Survival Analysis on New Users' Behaviours

1. Introduction

The survival analysis is defined as the method to analyze data of the time to event outcomes. Usually, the response in the analysis is about the survival time, the death time or the event time. The survival analysis has been widely used in medical research. However, such an analysis method can also be applied to other areas, as long as the target response can be measured as an occurrence of events. In this report, the survival analysis will be used to study new users' first purchase behaviours, which is known as the new users' transmission rate of purchasing.

This project's main objective is to study the transmission rate of purchasing by using the traditional methods in survival analysis. The transmission rate of purchasing can be observed by comparing the time needed for a new user to make the first purchasing. In this sense, the event in the study will be the status of purchasing, and the time of such occurrence of events will be the durations for purchasing. Several factors may affect the users' behaviours, including gender, age and living locations. An analysis of the influence of these factors will be conducted to suggest a set of characteristics of new users who have higher purchasing potential.

2. Exploratory Data Analysis (EDA)

2.1 Data Overview

The data used in the project records new users' behaviours in a shopping platform about the purchasing decisions. The main variables in the dataset are summarized in the following table 1.

Table 1 Description of the Data

Category	**Variables**	**Type**	**Range and Quantity**
User's Identity	User_id	Integer	506671 – 533230 (15217)
Event Time	Signup_time End_time Duration	Time Time Continuous	2017-09-01 – 2018-05-31 2017-09-01 – 2018-06-01 0.00110 – 272.63499 (days)
Event Status	Have_bought	Boolean	True (11659), False (3558)
User's Characteristics	Sex Birth_year Province	Binary Continuous Categorical	Female (9691), Male (5526) 1931 – 2007 (year) 30 provinces

The event time and event status are essential variables for survival analysis, while the user's characteristics will contribute to the difference between individuals or groups in purchasing behaviours.

In total, there are 15217 rows of records in the dataset. Missing values exist in the variable "birth_year", and there are 1294 missing observations among the 15217 records. The percentages of purchasing behaviours among the group with missing "birth_year" and the complete group are summarized in table 2.

Table 2 Summary of the Missing Values

Group	Missing "birth_year"	Complete
Percentage of Purchasing	76.89% (995/1294)	76.59% (10664/13923)

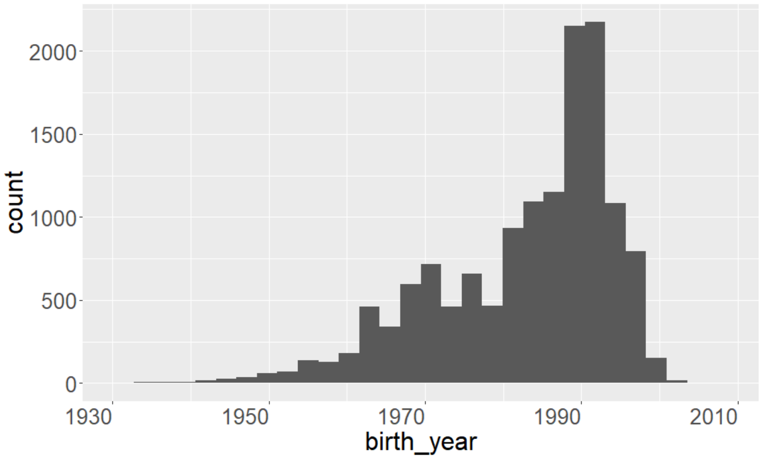
Accordingly, the purchasing behaviours in two groups of users are similar. The missing values tend to be missing in random and maybe ignorable when investigating the correlation between purchasing behaviours and birth years.

2.2 Demographic

The dataset contains records of 15217 users of various ages from 30 provinces. A demographic summary of the uses involved in the study will be necessary.

