Datasheet for FinSurvival

Aaron Green

May 2025

1 Datasheet Questions

Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

The FinSurvival dataset was created to provide large-scale, publicly accessible, realistic datasets for benchmarking artificial intelligence (AI) survival models, specifically in financial contexts using decentralized finance (DeFi) transaction data.

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

The dataset was created by Aaron Green, Zihan Nie, Hanzhen Qin, Oshani Seneviratne, and Kristin P. Bennett from Rensselaer Polytechnic Institute.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

The authors acknowledge the support from NSF IUCRC CRAFT center research grants (CRAFT Grants #22003, #22006) for this research. The opinions expressed in this publication and its accompanying code base do not necessarily represent the views of NSF IUCRC CRAFT.

Any other comments?

None.

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

Each instance represents a survival record derived from DeFi transactions on the Aave V2 Mainnet (Ethereum) protocol. Each instance includes an index event (e.g., borrow) and an outcome event (e.g., repay) with associated survival times and censoring status. Additional features for each record are included.

How many instances are there in total (of each type, if appropriate)?

The dataset comprises 7,698,497 instances which are organized into 16 distinct survival modeling tasks.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

The dataset contains all possible instances across the 16 survival models it contains.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.

Each instance consists of both raw data from DeFi transactions and processed, engineered features derived from the raw transaction data. The features include transaction attributes such as transaction amounts, currency details, timestamps, categorical transaction types, and engineered features summarizing user and market histories, as well as cyclical temporal attributes.

Is there a label or target associated with each instance? If so, please provide a description.

Each instance has two labels/targets which are standard for survival analysis: "timeDiff" representing how much time passed between the index and outcome events, and "status" representing whether the event saw an outcome or was censored.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

There is no missing information. However, there could be NA values for some features in some instances. These can occur in the engineered features because some of the user/market historical features have not yet seen activity to populate certain features.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

Yes, relationships between individual instances are explicitly defined by the combination of user and coin identifiers associated with each transaction. Each instance is linked to others through shared users, coins, and transaction histories.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

There is a recommended split of the data into training and testing sets. This split has been made explicit in how the data is structured for release. We have sixteen distinct datasets, each with a unique combination of index

and outcome event. There are four distinct index events, and the features are computed based on the index events. Thus, we have already split the features from the target variables in the form of X and y, respectively. Each dataset was split into training and testing data by using all the data prior to July 1, 2022 as the training data and all data after July 1, 2022 as testing data. This allocates approximately 60% of the data as training data and 40% as testing data.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

There should be no errors in the data. There may appear to be redundancies, as some IDs appear multiple times, but these are okay because multiple transactions can occur with the same blockchain transaction ID, so they can represent distinct data points even with the same ID on occasion.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

The dataset is self-contained.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals non-public communications)? If so, please provide a description.

The dataset is entirely derived from public blockchain transactions and so does not contain any confidential data.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

No, the dataset does not contain offensive, insulting, threatening, or anxiety-inducting data.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

Yes, the dataset relates to people and their transactions in a Decentralized Finance lending protocol.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

The dataset does not identify any subpopulations. It contains no personal identification information.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.

The dataset contains user IDs which represent blockchain addresses on the

Ethereum blockchain. There is no direct way to identify individuals from this unless they have made their blockchain address public.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

This dataset does contain real financial data and thus is sensitive in that regard. However, since the data comes from the public Ethereum blockchain and contains no personal identification information, we do not believe this to be sensitive data.

Any other comments?

None.

Collection Process

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

The raw data was collected from TheGraph[1]. The developers of the Aave protocol maintain a subgraph that is freely accessible where all the data from the protocol can be pulled. Engineered features were created using only the raw data from this source.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

The data was collected using a software API. It used an API from The Graph, programmatically accessing the API using the R programming language. The API allows access to a set of data tables from which the raw data was collected and aggregated.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

The dataset is not a sample; it contains all instances of transactions from Aave.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

The data was collected by undergraduate and graduate students at Rensselaer Polytechnic Institute as part of undergraduate research and PhD research. The students were compensated with typical stipend amounts or credit as part of their overall scholarly activities.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the

[1] thegraph.com

data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

The data was originally collected in the summer of 2021, acquiring all transactions from Aave V2 dating back to its deployment in November 2020. The data was periodically updated with new transactions. The final update of the data from which our dataset is derived was in September 2024.

