Erasmus University Rotterdam
Erasmus School of Economics
Seminar Case Studies in Quantitative Marketing

Modelling Car Insurance Churn at AEGON: A Comparison of Approaches, from Parametric to Non-Parametric

Hanneke Tiktak (455549)
XingBao Fu (534823)
Sem Koelewijn (494615)
Sjoerd Bommer (621763)

Contents

1 Introduction

Insurance companies are involved in managing risks for individuals and businesses by offering protection against potential losses. They collect premiums to create a pool of funds capable of compensating those who experience insurance related unforeseen events, thereby consolidating risk. However, pricing insurance products poses a challenge, requiring a cautious balance between competitiveness and financial sustainability.

Furthermore, retaining customers over extended periods is advantageous for insurance companies. These long-term customers offer a stable revenue stream and typically yield higher lifetime value for the company [1]. Conversely, customers who frequently switch insurers continually seek the best available deals in the market. Therefore, it is crucial for insurance companies to develop pricing models that not only consider customers' price sensitivity but also anticipate when they may leave over time.

In this case study, we will therefore investigate the likelihood of customers cancelling their insurance policy. In marketing, this can be described as customer churn. We will investigate how the characteristics of customers, the insured object, and the insurance itself may influence the likelihood of customer churn. We will mainly investigate the effect of two important pricing mechanisms; applying welcome discounts in the first year, and (permanent) list price adjustments. In doing so, we will be considering car insurances specifically, though many of the questions could be generalised to other non-life insurance products. Concisely, the research question is as follows: does churning behaviour of customers from year to year change if a welcome discount or list price adjustment is applied, and if so, how?

To answer this question properly, we wish to obtain causal estimates of these effects, which will require a careful assessment of the assignment process of the pricing mechanism. In addition, given that car insurance is often mandatory, there are different kinds of customers that would want insurance, and they may display different churning behaviour as well. It might not be sensible to develop a model for all insurance policyholders at once. Various techniques exist within quantitative marketing research and actuarial sciences to model customer churn or death [2]. Although the latter are used in the context of life insurance, they could be used to model the departure of policyholders as well. In this case study, we therefore try to bridge the gap between these different issues and areas of research and use several (combinations of) models to answer our research question and compare their results. We therefore investigate the following sub-questions as well:

- Can different segments of customers be found, and is their churning behaviour different?
- What is the effect of welcome discounts and list price adjustments on churning behaviour, and can this be determined causally?
- From a parametric to a non-parametric approach, how can predictions for churn be done in the future?

In short, we will apply clustering techniques to separate the customer base into a number of segments. Next, we will focus on finding causal relations by applying a hierarchical Bayesian model. Finally, we will use two different models that incorporate the concepts of a survival analysis, as is often featured in both actuarial sciences, using concepts that are also prevalent in quantitative marketing areas. One method will rely on more parametric assumptions and

allows for more interpretation of the results, while another will make use of machine learning techniques. We compare these methods and their results to determine the impact of model selection on the answer to our research question.

In what follows, we first outline the provided data in detail. After this, we present the various methods used to answer the research question, including a discussion of relevant literature where these techniques were developed and/or used. Once the models have been explicated, we present and discuss the results of our analyses. In the final section, we discuss possible routes for further investigation and address any key limitations related to the dataset and the chosen methods.

2 Data

This section aims to provide a thorough overview of the car insurance churn data set, including its variables and observations. Besides, some variable strategies and data pre-processing methods are introduced.

2.1 Data description

The dataset provided by AEGON provides customer churn for car insurance, represented in the form of unbalanced panel data. It contains 162,454 unique customers who hold a policy from 2019 to 2024, with a total number of observations of 437,514. Each observation documents whether a customer has churned in a time interval, alongside relevant information about the insurance and customer. Most intervals represent a year. However, contracts started before 2019 are summarized in a single observation. Once customers churn, they are not further recorded in the data set, and their identifiers change if they return later.

The dataset shows a right-censoring pattern, which is defined as the pattern in churn that has not happened during the data collection period but is expected to happen in the future. Figure 1 illustrates an example of ten customers, where red lines with dots denote right-censored customers who have not churned yet by the end of the collection period. In the whole data set, there are 95,094 customers who are censored while 67,360 churned.