Here is a plot of the distribution of "birth_year". (figure 1)

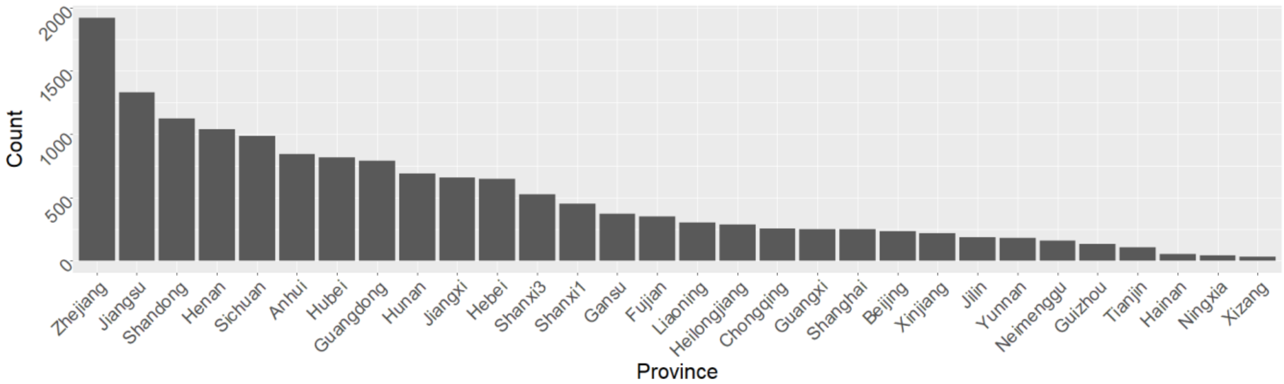
Figure 1 Histogram of the Birth Year



The distribution of the birth year is left-skewed, and most of the users are born from the 1980s to the 2000s. Some users of higher ages also participate in the data collection.

Here is a summary of the population from 30 provinces. (figure 2)

Figure 2 Count Plot of the Population from Provinces

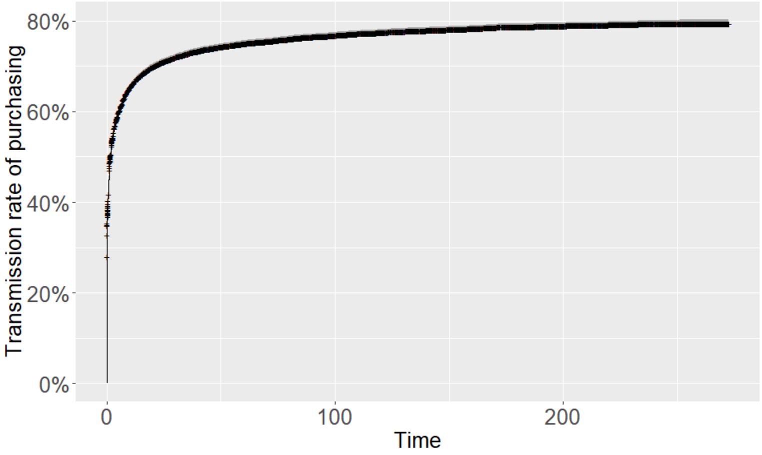


According to the count plot, users from the Yangtze River Delta area occupy a pretty large percentage of the total investigated population. The proportion of users from different regions corresponds to the actual population proportions of provinces in China.

2.3 Variables and Transmission Rate of Purchasing

The main interest in the study is the transmission rate of purchasing, which indicates the efficiency of the product marking to attract new users to revenues. The overall transmission rate of purchasing versus time in days is summarized in figure 3.

