# Churn Prediction in Mobile Social Games: Towards a Complete Assessment using Survival Ensembles

1. Overview

* Churn prediction models -> understand player loyalty and anticipate when they will stop playing a game -> actions taken to retain players
* Survival analysis: focus on predicting the time of occurrence of a certain event
* Classical method (e.g., regression): could be applied only when all players have left the game
* Challenge: incomplete churning data for all players (most of them still play the game) – censored data problem
* Censoring – commonly dealt with by survival analysis techniques >< inflexibility of survival statistical algorithms – poor accuracy
* Novel ensemble learning techniques -> high-class prediction results
* Survival ensemble model – comprehensive analysis & accurate prediction of churn
* Predict prob of churning as function of time -> distinguish various levels of loyalty profiles
* Assess risk factors that explain the predicted player survival times
* Significant improves accuracy and robustness of traditional analyses (e.g., Cox regression)

1. Introduction

* Social games – target new audience of players: causal players & new monetization model: free-to-play (F2P or freemium)
* Ability to predict when a player will leave a game -> take incentive actions to re-engage them and prevent churn or move them to another game of the company
* F2P monetization – main model used by mobile social games, involves a non-contractual relationship
* Churn is not clearly determined by an explicit statement ending the contract
* Define churn as a prolonged period of inactivity
* Focus: predicting churn for high-value players (whales)
* Behave differently than average players
* Often the most active players -> easily define churn as a prolonged period of inactivity
* High level of engagement -> collect more data about activity & more likely to answer positively to actions taken in order to prevent churn
* Top spenders -> revenues for business
* Classical approaches to churn: binary classification problem
* Intuitive >< not able to predict when the player will stop playing
* Features are limited to provide static (non-temporal) info
* Model time until churn:
* Regression – only appropriate when all players have stopped playing the game
* Challenge: incomplete data
* Survival ensembles:
* Outputs accurate predictions of when players churn
* Info about risk factors affecting churn
* Identify possible churners & survival prob function for each player over time
* Distinguish various levels of loyalty profiles: upcoming, near-future, far-future churners
* Variables influence survival behavior
* Median survival time – life expectancy threshold
* Label players as being at risk of churning
* Take action to retain

1. **Survival ensemble models**
2. **Survival analysis**

* Focus on studying time until event of interest happens & relationship with different factors
* Time-to-event outcome -> survival time
* Censored data – measurements only contain info if event occurs or not before given time t
* **Survival function – likelihood of survival at a certain time t** – estimated through the non-parametric Kaplan-Meier estimator
* **Cumulative survival prob**:
* Presence of competing risks – more than 1 possible failure event
* Alternative events can prevent the observation of main event of interest
* Here we focus on **loss of interest in a game – main cause of churn**
* **Cox proportional-hazards** – semi-parametric survival techniques: **Estimated hazard for k individual players and p covariate vectors**  -> Solves censoring problem by maximizing the partial likelihood
* Allow regressions to work with censored data & permit intuitive interpretation of the impact of features >< assume a fixed link between the output and the variables (assuming them additive and constant over time)
* Requires explicit specification of the relationship by researcher and involves important efforts in terms of model selection and evaluation
* Difficulties to scale with big data problems -> **alternative regularized versions of Cox regression** proposed to amend this >< still based on restrictive assumptions that are not easy to fulfill
* **Accelerated failure time models** – parametric approach: Type of distribution previously determined
* **Suboptimal** because it is uncommon that the data follow these specific distribution shapes

