Surviving the teenage years without smoking marijuana

A Decision Support Systems Project

Team 2:
Simon Fogh Thomsen (201906472)
and
Martin Lilleholt Frederiksen (201906635)

AARHUS UNIVERSITY

March 2023

Contents

1 Introduction

Despite general increasing acceptance of the use of marijuana as a recreational drug and as (self-)medication there exists broad scientific consensus that marijuana consumption is damaging to the brain [6] - especially through long-term use and even more when the consumption starts in the early life of people [8].

Motivated by understanding the factors leading to marijuana consumption in the teenage years, this projects takes an survival analysis-approach to the problem and seeks to explain the different in age of first marijuana consumption based on covariates such as income, demographic etc.

2 Methods and Material

This section present the dataset and the methods used to obtain the results in the project.

2.1 Dataset

In this section the dataset and the selected features will be described, where relevant it will also be described how the features are reworked.

2.1.1 National Survey on Drug Use and Health

Every year the NSDUH[1] interviews approximately 70.000 Americans of the age of 12 specifically with questions around alcohol, tobacco and drug consumption, mental health, general health, upbringing, demographics etc.

All data is pseudonymised with per individual-information and with hundreds of unique questions of which some are deselected based on previous answers.

2.1.2 Feature Selection and explanation

The feature selection is purely made based on the projects intuition of features relevant to the age of onset of marijuana consumption, with a focus on features that subdivides the population into most distinct groups. All features are thoroughly described in[10]. Selected features distribution in the filtered dataset can be seen in figure 2.

Ages for substance consumption Three features has been selected within the range of ages for substance consumption, amongst substances that intuitively are consumed at an earlier age than marijuana. The features are here listed below with the question they are based on.

[1] National Survey on Drug Use and Health

1. **ALCTRY:** *"Think about the first time you had a drink of an alcoholic beverage. How old were you the first time you had a drink of an alcoholic beverage? Please do not include any time when you only had a sip or two from a drink"*
2. **CIGTRY:** *"How old were you the first time you smoked part or all of a cigarette?"*
3. **CIGUSE:** *"How old were you when you first started smoking cigarettes every day?"*

Censored covariates Some of the events has not transpired (yet) at the age of first marijuana consumption. To deal with this censoring in the covariates each of the age-features is encoded in a way where they represent the amount of years of the event happening prior to the year of 17, if the event happened prior to the age of first marijuana consumption - if it did not happen it is then represented with a zero. For these experiment and the motivation described it is only interesting what has happened prior to the event, as to describe factors leading to the event as to not unintended describe any reverse causality effects.

Age The age of which the questionnaire was completed is found encoded in the features **AGE2** and decoded in the feature **AGE**. This feature has solely been used as a mean to filter the respondents. The project focuses on respondents at the age of 17 years ($n = 2260$), as the questions of the questionnaire changes for respondents at higher ages.

General demographics The following list of features has been selected to group the respondents.

1. **IRFAMSOC:** Whether the respondents family receives Social Security Payments, as an indicator of low income.
2. **IRFAMSSI:** Whether the respondents family receives Supplemental Security Payments, which is an extra social security benefit payed to those with a combination of low income and disabilities. The size of the groups and their overlap is shown in figure 1.
3. **IRFAMIN3:** Encoded income in family (1: $\leq$10k USD, 2: 10k-20k USD, 3: 20k-30k USD, 4: 30k-40k USD, 5: 40k-50k USD, 6: 50k-75k USD, 7 $\geq$75k USD)
4. **PDEN10:** Core Based Statistical Area classifications (CBSA) as an indicator of the size of the county of the respondent. [3]
5. **YEPGDJOB:** How often the respondent is told by its parents when it has been doing a good job the last year, under the consumption that this behaviors from parents has been constant from the age of 12 to the age of 17 (hence was the same when the respondent did consume marijuana).