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

None.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

The dataset does relate to people.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

The data came from a subgraph on The Graph maintained by the developers of Aave, the protocol from which all of this data is derived.

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

No, individuals were not explicitly notified because the dataset consists entirely of publicly available blockchain transaction data.

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

By participating in transactions on a public blockchain, users implicitly consented to their data being publicly accessible and available for analysis.

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

No, explicit consent was not obtained because the data originates from public blockchain transactions. By engaging with the public Ethereum blockchain, users implicitly consented to the permenant and transparent recording of their transaction data. Due to the immutable nature of blockchain records, users do not have a mechanism to revoke consent or remove their data from the public record once the transactions are completed.

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

No, a formal analysis of the potential impact of the dataset on data subjects has not been conducted. Given that the dataset is derived entirely from publicly available blockchain transaction data containing no personally identifiable information or intellectual property, such analysis was deemed unnecessary.

Any other comments?
None.

Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

Yes, significant preprocessing and cleaning was applied to create the FinSurvival dataset from raw DeFi transaction data. These steps included filtering and joining transaction records to define time-to-event instances and hand-engineering additional features based on the raw form of the data.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

The raw transaction data was saved in addition to the FinSurvival dataset. We provide a 5% sample of the raw transaction data in the Github repository [2] alongside the survival data, but the full data can be acquired through The Graph [3].

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

Yes, the code used to turn the transaction data into the FinSurvival dataset is available at the Github repository [4] associated with this dataset.

Any other comments?
None.

Uses

Has the dataset been used for any tasks already? If so, please provide a description.

Early versions of the dataset were used in two papers by the authors presenting preliminary results on the application of survival analysis to DeFi transactions.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

Yes, there is a page on the Github with references to the other papers that used early versions of this dataset.

What (other) tasks could the dataset be used for?

The dataset could be adapted to incorporate competing risks, allowing for more than one possible outcome event for each index event.

[2] https://github.com/Large-Transaction-Models/DMLR_DeFi_Survival_Benchmark
[3] https://thegraph.com/explorer/subgraphs/C2zniPn45RnLDGzVeGZCx2Sw3GXrbc9gL4ZfL8B8Em2j?view=Query&chain=arbitrum-one
[4] https://github.com/Large-Transaction-Models/DMLR_DeFi_Survival_Benchmark

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

None.

Are there tasks for which the dataset should not be used? If so, please provide a description.

Yes, the dataset should not be used for tasks that require personally identifiable information, behavioral profiling tied to real-world identities, or any analysis that implies private user attributes. Since the data is derived from pseudonymous blockchain activity, it is inappropriate to use this dataset for user-level targeting, deanonymization, or surveillance-related tasks. The dataset is intended solely for research and benchmarking in machine learning, particularly in the context of survival modeling and financial behavior in DeFi protocols.

Any other comments?

None.

Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

No.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub) Does the dataset have a digital object identifier (DOI)?

The dataset will be distributed on GitHub. It does not have a DOI.

When will the dataset be distributed?

The dataset will be distributed alongside the submission of the associated paper to DMLR in May 2025.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

No.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

No.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

No.

Any other comments?

None.

Maintenance

Who will be supporting/hosting/ maintaining the dataset?
The dataset will be hosted on GitHub and maintained by the authors of the dataset.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?
The manager of the dataset can be reached at aaronmicahgreen@gmail.com.

Is there an erratum? If so, please provide a link or other access point.
No.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?
There are no plans to update the dataset, but if mistakes are found and communicated to the owner/manager of the dataset, we will do our best to fix the errors and push the changes to the GitHub repository.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.
There are no limits on retention of the data.

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.
There is only one version of the dataset. We don't plan to update the dataset, but may fix any errors in the data if users find them. These updates will be posted on the Github where the dataset is maintained.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.
The code for creating the dataset based on tabular transaction data is included in the Github repository, and as such, others could conceivably extend the dataset or create similar datasets using the same methods. We don't plan to officially incorporate updates in the dataset, however.

Any other comments?
None.