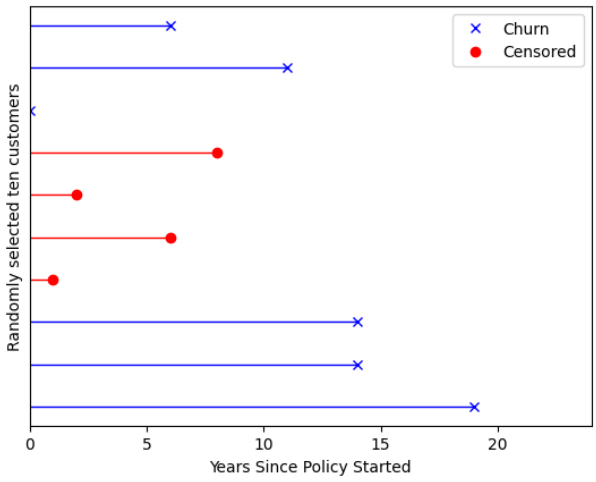


Figure 1: Visualization of customer churn and censoring status over time. Red lines with dots indicate right-censored customers and blue lines with crosses are churning customers

The dataset consists of 57 variables, which can be summarised into five groups: customer/policy identifier, time indicators, churn indicators, insurance details, and customer characteristics, as detailed in Table 3 in the Appendix, which also includes the data types and properties of the time dependencies. Churn indicators represent whether a customer churns in a certain year. Three premium-related variables are calculated based on the customers' characteristics and their choice of insurance.

Moreover, customers are pseudo-randomly segmented into four subgroups based on their postal code to determine whether they are eligible for a welcome discount and list price adjustment. The pivotal variable of our research question, welcome discount, indicates the magnitude of the welcome discount, ranging from 0 (no discount) to 0.3 (30% discount). Eligible customers who initiated their contracts in 2021 or later may receive this discount during the first year of their contract.

2.2 Data pre-processing

Different subsets of the dataset are selected to address different research questions. To answer the questions about customers' heterogeneity and the influence of welcome discounts on their churn probability, customers who started policies in 2019 and onwards were selected to avoid underestimating long-term churning. Additionally, we selected customers from 2021 onwards to assess the causal effect of welcome discounts, which were introduced in 2021.

This dataset contains variables that are a linear combination of other variables, such as the number of coverage, which is the summation of the number of the main coverage and the number of supplementary coverage. For these variables, we drop one of the variables to get rid of the multicollinearity issue, especially for the discrete-time survival analysis. Besides, several variables include missing values. Table 4 in the Appendix displays the variables with missing or inaccurate values, influencing our variable selection process. For example, the premium of main coverage and the premium of supplementary coverage include many missing values, which suggests we drop them and keep the variable total premium, which is the summation of the two with no missing values.

However, we have been informed that some values in total premium are wrongly assigned to zero, which exclusively coincides with the year of customer churn, suggesting an intentional recording practice. However, this can introduce reverse causality biases in our case, as a regression may wrongly attribute zero premium as the cause for churn rather than its consequence. Thus, imputing missing values plays a significant role in our analysis. Many methods can be used for this task, such as column-wise mean/median imputation, distance-based imputation methods, and model-based methods [3]. In this case, because the premium depends on the customer's characteristics and insurance package instead of a random number, and the relationship between premium and characteristics is not always linear, such as age and premium, we will use the class mean to fill the customers who churn within a year and use the "Last observation carried forward" method to fill the others, following the suggestions by [4]. Classes are determined by the cluster method, which will be introduced in the methodology section.

3 Literature Review

The literature review section provides an overview of existing research related to the topic of this paper, which includes clustering, causality, and modelling customer lifetimes in the context of predicting customer churn. We examine clustering techniques applied in previous studies, such as K-Prototypes and DBSCAN, along with the evaluation methods employed to measure the effectiveness of these clustering approaches. Additionally, we explore the complexities of causal inference, discussing frameworks like the potential outcome framework and Bayesian regression techniques. Furthermore, we investigate different modelling strategies, including survival models and machine learning algorithms like Random Survival Forest, for handling censored time-to-event data in churn prediction scenarios. This literature review will inform our subsequent analyses.

3.1 Clustering

Multiple papers have proposed integrating clustering in customer churn prediction-related research [5, 6, 7]. James et al. proposed in their research related to Customer Churn Prediction in Automobile Insurance to use K-prototypes for identifying customer segments [5], Bose et al. evaluated the effect of implementing Hierarchical Clustering on customer churn prediction [6], and Matuszelański at al. applied Density-Based Spatial Clustering of Applications with Noise (DBSCAN) in the context of customer churn in retail E-commerce business [7].