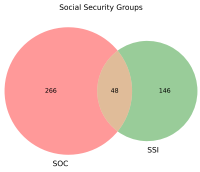
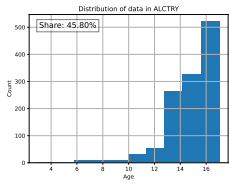
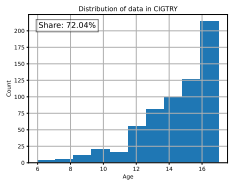


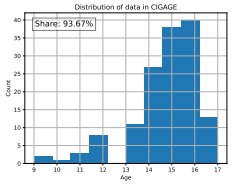
Figure 1: Social Security Groups



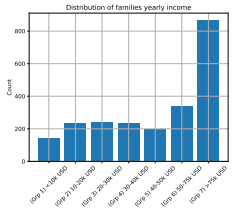
(a) Distribution of feature ALCTRY



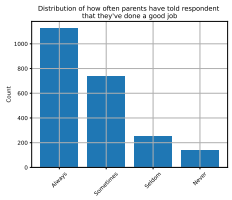
(b) Distribution of feature CIGTRY



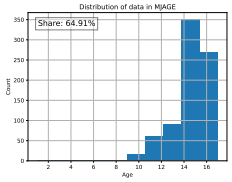
(c) Distribution of feature CIGAGE



(d) Distribution of feature IRFAMIN3



(e) Distribution of feature YEPGDJOB



(f) Distribution of feature MJAGE

Figure 2: Distribution of selected features from the 17 years old respondents ($n = 2260$)

Age of first Marijuana consumption The dependent variable **MJAGE** gives the age of the first consumption of marijuana of the respondent.

Censoring Only a part ($\approx 65\%$) of the respondent has at the time of the questionnaire used marijuana, which makes the dependent variable subject to right-censoring [7]. For this group ($\approx 35\%$) we don't know if they will ever start using marijuana, we only know that it will be later than at the age of 17. Further we deal with a type of censoring, where all the censored observations are censored at the same time (age 17).

2.2 Methods

In this section the method used to conduct the results will be introduced and described. The section will utilize the general nomenclature in the survival analysis domain by referring to not smoking marijuana as surviving.

2.2.1 Kaplan Meier Estimator

The Kaplan Meier Estimator is a non-parametric statistic used to quantify the probability of surviving past a certain time [5]. The decreasing function is defined as

$$S(t) = Pr(T > t) \tag{1}$$

And is given by

$$\hat{S}(d_k) = \prod_{j=0}^{k} \frac{r_j - q_j}{r_j} \tag{2}$$

Such that at any time t between d_k and d_{k+1} the $\hat{S}(d_k)$ is the probability of being alive. The confidence interval used is calculated point-wise and is based on Greenwoods Exponential Formula [13] given by

$$\hat{S}(t) \pm Z_q \sqrt{\hat{V}[\hat{S}(t)]} \tag{3}$$

Where $\hat{V}[\hat{S}(t)]$ represents the variance and is given by

$$\hat{V}[\hat{S}(t)] = [\hat{S}(t)]^2 \sum_j \frac{q_j}{q_j(r_j - q_j)} \tag{4}$$

2.2.2 Log-Rank test

To compare various groups survival time, the experiments uses the log-rank[1] test with the null-hypothesis, H_0 that the two groups have the same survival curve (hence H_A being that they differ). The log-rank test statistic is given by

$$W = \frac{X - E(X)}{\sqrt{Var(X)}} \tag{5}$$

Which corresponds to a chi-squared test value, that can be converted to a p-value, which is used to accept or reject the null-hypothesis. If the p-value is found to be less than the confidence level chosen, the null hypothesis is rejected - hence we cannot confirm that the survival curves are similar.

2.2.3 Cox Proportionality Hazard

Using Cox Regression (Proportionality Hazard Regression) covariates can be examined to determine their respective effect on the dependent variable MJAGE.

Hence the model expresses the risk of starting to smoke marijuana at a certain age, t, given the covariate x_i, such that

$$h(t|x_i) = h_0(t) exp(\sum_{j=1}^{p} x_{ij}\beta_j) \tag{6}$$

Where x_{ij} is the covariate j for individual i and β_j represents the influence of the j'th covariate in the predicted age. Further $h(t)_0$ is the baseline hazard, being the hazard at time t where the influence from all covariates equals zero.

The quantity $HR = exp(b_i)$ is the Hazard Ratio and describes how much a covariate effects the overall hazard, where:

- $HR = 1$: No effect
- $HR < 1$: Reduction of hazard
- $HR > 1$: Increase in hazard

Cox Proportionality Hazard Models Assumption The Cox Model inherently assumes that for all the covariates x_{ij}, a one-unit increase, corresponds to an increase in $h(t|x_i)$ by a factor of β_j.

