{\beta}(m) \cdot X_{i,t}) \]

\[
(f_m^{\text{MLP}}(X_{i,t}), q_m^{\text{MLP}}(X_{i,t}))_{m=0,1,\ldots,n-1} \rightarrow \Theta^{\text{MLP}}(X_{i,t}; m = 0,1,\ldots,n-1)
\]

- The MLP model generates the two types of forward intensities for all prediction horizons at once [2].
- \( \Theta^{\text{MLP}} \) 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.

Capture Time Dynamics of Covariates

\[
\left( f_m^{\text{MLP}}(X_{i,t}), q_m^{\text{MLP}}(X_{i,t}) \right)_{m=0,1,...,n-1} \rightarrow \Theta^{\text{MLP}}(X_{i,t}; m = 0,1,...,n-1)
\]

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

- 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.

**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).
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.

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].

<table>
<thead>
<tr>
<th>Year</th>
<th>Active Firms</th>
<th>Default/bankruptcies (%)</th>
<th>Other exit (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1994</td>
<td>6915</td>
<td>17 0.25</td>
<td>223 3.22</td>
</tr>
<tr>
<td>1995</td>
<td>7395</td>
<td>16 0.22</td>
<td>362 4.90</td>
</tr>
<tr>
<td>1996</td>
<td>7947</td>
<td>17 0.21</td>
<td>401 5.05</td>
</tr>
<tr>
<td>1997</td>
<td>8305</td>
<td>18 0.58</td>
<td>568 6.84</td>
</tr>
<tr>
<td>1998</td>
<td>8270</td>
<td>75 0.91</td>
<td>891 10.77</td>
</tr>
<tr>
<td>1999</td>
<td>7961</td>
<td>85 1.07</td>
<td>921 11.57</td>
</tr>
<tr>
<td>2000</td>
<td>7624</td>
<td>106 1.39</td>
<td>782 10.26</td>
</tr>
<tr>
<td>2001</td>
<td>6930</td>
<td>174 2.51</td>
<td>757 10.92</td>
</tr>
<tr>
<td>2002</td>
<td>6229</td>
<td>118 1.89</td>
<td>533 8.56</td>
</tr>
<tr>
<td>2003</td>
<td>5825</td>
<td>80 1.37</td>
<td>472 8.10</td>
</tr>
<tr>
<td>2004</td>
<td>5664</td>
<td>37 0.65</td>
<td>371 6.55</td>
</tr>
<tr>
<td>2005</td>
<td>5649</td>
<td>35 0.62</td>
<td>384 6.80</td>
</tr>
<tr>
<td>2006</td>
<td>5591</td>
<td>21 0.38</td>
<td>392 6.83</td>
</tr>
<tr>
<td>2007</td>
<td>5611</td>
<td>23 0.41</td>
<td>463 8.25</td>
</tr>
<tr>
<td>2008</td>
<td>5275</td>
<td>58 1.10</td>
<td>382 7.24</td>
</tr>
<tr>
<td>2009</td>
<td>4983</td>
<td>105 2.11</td>
<td>322 6.46</td>
</tr>
<tr>
<td>2010</td>
<td>4855</td>
<td>29 0.60</td>
<td>313 6.45</td>
</tr>
<tr>
<td>2011</td>
<td>4704</td>
<td>32 0.68</td>
<td>304 6.46</td>
</tr>
<tr>
<td>2012</td>
<td>4591</td>
<td>39 0.85</td>
<td>262 5.71</td>
</tr>
<tr>
<td>2013</td>
<td>4621</td>
<td>28 0.61</td>
<td>239 5.17</td>
</tr>
<tr>
<td>2014</td>
<td>4772</td>
<td>27 0.57</td>
<td>212 4.44</td>
</tr>
<tr>
<td>2015</td>
<td>4858</td>
<td>40 0.82</td>
<td>275 5.66</td>
</tr>
<tr>
<td>2016</td>
<td>4802</td>
<td>65 1.35</td>
<td>362 7.54</td>
</tr>
<tr>
<td>2017</td>
<td>4710</td>
<td>42 0.89</td>
<td>311 6.60</td>
</tr>
<tr>
<td>2018</td>
<td>4737</td>
<td>20 0.42</td>
<td>262 5.53</td>
</tr>
<tr>
<td>2019</td>
<td>4772</td>
<td>33 0.69</td>
<td>292 6.12</td>
</tr>
<tr>
<td>2020</td>
<td>4967</td>
<td>70 1.41</td>
<td>238 4.79</td>
</tr>
<tr>
<td>2021</td>
<td>5785</td>
<td>17 0.29</td>
<td>242 4.18</td>
</tr>
</tbody>
</table>

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.

- **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 have similar distributions.

Objective: Test the capability of the DNN models

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.