1. **Survival trees and ensembles**

* **Decision trees:**
* Classification and regression trees (CART) – **non-parametric techniques**
* **Split the feature space recursively** to group subjects with homogeneous characteristics and separate those with bigger differences based on the outcome
* To perform the nodes classification and maximize homogeneity within the nodes -> **minimize impurity** (e.g., **cross-entropy, sum of squared errors**)
* **Survival trees:**
* Constructed as a **set of binary trees that grow by recursive partitioning** of the sample space χ, where tree nodes are subspaces of χ.
* Tree splitting starts in the root node
* Based on a **survival statistical criterion** (e.g., cumulative hazard function or Kaplan-Meier estimates), root node is then divided into 2 children nodes
* **Principle for partitioning: maximize the survival difference between 2 groups & maximizing homogeneity among nodes,** based on survival experience
* Best split achieved by exploring all combinations, considering all the predictor variables and all the possible splits
* Limitation: employing a single tree can produce **instability in predictions**
* Use an ensemble of them, instead of using a single tree
* **Survival ensembles:**
* Using an ensemble of models, instead of a single one
* Ensemble-based learning methods – growing **a set of survival trees**, instead of a single one
* **2 main techniques:**
  + Random survival forest
  + Conditional inference survival ensembles
* **Conditional inference survival ensembles** – chosen method
* Uses a **weighted Kaplan-Meier function** based on the measurements used for training
* **Ensemble survival function:**

Where:

* + n: number of trees in the ensembles (n = 1, 2,…, N)
  + : covariates
  + : uncensored events until time t
  + number of individuals at risk at time t
* Introduces **additional weight to the nodes where there more subjects at risk**
* Uses **linear rank statistics as splitting criterion** to grow the trees
* **Random survival forests:**
* Based on **Nelson-Aalen estimates**
* **Split criterion**: Maximum of the **log-rank statistical test**
* **Biased results** in favor of covariates with many splits

1. Dataset

* Data from a major mobile social game. Several churn predictors or risk factors investigated
* **Game-independent features** (i.e., features not related to the game mechanics and can be measured in any game) -> model applicable to other games
* Player attention: Time component of the player accessing the game
* Player loyalty: Frequency of the player access to the game
* Playing intensity: quality of the playing sessions
* Player level: Value of this variable and its evolution depends on the game (player level is present and measurable in many games, can be considered as a game-independent predictor)

|  |  |  |
| --- | --- | --- |
| **Feature group** | **Feature** | **Description** |
| Player attention | Time spent per day in the game | Averages over the first weeks and moving average over the last weeks |
| Lifetime | Number of days since registration until churn (in case of churn) |
| Player loyalty | Number of days with at least one playing session |  |
| Loyalty index | Number of days played / Lifetime |
| Days from registration to first purchase |  |
| Days since last purchase |  |
| Playing intensity | Number of actions |  |
| Number of sessions |  |
| Number of in-app purchases |  |
| Amount of in-app purchases |  |
| Action activity distance | Euclidean distance between average number of actions over lifetime and average number of actions over the last days |
| Player level | Player level |  |

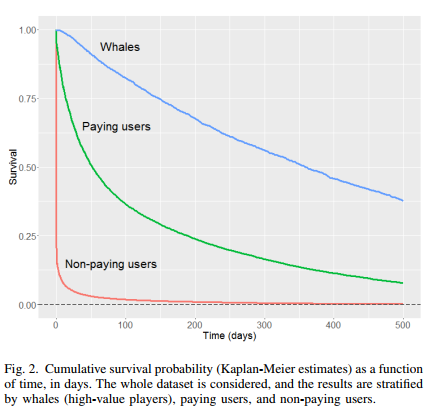
* **Game-dependent features** (did not keep in model)
* Participation in a guild
* Measure of number of actions by category (shop, battle, mission)

1. Modeling
2. Churn definition

* We consider that a player has churned if they do not connect to the game for 10 consecutive days
* Conventional churn prediction – solved from a static POV, binary classification problem
* Our Focus: when churn will happen
* Model churn behavior from the perspective of survival analysis
* Learning samples: n = 2500 whales