Covariates can be both categorical and continuous, but they must comply with the two points, following the inherently assumption in the cox model[4]:

1. Survival curves for different strata[2] must have hazard functions that are proportional over the time t
2. The relationship between the log hazard and each covariate is linear

2.2.4 Other parametric models

Further the following parametric models has been used: Wiebull, Exponential, Lognormal, Log-logistic, Piecewise exponential, Generalized Gamma and Splines [2].

2.2.5 AIC

AIC or the Akaike's Information Criterion is used to evaluate the fit to the parametric models and is an error metric (hence lower values corresponds to a better fit) and is defined as:

$$AIC = 2k - 2ln(L) \tag{7}$$

Where k is the amount of degrees of freedom in the model and L is the (partial) likelihood of the model. This metric is chosen to balance the tradeoff between amounts of degrees of freedom and the fit.

2.2.6 Concordance Index

The concordance index will be used to measure the non-parametric models and is defined as the concordant pairs divided by the total number of possible pairs.

[2]Subgroups of population

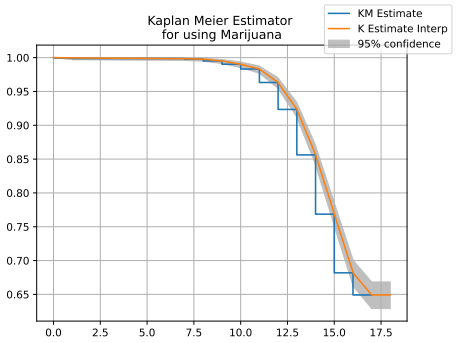


Figure 3: Kaplan Meier Estimator for using Marijuana, with linear interpolation and confidence-intervals

2.2.7 Random Survival Forest

The experiments will use a Random Survival Forest model[9], which is a version of the popular tree-models, suited for survival analysis. Further the RSF-model is of the random tree-models that, compared to the classic tree-models, are less prone to overfitting, due to the randomness in their training.

2.2.8 XGBSE: XGBoost for Survival Analysis

The popular XGBoost is amongst the best state-of-the-art machine learning models a broad range of fields[14]. A model has been found that builds on top of this to improve its performance even further specifically in the area of survival analysis[12], that's called XGBSE[3]

3 Experiments, Results and Discussions

In this section the various experiments will be presented with the related results and relevant parts will be discussed.

3.1 The risk of anyone using marijuana

Initially it is being assessed how the overall (for all stratum and covariates) survival curve smoking marijuana looks up to the age of 17. Since the duration's (ages) are heavily quantize into integer values of the age of onset of substance use, linear interpolation is used to show the survival curve and 95% confidence-intervals. The Kaplan Meier Survival Curve is showed in figure 3.

The right-censoring mentioned in paragraph 2.1.2 is seen as the curves ends at 17 years.

[3]XGBoost Survival Embeddings

3.1.1 Predictions on events happening in the future

The plot is generated on the Kaplan Meier model, as seen in code listing 1. It can be noted that all events happening after the age of onset of marijuana use is being artificially sensored (by setting it to 'NaN'), to avoid the models using events happening after the onset of marijuana.

```
from lifelines import KaplanMeierFitter

kmf = KaplanMeierFitter()
dur = df.apply(lambda x: x['MJAGE'] if not np.isnan(x['MJAGE'])
    else x['AGE'],axis=1)
event_obs = df.apply(lambda x: 1 if not np.isnan(x['MJAGE']) else
    0,axis=1)

kmf.fit(durations=dur, event_observed=event_obs)
```

Listing 1: Python example

3.1.2 Modelling the survival function as parametric model

Experiments on modelling the survival function one of the parametric models mentioned in 2.2.4 has been performed, as seen in figure 4. The AIC has been calculated for all the fits are as follows:

1. **Weibull**: 5850
2. **Exponential**: 7662
3. **LogNormal**: 5927
4. **LogLogistic**: 5835
5. **Piecewise Exponential**: 7666
6. **Generalized Gamma**: 5843
7. **Spline**: 5843

It is noted that best fits are found in the Weibull, Generalized Gamma and Spline-model; which makes sense as the Generalized Gamma function is a generalization of the Weibull-distribution[11]. It can also be observed in the figure 4 that these are the models that model the survival curve ground truth provided by the Kaplan Meier-estimator, the best.