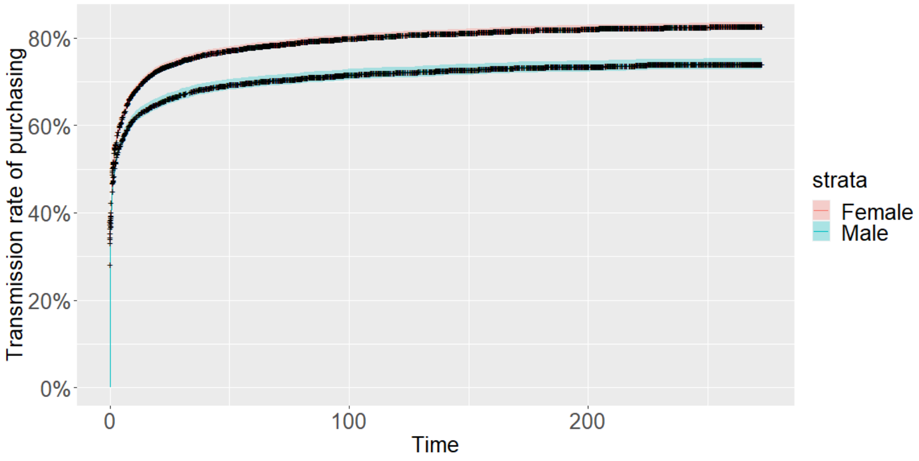
Figure 3 Overall Transmission Rate of Purchasing



Based on the overall transmission rate of purchasing, most of the new users would make the first purchasing in 50 days, and the transmission rate increases much slower after 50 days.

Three users' characteristics are included in the dataset, and the common factor in splitting the population is gender. Gender may affect the transmission rate of purchasing, and such difference can be presented through the Kaplan-Meier plot of the two groups. (figure 4)

Figure 4 Transmission Rate of Purchasing vs. Gender



The transmission rate of purchasing seems to differ between the two genders. The female new users tend to have a higher transmission rate and may make purchasing decision earlier than the male users. A log-rank test can be conducted to check the difference in transmission rate between the two genders, and the results are reported in table 3.

Table 3 Log-Rank Test on Transmission Rate between Genders

Groups	Chi-Squared Statistic	P-Value	Significance
Female/Male	87.5	<2e-16	True

Accordingly, the p-value reported in the log-rank test is smaller than the significant level of 0.05. There is strong evidence that the transmission rate varies between groups of female users and male users.

Considered that the variables “birth_year” and “province” have multiple levels or values, the effects from the two variables can be observed by selecting pairs of levels into comparison. The transmission rates of purchasing versus birth years and provinces are presented in figure 5 and figure.6. The pair of birth years “1970” and “1995” and the pair of province “Guangdong” and “Zhejiang” are selected in the comparisons.

Figure 5 TRP vs. Birth Year “1970” and “1995”

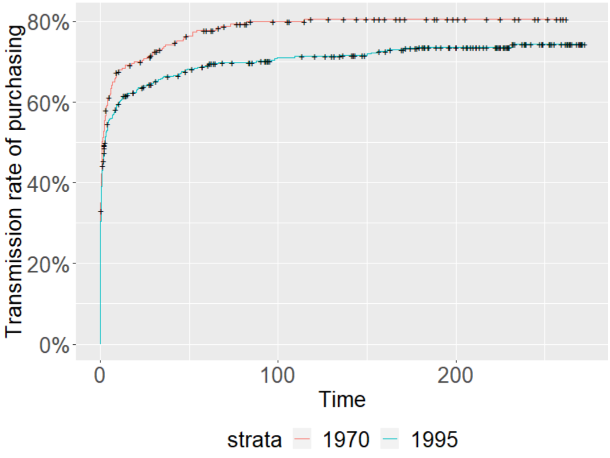
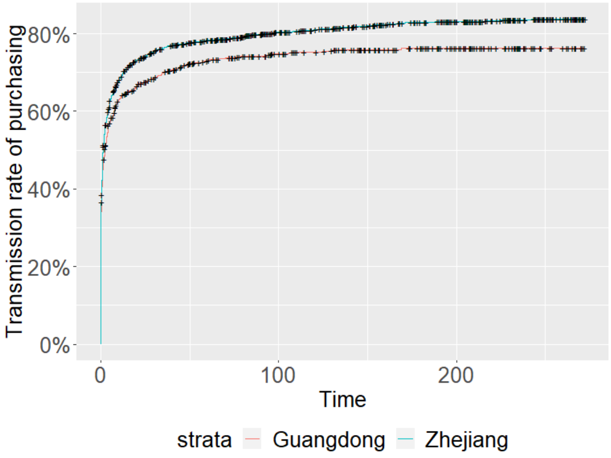