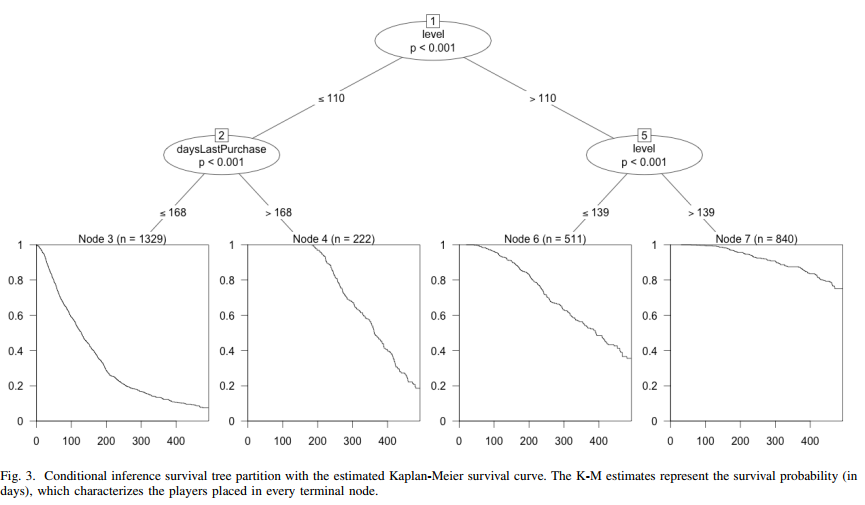
1. Kaplan-Meier estimates

* Plot KM survival curves stratified by whales, normal paying users, and non-paying users (Sample of 1500000 players)

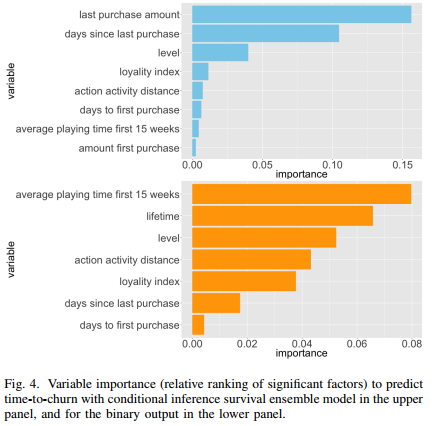


1. Churn model as a censored data problem

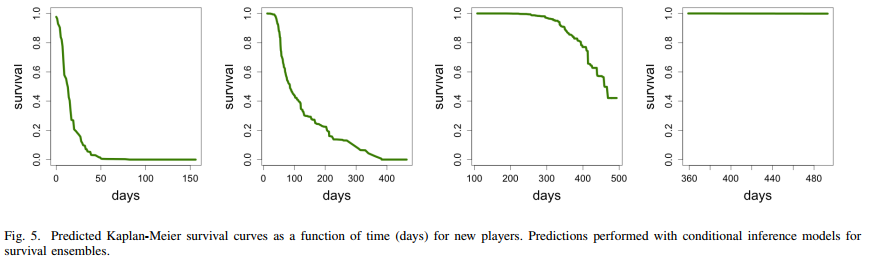
* Conditional inference survival ensembles to model game churn
* 1000 conditional inference trees used as a base learner to predict exit time of whales from the game
* How conditional inference trees work: simple partition with 2 terminal nodes
* Root node var: last level the player reached in the game
* Children nodes partitions: one also based on level and another based on the number of days since last in-app purchase



* Outcome: Overall survival time
* Most significant predictors – Variable importance computed using Brier score (IBS) & Feature selection performed based on it