3.1.3 Weibull AFC

Since the Weibull model seemed to have the best fit on the survival curve, a Weibull AFC[4]-model has been fitted to the data, but the underlying baseline model was so dominating that none of the coefficient for the covariates seemed to have any effect, for which reason the experiment is left out of this report.

[4]Accelerated Failure Time

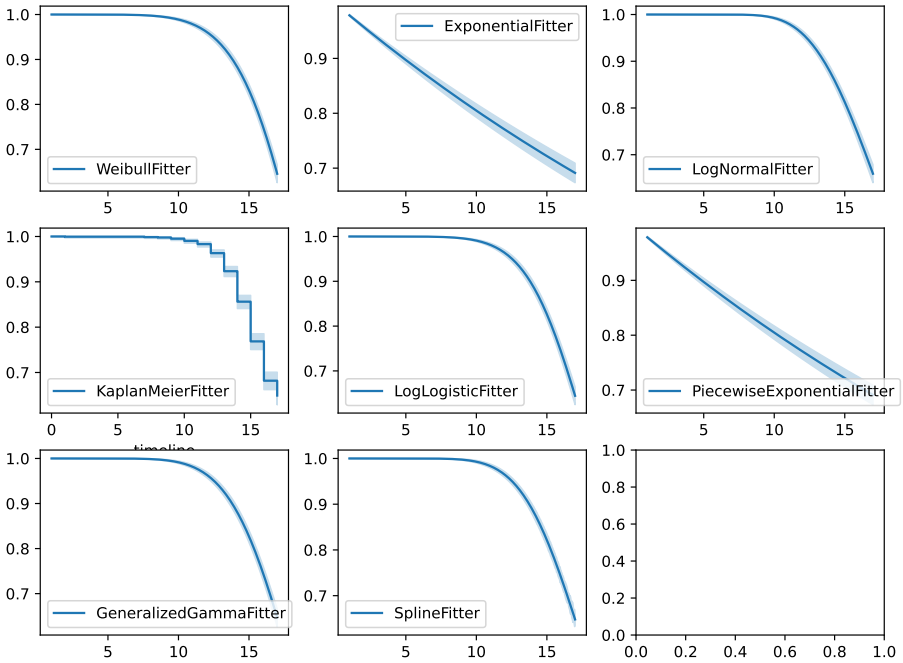


Figure 4: Fits of the survival function to parametric models

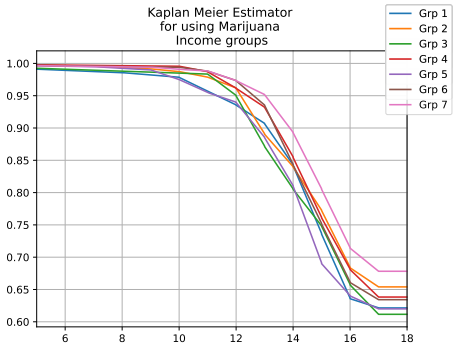


Figure 5: Interpolated Kaplan Meier Estimator for population grouped by income-group (IRFAMIN3)

3.2 The risk of using marijuana based on income groups

To further examine the covariates, the population has been split into stratum based on their income group. For each group a Kaplan Meier Estimator has been fitted. The a plot of the linearly interpolated function can be seen in figure 5. It could seem that there is a slightly significant slower tendency for respondents from the richest income group to start the use of marijuana - and they also have the best statistical chance of not smoking marijuana before the age of 18. The second to best chance of not smoking marijuana is of the second to poorest income group, which points towards rejecting a hypothesis that the more wealthy background, the less prone you are to start smoking marijuana at an early age.

3.2.1 Social Security Payments

Another way the questionnaire quantifies the financial aspect of the respondents background is by categorizing them into the categories of which social security plans they might be, as described in section 2.1.2. In the same manner as with the family income group from the previous section, the Kaplan Meier Estimators survival function is plotted in 6.

3.2.2 Statistical Significant difference between income groups

To assess if there is really an statistical significance between the survival curves for the different income groups, a log-rank test is performed on the Cartesian product of the sets of income groups. As its seen in figure 7, only a very few combinations yields a log-rank test-statistic score below the classical threshold of a $CI = 0.05$; meaning that they cannot be confirmed to be different. However, all of the lowest test-statistics are to be found amongst combinations of income group 7 and other income groups, which could lead to the assumption that this income group stands out from the rest.