An essential aspect of clustering is determining the optimal number of clusters. Several papers on customer churn have employed clustering methods with cluster numbers ranging from 2 to 15 [8, 9, 10]. In one study, Li et al. utilized 7 clusters as the maximum number of considered clusters [11]. In this paper, we considered a cluster range of 2 to 7 clusters since using clustering methods with more than 7 clusters did not result in better clusterings.

Lastly, clustering methods must be evaluated and compared. Various cluster evaluation methods are available for selecting the optimal number of clusters. Wu et al. proposed, in their research related to customer churn, using the Silhouette Score for evaluating clustering methods [10]. In a similar research, Abdi et al. proposed to use the Davies-Bouldin Index for cluster evaluation [8] and Łapczyński et al. used both the Davies-Bouldin Index and the Calinski-Harabasz Index for cluster evaluation [9]. Therefore, these 3 methods are applied in this report for selecting the optimal number of clusters.

3.2 Causality

To determine the causal effect of the welcome discount and list price adjustment on churning behaviour, we can use the fact that the eligibility for either of these is determined quasi-randomly. As such, we can consider the groups that are eligible to receive a discount and those that are not to be similar to a treatment and control group, and attempt to make causal inferences by inspecting the differences in these groups. Similarly, we can do this for the group that is eligible for the list price adjustment and those that are eligible for both the discount and the adjustment. However, a policyholder being eligible to receive the treatment does not imply that the policyholder gets a discount or adjustment. In fact, when inspecting the data, most

customers that are eligible to receive a discount or adjustment do not receive one. From a financial perspective it is sensible that the insurer does not give a discount to anyone randomly, as they might already be the cheapest on the market, for instance. Unfortunately, this nuance complicates causal inference.

In experimental designs, the potential outcome framework is often used to aid in interpretations of causal effects of a treatment in the face of non-compliance. Popularised by Angrist, Imbens, and Rubin amongst others [12], it distinguishes the effect of the 'treatment on the treated' (TOT) from the more general effect of the 'intention to treat' (ITT) that we would find if we were to simply compare the treatment and control group. Importantly, this TOT effect concerns the effect of the treatment on the people that actually comply with their assigned treatment, those that would always or never take the treatment are not taken into account here. To estimate this effect, an instrumental variable (IV) approach is frequently used, as discussed in [13]. Bayesian approaches to instrumental variable regression were also developed (for a discussion, see [14]), and more specifically, under the potential outcome framework by Imbens and Rubin [15].

There are two challenges in applying these techniques to our problem. Firstly, our outcome variable is binary as opposed to continuous. Clarke and Windmeijer discuss a number of modifications to the IV approach to resolve the arisen estimation issues [16], and show by means of a simulation study that these different methods can yield different results. Two common link functions that are used to perform binary regressions are the logit and probit link functions. Nichols compares the implications of the choice for either of these link functions for causal inference [17]. Secondly, it holds that the level of the discount, and similarly the list price adjustment, may vary and is based on a number of customer characteristics included in our dataset. Since a customer can also be given a discount of zero, even if they are eligible, receiving treatment depends on these covariates. Given that these variables are likely to influence churn probabilities more generally, there is a crucial endogeneity issue here. One key assumption in IV-estimation is the exclusion restriction, which means that the variables used to determine if the treatment is actually applied only affect the outcome variable of churn through this treatment variable [18]. This restriction cannot be satisfied, and as such, commonly used IV approaches may be unfit for our case study.

Another approach to causal modelling is to use Bayesian regression techniques, as outlined by McElreath [19]. Bayesian techniques are widely applicable in the area of churn modelling and marketing in general. Notable examples include modelling engagement in TV-watching and video games [20, 21], customer lifetime value [22], as well as customer and employee churn [23, 24]. Bayesian regression analyses allow for flexibility in the modelling procedure, allowing for the incorporation of (layers of) prior knowledge in the form of probability distributions, which, in turn, also makes the modelling more transparent [19, 25]. AEGON can use their own current knowledge on churning behaviour in the past to tailor these models. The fact that our target variable is binary is less of an issue in Bayesian analysis; although the choice of link function is not trivial, they are theoretically equivalent in the framework of (hierarchical) Bayesian modelling [26].