<table>
<thead>
<tr>
<th>Data arrival time</th>
<th>Prediction time</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Train</strong></td>
<td><strong>Test</strong></td>
</tr>
<tr>
<td>1994/1/1 - 2003/12/31</td>
<td>2004/1/1 - 2004/12/31</td>
</tr>
<tr>
<td>1994/1/1 - 2004/12/31</td>
<td>2005/1/1 - 2005/12/31</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>1994/1/1 - 2019/12/31</td>
<td>2020/1/1 - 2020/12/30</td>
</tr>
<tr>
<td>1994/1/1 - 2020/12/31</td>
<td>2021/1/1 - 2021/11/30</td>
</tr>
</tbody>
</table>

Validation

Cross-sectional: 9:1
Evaluation Metrics

Order

Accuracy Ratio (AR, %)

Value

R-square (Compared with FIM)

\[ SS_{DKT} = \sum_{i=1}^{n} (y_i - f_{i,DKT})^2 = \sum_{i=1}^{n} e_i^2 \]

\[ SS_{FIM} = \sum_{i=1}^{n} (y_i - f_{i,FIM})^2 = \sum_{i=1}^{n} e_i^2 \]

\[ R^2 = 1 - \frac{SS_{DKT}}{SS_{FIM}} \]

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.

\[
R^2 = 1 - \frac{SS_{DKT}}{SS_{FIM}}
\]

\[
SS_{DKT} = \sum_{i=1}^{m} (y_i - f_{i,DKT})^2 = \sum_{i=1}^{m} e_i^2
\]

\[
SS_{FIM} = \sum_{i=1}^{m} (y_i - f_{i,FIM})^2 = \sum_{i=1}^{m} e_i^2
\]

The higher the better.
Results — Cross-sectional Experiments

<table>
<thead>
<tr>
<th>Horizons (months)</th>
<th>1</th>
<th>3</th>
<th>6</th>
<th>9</th>
<th>12</th>
<th>24</th>
<th>36</th>
<th>48</th>
<th>60</th>
</tr>
</thead>
<tbody>
<tr>
<td>Panel A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FIM</td>
<td>95.443</td>
<td>93.337</td>
<td>91.178</td>
<td>86.746</td>
<td>86.192</td>
<td>76.925</td>
<td>69.649</td>
<td>64.687</td>
<td>60.070</td>
</tr>
<tr>
<td>MLP</td>
<td>96.144</td>
<td>94.317</td>
<td>92.538</td>
<td>89.174</td>
<td>88.693</td>
<td>81.771</td>
<td>75.783</td>
<td>70.794</td>
<td>66.418</td>
</tr>
<tr>
<td>GRU</td>
<td>97.346</td>
<td>95.025</td>
<td>93.787</td>
<td>91.591</td>
<td>91.302</td>
<td>86.342</td>
<td>81.375</td>
<td>76.863</td>
<td>73.079</td>
</tr>
<tr>
<td>DKT_GRU</td>
<td>97.330</td>
<td>94.912</td>
<td>93.364</td>
<td>90.645</td>
<td>90.311</td>
<td>84.844</td>
<td>79.678</td>
<td>74.807</td>
<td>70.666</td>
</tr>
<tr>
<td>Panel B</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MLP</td>
<td>0.037</td>
<td>0.059</td>
<td>0.096</td>
<td>0.176</td>
<td>0.193</td>
<td>0.280</td>
<td>0.354</td>
<td>0.320</td>
<td>0.273</td>
</tr>
<tr>
<td>GRU</td>
<td>0.025</td>
<td>0.205</td>
<td>0.231</td>
<td>0.360</td>
<td>0.402</td>
<td>0.578</td>
<td>0.579</td>
<td>0.479</td>
<td>0.431</td>
</tr>
<tr>
<td>DKT_GRU</td>
<td>0.040</td>
<td>0.177</td>
<td>0.223</td>
<td>0.332</td>
<td>0.379</td>
<td>0.553</td>
<td>0.583</td>
<td>0.496</td>
<td>0.446</td>
</tr>
</tbody>
</table>

- 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

<table>
<thead>
<tr>
<th>Horizons (months)</th>
<th>1</th>
<th>3</th>
<th>6</th>
<th>9</th>
<th>12</th>
<th>24</th>
<th>36</th>
<th>48</th>
<th>60</th>
</tr>
</thead>
<tbody>
<tr>
<td>Panel A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FIM</td>
<td>93.538</td>
<td>92.191</td>
<td>90.040</td>
<td>86.383</td>
<td>85.619</td>
<td>76.410</td>
<td>68.086</td>
<td>60.356</td>
<td>53.915</td>
</tr>
<tr>
<td>MLP</td>
<td>93.445</td>
<td>92.195</td>
<td>89.856</td>
<td>85.830</td>
<td>85.000</td>
<td>74.169</td>
<td>65.814</td>
<td>58.851</td>
<td>52.765</td>
</tr>
<tr>
<td>GRU</td>
<td>94.268</td>
<td>93.143</td>
<td>91.515</td>
<td>88.667</td>
<td>88.018</td>
<td>78.472</td>
<td>70.856</td>
<td>64.483</td>
<td>59.294</td>
</tr>
<tr>
<td>DKT_GRU</td>
<td>94.767</td>
<td>93.559</td>
<td>92.000</td>
<td>89.301</td>
<td>88.693</td>
<td>80.379</td>
<td>73.681</td>
<td>67.330</td>
<td>61.914</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Improvement (%)</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Panel B</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GRU</td>
<td>-0.001</td>
<td>-0.046</td>
<td>-0.036</td>
<td>-0.101</td>
<td>-0.144</td>
<td>-0.092</td>
<td>0.053</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DKT_GRU</td>
<td>-0.470</td>
<td>-0.486</td>
<td>-0.770</td>
<td>-0.594</td>
<td>-0.557</td>
<td>-0.475</td>
<td>-0.329</td>
<td>-0.243</td>
<td>-0.081</td>
</tr>
</tbody>
</table>

- 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

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

- 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
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

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)