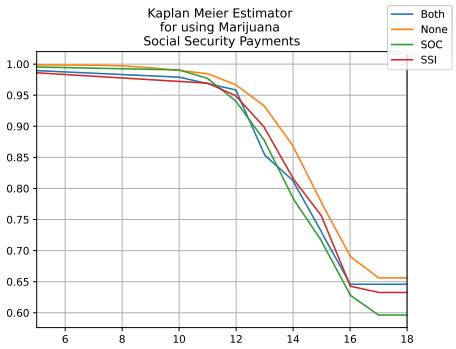


Figure 6: Interpolated Kaplan Meier Estimator for population grouped by social security category (IRFAMSSI/IRFAMSSO)

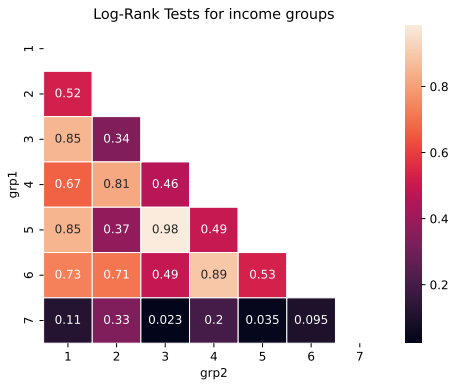


Figure 7: Heatmap of log-rank test-statistic on all combinations of income-groups without duplicates (hence the lower triangle matrix)

3.3 How lower my risks of starting to smoke marijuana early?

To assess what covariates play a role in your hazard of starting to smoke marijuana early, a Cox Proportional Hazard model is being trained.

3.3.1 Feature Design

The covariates describing ages of trying or using other substances is being transformed from being ages, to being ages prior to turning 18 years. In this way the higher the number, the more extreme.

3.3.2 Cox Proportional Hazard model

After the model is trained, it can be seen that the concordance index is $CI = 0.72$. Further the feature importance can be evaluated in the plot seen in figure 8, from which it, amongst other things, can be seen that:

- **CIGTRY_P, ALCTRY_P**: The earlier you try to smoke a cigaret (or try alcohol), the more prone you are to start using marijuana early
- **IRFAMSOC**: If your family receives social security, there is a higher chance you will start to smoke marijuana early
- **CIGUSE_P**: The later you start smoking cigarets regularly, the less prone you are to start using marijuana early
- **PDEN10**: The more dense populated area you live in, the

3.3.3 Analysis of the covariates?

To examine the partial effects on the overall hazard function, these are plotted for each covariate with the following findings:

- IRFAMSSI: The feature didnt show any significant effect in the log(HR) plot (figure 8), hence the partial effect from belonging to the two groups seems to have the same survival curve according to the model as seen in figure 9a.
- CIGTRY_P: The age of trying to smoke a cigarette seems to have a large influence on the survival function - according to the model, you are almost certain to start smoking marijuana before you turn 18, if you have tried to smoke a cigarette prior to you turning 10.
- CIGTRY_P: There seems to be a clear trend: the earlier you try a cigarette, the earlier you are prone to smoke marijuana. However, it can be noticed that there is some of the survival curves that are out of order with that trend (eg. CIGTRY age 16 and 12).