Since the obtained parameter estimates in a Bayesian regression are probability distributions,

we can obtain more nuanced information regarding the certainty of the forecasts, instead of just obtaining a point estimate and associated standard error. Given these advantages, we expect a model that can predict churn like this to obtain useful results. The main challenge in performing a Bayesian causal analysis is setting up an adequate causal model, and deriving an appropriate regression equation [19]. Only if this is done correctly, the coefficients may be interpreted causally. As such, this part of the modelling procedure requires special attention.

3.3 Modelling customer lifetimes

The dataset captures the pattern "time-to-event", which records the customers' information before their churn and whether the customer churns at each given time period [27]. The dataset is right censored due to the limited data collection period, which is particularly notable for the subgroup that received the welcome discount. In other words, numerous observations have not yet experienced churn within this temporal constraint, which is expected in the long term. To estimate the probability of customers churning at a certain year and compare the difference between groups of customers in terms of the probability of churning, the right censored issue should be treated properly, as it triggers substantial bias and leads to low-accurate predictions [28]. Therefore, it is crucial to employ a modeling framework capable of handling censored data and accurately estimating churn probabilities.

Survival models are widely applied to handle the right-censored time-to-event data effectively [29]. Many variants of the survival models have been developed and applied to cope with right censored data, from single sample parametric models to machine learning algorithms [27]. The Cox proportional hazards model is one of the most widely applied survival models in client management and churn prediction [30, 31], as the semi-parametric model is flexible and able to handle multiple covariates[30, 32, 33]. Instead of assuming distributions, like conventional parametric models, it relies on the hazard rate, providing a less restrictive framework. Besides, it accommodates customer heterogeneity by incorporating covariate effects, enhancing model flexibility, and allowing predictions to adapt to individual customers [27]. However, the original Cox model was primarily used in the context of continuous time. As the data is recorded yearly, discrete-time survival analysis is required [34], which has been used in predicting monthly/yearly recorded data [35, 36].

Machine learning algorithms have gained increasing popularity in predicting censored data, as illustrated by recent research [27]. Among these algorithms, Random Survival Forest (RSF) stands out as one of the most effective methods for predicting churn rates [30]. For instance, in the medical field, RSF has been used to predict the 30-day mortality risk in elderly patients with sepsis, highlighting its versatility [37]. The study focused on predicting the survival rate of patients and ranking the risk factors that affected the patients. Additionally, RSF has been applied in economics research, particularly in the field of credit risk default prediction for small, medium, and large enterprises [38]. Here, the emphasis was on predicting the time until bankruptcy using financial explanatory variables and comparing the predictive performance with traditional models such as the logit model.

RSF is also applied in studies concentrating on the time until customer churn. For instance, a study predicts churn among telecommunication industry customers using an RSF model [39].

This model's performance is then compared with that of the Cox proportional hazard model to assess prediction accuracy. Across all these research studies, RSF consistently demonstrates relatively high forecasting performance compared to other models in survival analysis. Therefore, RSF emerges as a relevant and effective method when compared to other survival models.

4 Methodology

Transitioning from the literature review, the methodology section combines Bayesian and machine learning approaches to predict customer churn. Initially, we incorporate clustering techniques to identify customer segments and explore how treatment, in the form of discount, affect customer churn. Subsequently, we employ Bayesian inference with weakly informative and stronger priors to estimate treatment coefficients and explore customer heterogeneity through clustering. Next, we implement discrete-time survival analysis and random survival forest algorithms to model time-to-event data. Finally, we develop a method to compare the outcomes of the different models. This broad methodology section incorporates various model settings, estimation procedures, and evaluation metrics. By approaching the problem from these different angles, we hope to establish which model serves which goal best given the data that has been provided.

4.1 Clustering

The motivation behind employing clustering techniques is to compose meaningful subgroups or clusters within the data that share similar characteristics or behaviours. By partitioning the data into distinct clusters, we can possibly obtain deeper understanding of the inherent variability present in the dataset.

The primary application of clustering in this report is inference-related. Customers within the same segment are considered similar to a certain extent, allowing us to extrapolate the effects of a welcome discount. This enables the distinction between similar customers who received treatment and those who did not, allowing us to extrapolate the impact of welcome discounts on customer churn and gain valuable insights into metrics such as customer lifetime value.