Figure 6 TRP vs. Province “Guangdong” and “Jiangsu”



Based on the plots of the transmission rate of purchasing between pairs of birth years and provinces, it seems that there are considerable effects from the two variables on the transmission rate. The log-rank tests will check the effects in these two specific pairs. (table 4)

Table 4 Log-Rank Tests on pairs of Birth Year and Province

Groups	Chi-Squared Statistic	P-Value	Significance
“1970” / “1995”	5.1	0.02	True
“Guangdong” / “Zhejiang”	10.4	0.001	True

The log-rank tests show a significant difference among the two pairs. Since there are much more levels and values in the two variables, a Cox Proportional-Hazard model should be made to illustrate the effects of these variables.

3. Confirmatory Data Analysis

3.1 Cox PH Model

The Cox Proportional-Hazard model is a classical semi-parametric method that relates the "time to the event" to explanatory variables. Generally, the Cox PH model can be written as follows:

$$h(t) = h_0(t)exp\left(b_1X_1 + b_2X_2 + \cdots + b_pX_p\right)$$

where $h(t)$ is the expected hazard at time t; $h_0(t)$ is the baseline hazard when all the predictors $X_1, X_2, \ldots, X_p$ are zero.

Under the full model assumption, all three variables, "sex", "birth_year", and "province", will be included in the regression. A summary of the full model is recorded in table 5.

Table 5 Cox PH Full Model

Variable	Coef. Estimator	Standard Error	P-Value	Significance
sexMale	-0.1707248	0.0205737	< 2e-16	True
birth_year	-0.0077861	0.0008579	< 2e-16	True
province_xxx	NA	NA	> 0.1	False

According to the Cox PH model, the variables "sex" and "birth_year" have significant effects on the transmission rate of purchasing. The gender "Female" is seen as the baseline of the model, while the gender "Male" has a negative effect on the TRP. The "birth_year" shows a negative correlation to the TRP, which means this product will gain faster purchasing decision among older people. Both results of "sex" and "birth_year" are corresponding to the findings in EDA. However, though some of the "province" may have relatively significant effects than the others, there is weak evidence that the difference of "province" may result in a substantial variety of

TRP. The regression result under the full model gives a brief conclusion on the correlation between variables and the outcome TRP, in which the “province” may be dropped.

A simpler Cox PH model with “sex” and “birth_year” is constructed according to the full model. The result of the simpler model is recorded in table 6.

Table 6 Cox PH Simpler Model

Variable	Coef. Estimator	Standard Error	P-Value	Significance
sexMale	-0.1742161	0.0204674	< 2e-16	True
Birth_year	-0.0080673	0.0008398	< 2e-16	True

After dropping the “province”, the reported estimators for both “sex” and “birth_year” have a slight increase, while the significances, as well as the conclusions, do not change.

Based on the Cox PH model, it implies that female new users with higher age may have a faster transmission rate to make the first purchasing. In contrast, young and male users may not quickly purchase the product after signing up.

3.2 Aalen’s Additive Regression Model

The Cox PH model assumes that the covariates do not change with time, while in the previous analysis, this assumption has not been confirmed. Hence, Aalen’s additive regression model may be utilized to investigate the possible time-dependent variables. Aalen’s additive model assumes that the cumulative hazard function can be expressed as follows:

$$h(t) = a(t) + XB(t)$$

where $a(t)$ is a time-dependent intercept term; X is the vector of covariates which are possibly time-dependent; $B(t)$ is the time-dependent matrix of coefficients.