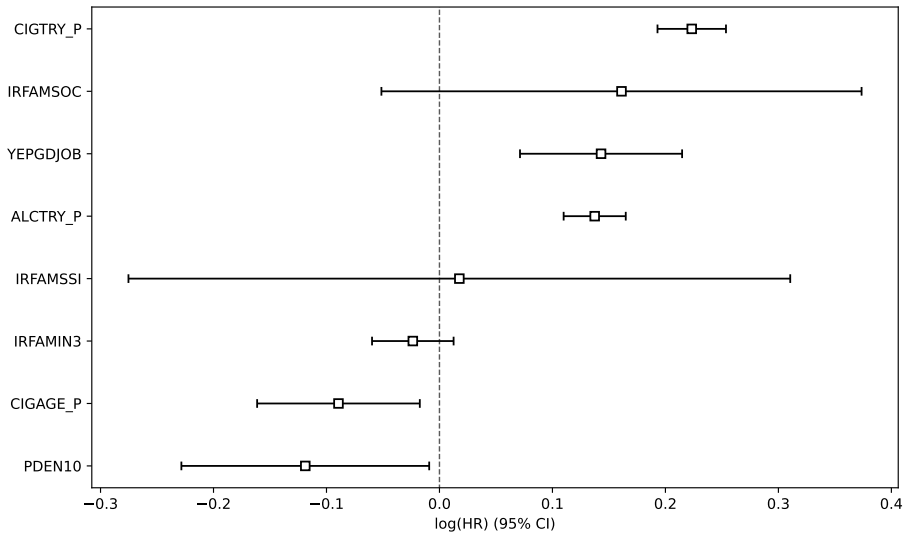
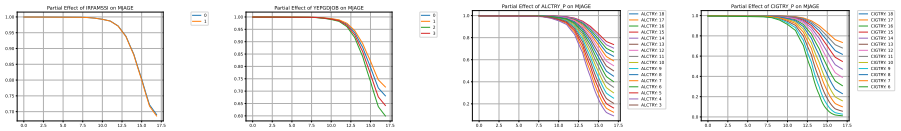
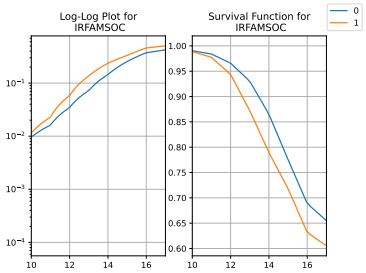


Figure 8: Log(HR) with confidence interval for Cox Proportional Hazard model

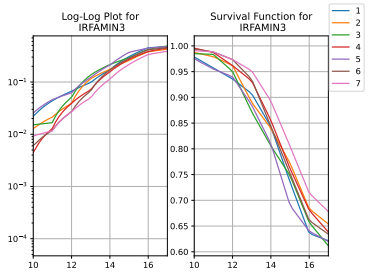


(a) (b) (c) (d)

Figure 9: Partial Effect on the age of using marijuana of selected covariates in the Cox Proportionality Hazard model



(a)



(b)

3.3.4 Cox Proportionality Assumptions

To check whether the assumptions embedded in the CPH-model are being violated, p-value for the proportionality hazard test are calculated and it seen that the assumptions are violated for the covariates:

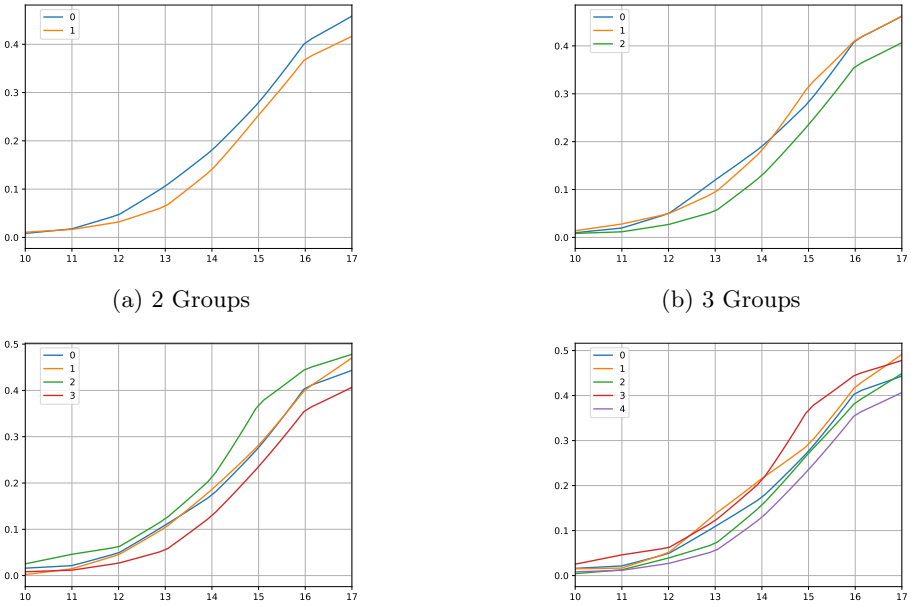
- IRFAMIN3 (p-value: 0.0445)
- ALCTRY_P (p-value: 0.0059)
- CIGTRY_P (p-value: 0.0501)

As an example the IRFAMIN3-feature hazard is showed in a log-log-plot next to the survival function in figure 10b, from which it is evident that the hazard is not proportional throughout the timeline evaluated (10-18 years) - as it is the case for the feature IRFAMSOC as clearly seen in figure 10a.