For this paper, we investigated K-prototypes and DBSCAN. Due to its computation intensity and storage space demanding nature, Agglomerative Hierarchical Clustering was not pursued further. All clusterings consider data ranging from 2019 to 2023. Variables included in the clustering methods can be found in Table 6 in the Appendix. Ultimately, the clustering method that results in the best performance is selected for further analysis.

Furthermore, within the clustering framework of this paper, some assumptions are made. Firstly, unique customers are only part of a certain cluster for their complete customer lifetime. This is done by only considering observations with "years_since_policy_started" equal to zero for the clustering. Therefore, for example, a costumer can not be part of cluster 0 in their first year of their lifetime and be part of cluster 1 in their second year of their lifetime. The causal analyses are preformed separately on the 2 clusters, emphasizing the need to assign all observations of a certain customer to either one of the two clusters. Secondly, variables directly related to

discounts are not used for clustering. This choice is made to ensure that clustering can effectively compare customers with similar characteristics, regardless of whether they have received discounts. By implementing this approach, we aim to enhance the comparability between the treatment and non-treatment groups. Lastly, prior to each clustering procedure, all numerical variables are normalized.

4.1.1 K-prototypes

The initial clustering method examined in this report is K-prototypes. K-Prototypes clustering is a hybrid clustering algorithm that combines K-Means for numerical data and K-Modes for categorical data. K-means clustering is a widely used machine learning technique that divides a dataset into 'k' groups based on variable similarities. The algorithm iteratively assigns data points to clusters, such that within-clusters these clusters the sum of squared distances are minimized until convergence [40]. K-Prototypes extends the K-Means algorithm to handle mixed-type data, effectively clustering observations based on both numerical and categorical variables. K-Prototypes assigns centroids for numerical variables and modal values for categorical variables [41]. This method is particularly useful for car insurance data, where a mix of numerical features (e.g. car value) and categorical variables (e.g. car brands) are present, providing a clustering solution that considers both data types. The precise structure of the K-prototypes algorithm can be found in algorithm 2 in the Appendix.

To apply the K-prototypes clustering, firstly the data is separated in numerical and categorical variables and then labelled such that they can be inputted in the algorithm. In our application of K-Prototype clustering, we explored cluster sizes ranging from 2 to 7 clusters.

4.1.2 DBSCAN

The second clustering method considered in this report is Density-Based Spatial Clustering of Applications with Noise (DBSCAN), a machine learning clustering algorithm that groups data points based on their density distribution in the variable space. Unlike K-means, DBSCAN does not require specifying the number of clusters beforehand. It identifies dense regions as clusters and marks less dense areas as noise. DBSCAN considers data points within a specified radius and creates clusters by connecting neighbouring points with sufficient density [42]. In the case of car insurance data, DBSCAN can reveal dense groups of customers without predetermined cluster numbers. The precise structure of the DBSCAN algorithm can be found in Algorithm 1 in the Appendix.

DBSCAN does not require a predefined number of clusters, making the cluster size unknown until the algorithm completes. However, adjustment of the hyperparameters `epsilon` and `min_samples` allows for steering DBSCAN towards generating a desired cluster range. Where epsilon is described as: the maximum distance between two samples for them to be considered neighbours. And `min_samples` is described as: the number of samples (or total weight) in a neighbourhood for a point to be considered as a core point [43]. In our specific case, setting `epsilon` to 3 and 2.5 and `min_samples` to 5000 allowed us to achieve a cluster range of 3 to 4 respectively.

4.1.3 Clustering evaluation

Multiple cluster evaluation metrics are available, with three commonly employed validation metrics being Internal, External and Relative validation. Internal cluster validation assesses clustering quality using intrinsic properties of clusters, relative validation compares different algorithms or settings to identify the most favourable algorithm and external validation compares results with external criteria, such as known labels. However, since we lack known labels in this context, external validation is not considered. This report combines internal and relative validation in the form of 3 cluster evaluation metrics for evaluation namely, the Silhouette Score, Davies-Bouldin Index and Calinski-Harabasz Index. These methods are applied to eventually determine the optimal number of clusters. In addition to these three criteria, cluster size must also be considered. Cluster size refers to the number of observations within a cluster. The clustering aspect of this report aims to compare the results of survival analysis and causal analysis within each cluster, therefore leveraging similarity that is present within these clusters. If the cluster sizes are too small, models with numerous coefficients can not be estimated. Apart from estimation issues, too little observations can also lead to poor model performance.