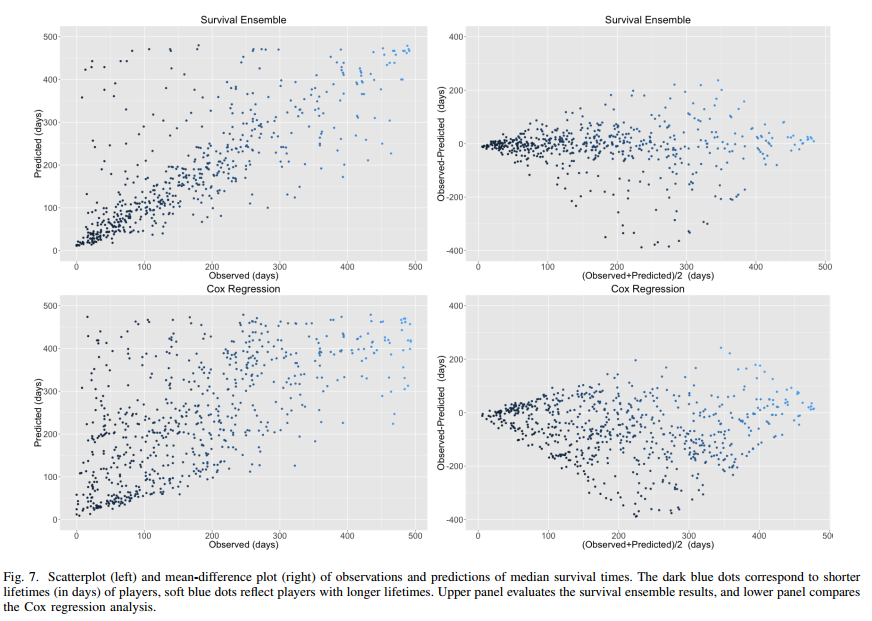
* Resulting prediction: For each player: A survival function – prob of churn as a function of time since registration of the game
* Distinguish different player profiles and survival behaviors:
  + Players who are going to churn soon
  + Players expected to churn in far future
  + Loyal player
* Classify and predict loyalty for every player, taking into account the temporal dimension



* Median survival time (time when the percentage of surviving in the game is 50%) – threshold to categorize a player as being at risk of churning

1. Model validation

* Censoring -> standard methods of visualizing and evaluating prediction performances are not suitable
* Fit of proposed conditional inference survival ensemble method and selected Cox regression (same predictors)
* As long as the censoring rate grows, the prediction capability diminishes



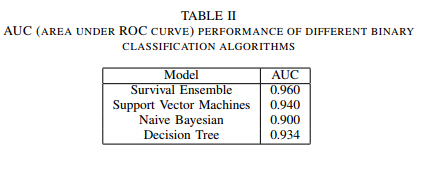
* Cumulative prediction error curve & Integrated Brier score
* Ensemble-based approach improves accuracy over the Cox model
* Prediction error function reaches max at median survival time of 304 days and 306 days for Cox regression and ensemble model

|  |  |
| --- | --- |
|  |  |

* Paired t-test (Welch Two sample t-test) – estimate whether the prediction ability of a model is statistically significant from another (95% confidence interval)
* Survival ensemble model is statistically significant

1. Comparison with other model approaches

* Binary classification model of churners
* Useful insight for a very short-term prediction
* Easy to interpret and implement
* Same algorithm of conditional inference ensembles >< different outcome
* Binary variable – whether player churns (yes or no)
* Train model with several sets of features to obtain final list of attributes (Fig. 4 above)
* Contrasting results of variable impact between the survival model and binary classification
* Comparison with other binary classification methods: SVM, naïve Bayesian classifier, decision tree



* Techniques applied above are powerful >< in their original form, they cannot handle the assimilation of information from censored data

1. Summary and conclusion

* Propose application of conditional inference survival ensembles to predict time-to-churn and survival prob of players in games in terms of game lifetime
* Main motivation: flexible technique that does not require a previous manipulation of data and able to deal with temporal dimension of the churn prediction problem
* Provided more accurate and more stable prediction results than traditional approaches
* Unbiased, does not overfit & provide robust info about the risk factors influencing players to churn
* Further on-going work: improvement of accuracy in the prediction of time-to-churn for players who stay longer in the game -> discover more significant features