That the hazards are proportional means for e.g. the IRFAMSOC, that being in a family that receives social security increases the risk of using marijuana equally throughout the depicted timeframe, as seen in figure 10a.

What to do? To mitigate that the features violates the assumption, one approach is to bin the values into groups. As seen in figure 11, this has been done for 2-5 equally sized bins. The p-value for the amount of bins can be seen in figure 12a. As a good example of the relationship between the log-log plot and the p-value of the proportionality hazard test, the figure 11a and 12a can be noted. However, in this experiment these possible improvements will not be explored further.

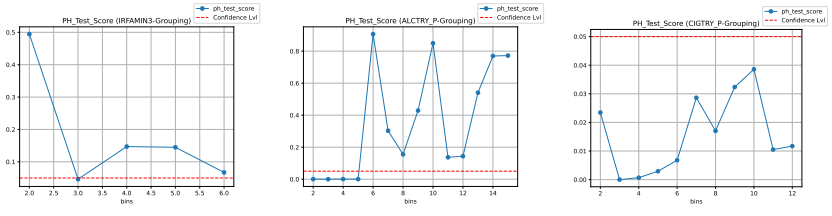
Is the violation a problem? It is not necessarily a problem that the covariates violates the assumptions of the model, as to the extend that the aim of the model is exact predictions - but the violations limits the interpretability and how the model can be communicated/understood; especially by people without insight into survival analysis.



(a) 2 Groups (b) 3 Groups

(c) 4 Groups (d) 5 Groups

Figure 11: IRFAMIN3 binned into fewer groups



(a) (b) (c)

Figure 12: P-Value from Proportionality Hazard Test Statistic related to how many bins the covariates that violates the CHP-model's assumptions are split into

Performance as a classifier The model has also been trained on a 75%/25% train-test-split and evaluated as a classifier (where the classes are based on wether an individual smokes marijuana prior to their 17th year birthday, with the following results:

- Accuracy: 0.73
- Precision: 0.73
- Recall: 0.95
- F1: 0.83

3.4 Machine Learning Models

Lastly two machine learning models have been trained on the data and evaluated by the concordance index. The models are both used with their default hyper parameters. The models and their performance are:

1. Random Survival Forest: 0.71
2. XGBSE: 0.68

4 Conclusion and Perspective

This project has chosen a complex dataset with a right-censored dependent variable, the age of which young people starts smoking marijuana, as well as a range of features that the project found interesting, that has been examined and transformed to the occation. Kaplan Meier Survival Curves has been showed for different financial backgrounds as well as the entire populator, which survival function has been examined under goodness-of-fit metrics to understand the underlaying distribution.

By log-rank test on the Cardinal product of income-groups it is shown that the highest income group has a distinct survival function to the remaining 6, and through the Kaplan Meier Estimator it's showed that this survival function has a longer survival time trough the teenage years.

The features are examined while a Cox Proportionality Hazard model has been fit. In particular the earlier the respondents has *tried* alcohol and cigarettes the more they are prone to an early event of using marijuana; in opposition to the age of which the respondent used cigarettes daily, which delays the event of marijuana use. Finally the assumptions behind the model has been examined and suggestions are given and showed to how these violations can be fixed.

Furthermore, two more complex machine learning models (Random Survival Forrest, XGBSE) has been fitted to the data.

The best CI-score obtained is by the Cox Proportionality Hazard model ($CI = 0.72$) and Random Survival Forest ($CI = 0.71$)

References

[1] J Martin Bland and Douglas G Altman. "The logrank test". In: *BMJ* 328.7447 (May 1, 2004), p. 1073. ISSN: 0959-8138, 1468-5833. DOI: 10.1136/bmj.328.7447.1073. URL: https://www.bmj.com/lookup/doi/10.1136/bmj.328.7447.1073 (visited on 03/22/2023).

[2] Cam Davidson-Pilon. *Survival regression*. URL: https://lifelines.readthedocs.io%20/en/latest/Survival%20%20Regression.html#accelerated-failure-time-models.

[3] PLANEY DONALD. *WHAT ARE CORE BASED STATISTICAL AREAS AND WHY DO WE USE THEM?* Oct. 2020. URL: https://carolinatracker.unc.edu/stories/2020/10/28/cbsa_geography/ (visited on 03/20/2023).