One of the three considered validation metrics is the Silhouette Score. The Silhouette Score is a clustering validation metric that quantifies the quality of clusters. For each data point, it measures the cohesion within its assigned cluster and the separation from neighbouring clusters. The score, ranging from -1 to 1, is computed by taking the difference between the average distance to points in the same cluster and the average distance to points in the nearest neighbouring cluster, divided by the maximum of these two distances. A higher Silhouette Score indicates well-defined and appropriately separated clusters. A score close to 0 suggests overlapping clusters, and a negative scores indicate that data points might have been assigned to the wrong cluster [44]. The exact formulation of the Silhouette Score can be found in mathematical details section of the Appendix.

The second validation method under consideration is the Davies-Bouldin Index. The Davies-Bouldin Index evaluates the quality of clustering by measuring the average similarity between each cluster and its most similar cluster, relative to the cluster's internal similarity. It quantifies the compactness and separation of clusters, with lower scores indicating better-defined clusters. The index considers both variation within clusters and separation between clusters, making it robust for assessing clustering performance. However, it is sensitive to the number of clusters and the shape of clusters [45]. The exact formulation of the Davies-Bouldin Index (DBI) can be found in mathematical details section of the Appendix.

Finally, the third validation technique examined in this report is the Calinski-Harabasz Index. The Calinski-Harabasz Index, also known as the Variance Ratio Criterion, evaluates the quality of clustering by comparing the ratio of between-cluster dispersion to within-cluster dispersion. It measures the compactness and separation of clusters, with higher scores indicating better clustering. The index is computationally efficient and robust to multiple cluster shapes and sizes. It favours clusters that are well-separated and compact, making it suitable for assessing clustering algorithms. However, it tends to favour convex clusters and is sensitive to outliers [46]. The exact formulation of the Calinski-Harabasz Index (CHI) can be found in mathematical details section of the Appendix.

4.2 Bayesian causal model

In order to draw accurate causal conclusions, it is essential to carefully analyse the underlying mechanisms that impact the overall probability of a customer leaving their insurance.

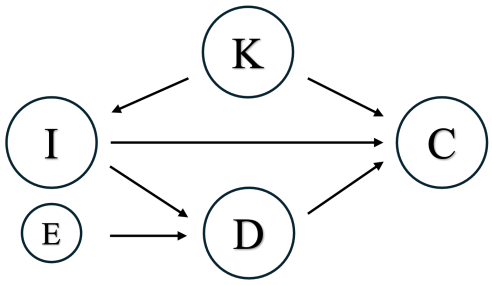


Figure 2: Directed acyclic graph of causal effects

Figure 2 provides an overview of the observed mechanisms within our dataset. Firstly, the variable set K, containing variables $m = 1, ..., 5$, includes key indicators for the four types of main coverages a customer may possess, along with the number of supplementary coverages. Variable set I, containing variables $j = 1, ..., 4$, includes relevant customer and car characteristics utilized by AEGON to determine eligibility for a welcome discount and/or list price adjustment. These characteristics consist of the number of claim-free years, customer and car age, car value, and geographical region, denoted by the customer's four-digit postal code. However, due to the encoding complexity of this categorical variable, it is omitted from the model. Component E denotes the pseudo-random determination made by the insurer regarding eligibility for the adjustment or discount. It can assume values from 0 to 3, representing different types of discounts available to customers. Specifically, 0 indicate no welcome discount or list price adjustment, 1 and 2 represent receiving either of these discounts individually, and 3 indicates receiving both simultaneously. The treatment variable D is determined by components E and I. While we possess data on the level of the welcome discount applied, information on the size of the list price adjustment is incomplete. Consequently, we will comprehensively model the treatment intensity utilizing E and I, without modelling the treatment intensity as two auxiliary outcome variables. Finally, the outcome variable of interest is c, indicating whether a customer churns or not.

For clarity purposes, the unobservable confound U is omitted from this graph. This confound is a catch-all term that indicates the influence of unobservable factors, and impacts all variables but D and R. One example of a known unobservable confound is the impact of the insurers pricing scheme or marketing strategy. This affects the type of coverage that new customers might opt for, as well as what kinds of characteristics they possess and their general expectancy to leave. We unfortunately cannot control for this unobservable factor, limiting the kind of causal inferences that we (or anyone) can perform. Any inferences regarding propensity to churn in the first year will be conditional on the customer taking out an insurance, and the estimated effects on the probability of churn in subsequent years will also be dependent on the customer remaining in the previous year(s). Once a customer has churned, there will be no new entries in our dataset for them.