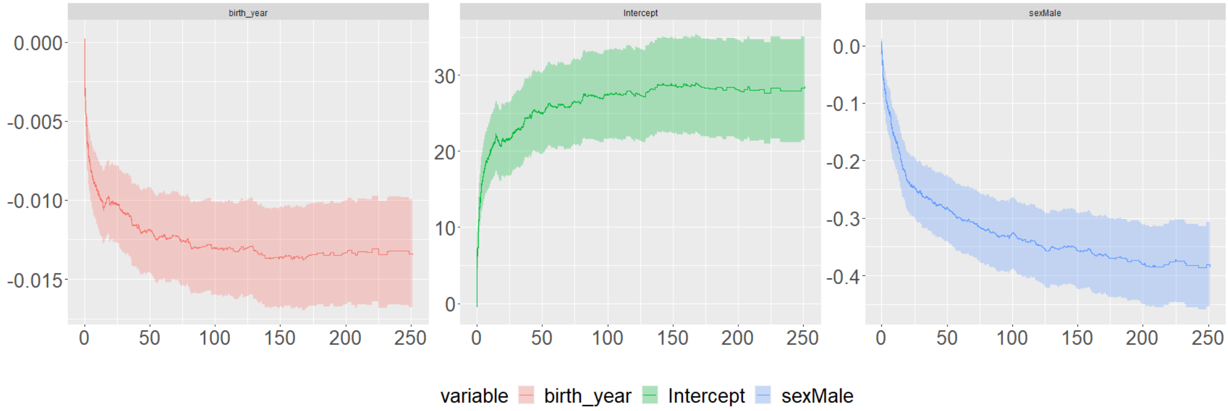
Since the “province” has been checked by the Cox PH model as a non-significant variable, Aalen’s additive will only consider the “sex” and “birth_year” variables. A summary of the regression is in table.7.

Table 7 Aalen's Additive Model

Variable	Slope Estimator	Coef. Estimator	Standard Error	Significance
Intercept	1.3020189	2.057277e-03	2.102398e-04	True
sexMale	-0.0099893	-2.011036e-05	2.329677e-06	True
Birth_year	-0.0006164	-9.748628e-07	1.059690e-07	True

The variety of effects of the covariates over time is shown in the figure.7.

Figure 7 Effects of Variables Change over Time



According to the regression results and the plots, the overall effects from the gender “Male” and the “birth_year” are negative to the transmission rate of purchasing. Moreover, such effects become even more prominent as time goes by. The Cox PH model’s conclusions should still hold, and Aalen’s additive model confirms the conclusion with a deeper explanation.

4. Conclusion

Accordingly, the transmission rate of purchasing is highly correlated with the new user's gender and age. The female new users are willing to buy products in the platform earlier than the male users. People who have higher ages tend to make their first purchase earlier in the platform as well. Hence, older female new users probably have the highest tendency to consume in the platform upon their signing up.

In conclusion, the platform can take different strategies on different population groups. The group of older female new users will be the most stable target consumers for the platform, and this platform should keep their exposure among the group. Though the group of young male new users seem to have the lowest potential of TRP, it implies that this group may have the highest growth potential in the future. In this sense, the platform may need to adjust its marketing strategy or product layout to accommodate the preference of young male users. Another insight of the dataset is that the young consumers occupy the essential part of the new users. It implies that the platform has successfully attracted the new generations and has good momentum in expanding the users' population. Hence, a marketing survey focusing on the young consumers is suggested to be conducted, as the young generation will be the primary users in the following years.

5. Discussion

In this report, the variable “province” is dropped in the Cox PH model. However, in the EDA part, a significant difference between users from “Guangdong” and “Zhejiang” is detected. This means there should be considerable effects from “province”, and a deeper analysis on this variable is suggested.

The Cox PH model implies low significances of the effects from most provinces. Part of the reason is that the “province” variable has too many levels, and a regroup of provinces may solve the problem. For example, the “province” can be reduced to “area” by grouping provinces based on geographical locations or cultures. After such a reduction, the 30 provinces will become a few areas, and the difference of effects from these areas may be more significant than the original “province”. This will also be more practical when designing different marketing strategies for various markets.

Appendix

Code in Github Project:

https://github.com/zhuzp98/Survival-Analysis-on-New-Users-Behaviours