[4] Dana Hashim and Elisabete Weiderpass. "Cancer Survival and Survivorship". In: *Reference Module in Biomedical Sciences*. Elsevier, 2017. ISBN: 9780128012383. DOI: 10.1016/B978-0-12-801238-3.65102-4. URL: https://linkinghub.elsevier.com/%20retrieve/pii/B9780128012383651024 (visited on 03/22/2023).

[5] Gareth James et al., eds. *An introduction to statistical learning: with applications in R*. Springer texts in statistics 103. OCLC: ocn828488009. New York: Springer, 2013. 426 pp. ISBN: 9781461471370.

[6] Jayalakshmi Krishnan. "Marijuana and Its Effects in Brain". In: *Asian Journal of Research and Reports in Neurology* (Sept. 8, 2022), pp. 64–69. URL: https://journalajorrin.com/index.php/AJORRIN/article/view/66 (visited on 03/20/2023).

[7] Kwan-Moon Leung, Robert M. Elashoff, and Abdelmonem A. Afifi. "CENSORING ISSUES IN SURVIVAL ANALYSIS". In: *Annual Review of Public Health* 18.1 (May 1997), pp. 83–104. ISSN: 0163-7525, 1545-2093. DOI: 10.1146/annurev.publhealth.18.1.83. URL: https://www.annualreviews.org/doi/10.1146/%20annurev.publhealth.18.1.83 (visited on 03/20/2023).

[8] Krista M. Lisdahl et al. "Dare to Delay? The Impacts of Adolescent Alcohol and Marijuana Use Onset on Cognition, Brain Structure, and Function". In: *Frontiers in Psychiatry* 4 (2013). ISSN: 1664-0640. DOI: 10.3389/fpsyt.2013.00053. URL: http://journal.frontiersin.org/article/10.3389/fpsyt.2013.00053/abstract (visited on 03/20/2023).

[9] Asif Newaz, Farhan Shahriyar Haq, and Nadim Ahmed. "A Case Study on Risk Prediction in Heart Failure Patients using Random Survival Forest". In: *2021 5th International Conference on Electrical Engineering and Information Communication Technology (ICEEICT)*. 2021 5th International Conference on Electrical Engineering and Information Communication Technology (ICEEICT). Dhaka, Bangladesh: IEEE, Nov. 18, 2021, pp. 1–6. ISBN: 9781665495226. DOI: 10.1109/ICEEICT53905.2021.9667933.

URL: https://ieeexplore.ieee.org/document/9667933/ (visited on 03/31/2023).

[10] NSDUH. *2015 NATIONAL SURVEY ON DRUG USE AND HEALTH, PUBLIC USE FILE CODEBOOK*. 2015th ed. NSDUH. URL: https://www.datafiles.samhsa.gov/sites/default/files/field-uploads-protected/studies/NSDUH-2015/NSDUH-2015-datasets/NSDUH-2015-DS0001/NSDUH-2015-DS0001-info/NSDUH-2015-DS0001-info-codebook.pdf (visited on 03/20/2023).

[11] *The Generalized Gamma Distribution*. The Generalized Gamma Distribution. URL: https://reliawiki.org/index.php/The_Generalized_Gamma_Distribution.

[12] . *XGBoost Survival Embeddings*. XGBoost Survival Embeddings. URL: http://github.com/loft-br/xgboost-survival-embeddings.

[13] Xiaobin Yuan and Shesh N. Rai. "Confidence Intervals for Survival Probabilities: A Comparison Study". In: *Communications in Statistics - Simulation and Computation* 40.7 (July 7, 2011), pp. 978–991. ISSN: 0361-0918, 1532-4141. DOI: 10.1080/03610918.2011.560732. URL: http://www.tandfonline.com/doi/abs/10.1080/03610918.2011.560732 (visited on 03/20/2023).

[14] Wengang Zhang et al. "State-of-the-art review of soft computing applications in underground excavations". In: *Geoscience Frontiers* 11.4 (July 2020), pp. 1095–1106. ISSN: 16749871. DOI: 10.1016/j.gsf.2019.12.003. URL: https://linkinghub.elsevier.com/retrieve/pii/S1674987119302361 (visited on 03/31/2023).