4.2.1 Model equation

Using the graph, we can construct the relation of the variables that we wish to estimate. Although we are primarily interested in estimating the relation between T and c, two paths of associations exist between T and c that run through I, and the changes in c can also be affected by K. To avoid the (direct) effect of the confounds I and K, they need to be included in our overall model. Next to this, the effect of I on D creates a so-called backdoor path [19] that creates further obfuscation of the effect of D on c. We can resolve this relation by conditioning on I *in* D. This is complicated slightly by the fact that D is also affected by E. Since E is categorical however, we can easily estimate the relation between I and D separately for each category. We therefore define the following model based on the graph.

$$\begin{aligned} c_{it} &\sim Bernoulli(p_{it}) \\ ln\Big(\frac{p_{it}}{1-p_{it}}\Big) &= \alpha_t + \phi_{E_i,t} + \beta_t' I_{it} + \gamma_t' K_{it} \\ \phi_{E_i,t} &= \alpha_{E_i,t} + \beta_{E_i,t}' I_{it} \end{aligned} \tag{1}$$

Here, c_{it} indicates whether an individual i churns in period t or not, where $t = 0, 1, 2$ and refers to the number of years since the start of the policy for each individual, depending on when the customer joined and potentially already left. AEGON's experiment has only run since 2021, so only three time periods are available at most. Since we are dealing with a binary outcome, we will model the log-odds ratio of the probability of churn. As discussed in section 3.2, we could also adopt a probit model instead of a logit-link function. We will use a logit model because the parameters are generally easier to interpret as an effect on the log-odds ratio of churning.

We allow for the coefficients to vary for each time period t, since the treatment might influence churn differently in different years. The variable sets I and K vary over time, as the specific characteristics of the policy, customer, and car may change slightly over time (such as their age or insurance configuration). In this equation, the dimension of the parameter vectors correspond to the cardinality of the variable sets. The key variable of interest will be $\phi_{E_i,t}$, as this will indicate to what extent the churn probability is impacted by the treatment compared to the base churn-level, adjusted for the intensity of said treatment based on the variables I. In addition, we can infer what effects the individual characteristics have by investigating the coefficients for $\beta_t + \beta_{E_i,t}$ and γ_t. For identification purposes, we let $\beta_{0,t} = \alpha_{0,t} \approx 0 \Rightarrow \phi_{0,t} \approx 0$ though the priors we use. Without this restriction, we cannot estimate general effect of the characteristic variables I in the overall equation for the log-link function. This assumption is also sensible because if an individual is determined to be ineligible for a treatment, its intensity if they were eligible is irrelevant, and this makes the general equation serve as the baseline churning model.

4.2.2 Estimation

We will employ a Bayesian approach to estimate this model across the entire dataset, aiming to derive causal effect estimates for the entire population. As we do not have much prior knowledge of the effect of our variables on churn from previous literature, we will use weak priors to estimate

the model. These are detailed as follows:

$$\alpha,\ \alpha_{E_i},\ \beta_j,\ \beta_{j,E_i},\ \gamma_m \sim \mathcal{N}(0, \Sigma_{(\cdot)})$$

$$\Sigma_{(\cdot)} = \sigma^2_{(\cdot)} R_3, \quad \sigma_{(\cdot)} \sim \text{half-Cauchy}(0, s_{(\cdot)}), \quad R_3 \sim LKJ(1)$$

where $(\cdot) = \alpha, \beta, \gamma$. To mitigate the uncertainty in estimating treatment coefficients, we employ coefficient priors that assume an indifferent prior belief regarding the marginal effect of all variables. These priors are designed with parameter vectors of size 3, corresponding to the number of time periods. To reduce the uncertainty of our estimation of the treatment coefficients, we pool the time-varying effects and assume they originate from the same distribution, a common approach in panel data analysis [25]. Variances are drawn from two different half-Cauchy distributions, following the approach outlined by Polson and Scott [47], with s_α set to 25, a frequently used default [48]. Additionally, we set c_β and c_γ to 10 to reflect our expectation of smaller variances for non-intercept coefficients, although this choice is somewhat arbitrary. The correlation matrix between time periods is defined using R, configured to the flattest setting using the Lewandowski-Kurowicka-Joe distribution, which approximates a uniform prior on the interval [-1, 1] but is better suited for estimation procedures [49]. Moreover, for α_{E_i} and β_{j,E_i}, we constrain the covariance matrix Σ to have infinitesimal values for $t = 0$. While it is feasible to incorporate more specific variance configurations for each parameter, for simplicity, we homogenize them into two broad categories. Furthermore, all β and γ coefficients are estimated separately for each variable, and no correlations between these variables are taken into account. Incorporating two layers of correlations would complicate estimation significantly.

To explore potential heterogeneity within the customer base, we will utilize the optimal clustering identified through our clustering methods. While we incorporate the assigned cluster as an additional variable in equation 1, a more intriguing approach is to segregate individuals into separate datasets based on clustering. This allows us to run the model independently on each dataset and assess whether the overall treatment effect α_{E_i} or $\alpha_{E_i,t}$ varies across different consumer clusters. We focus on the general effect, as the covariate spaces for the general model and cluster-specific models may not be comparable, resulting in differing coefficients for all β variables.

Subsequently, the model is employed to derive posterior densities of the likelihood to churn for each individual. These densities can then be aggregated at various levels, such as each treatment category and each cluster. This approach provides a comprehensive understanding of the expected churn rate from one year to the next, including all associated uncertainties. However, due to the dataset's limited number of time periods, the evaluation of the model and its forecasts is constrained to some extent.

To evaluate the predictive ability of the model, we use the 2023 data as hold-out sample. To evaluate the forecasts of the model for $t = 0$, we then train the model on other $t = 0$ data (with modified dimensions of the prior distributions), and then use the estimated posterior distributions to generate predictive densities and binary predictions for the 2023 data using its covariates. For $t = 1$, we can replicate this procedure, but include $t \leq 1$ data instead. For $t = 2$ the forecasts cannot be evaluated, as no data is available for $t = 3$ before 2023. To allow

for comparison between the causal model and other models, we randomly split the dataset into training and testing samples based on the policy identifiers. Further details will be provided in Section 4.4.

4.3 Survival Analysis

This section presents the application of discrete-time survival models and random survival forests applied in this case. We will divide this section into two parts. First, we will introduce discrete-time survival analysis, providing insights into the model's mathematical concept and underlying assumptions. Then, we will shift our focus to implementing random survival forests. These approaches introduce and compare econometric and machine-learning techniques, highlighting a trade-off between prediction accuracy and interpretability.

4.3.1 Discrete-time survival analysis

Model Setting

Given the data has been recorded yearly, and all time-varying information remains the same within a year, we will formalize the **discrete-time survival model** based on Singer & Willett's settings [34]. We denote the random minimum time elapsed before customer i's churn (T_i) or being censored (C_i) as $T_i^* = \min(T_i, C_i)$, and divide the whole continuous period of the customer holding the car insurance into time intervals, denoting as s, where ($s = (t_0, t_1], (t_1, t_2], \ldots, (t_{s-1}, t_s]$), where $t_0 = 0$. For time interval s, a constant hazard rate h_s is defined as a conditional probability that the event happens at the interval s if and only if the customer has not churned before the interval, as the event occurrence is intrinsically conditional [34], shown in equation 2.

$$h_s = Pr[T = s | T \geq s] \tag{2}$$

Then, to capture the customers' heterogeneity, we rewrite the probability as a deterministic logistics function depending on time indicator $Year_{is}$ and observable time-invariant (Z_{is})/time-varying (X_{is}) characteristics variables [50, 32]. Equation 3 gives the logistic transformation proposed by Cox [32] and the equivalent conditional log-odds form, where the parameters α's capture the baseline logit-hazard function as they capture the impact of time indicators on the log-odds ratio of the hazard probability given other variables fixed. Parameters β and γ show the influence of each independent variable on the log-odds ratio. For calculation simplicity, three assumptions have been made: 1) no unobserved heterogeneity exists; 2) the relationship between the log-odds ratio and the included variables is linear; 3) there is no time-varying coefficient.

$$\begin{aligned} h_{is} &= \frac{1}{1 + exp(-[\alpha^T Year_{is} + \beta^T Z_{is} + \gamma^T X_i])} \\ \iff ln(\frac{h_{is}}{1 - h_{is}}) &= \alpha^T Year_{is} + \beta^T Z_{is} + \gamma^T X_i \end{aligned} \tag{3}$$

Thus, the probabilities for censored and non-censored individuals can be defined. For an uncensored individual churning in the time period s_i, the probability can be written as equation 4. For a censored individual, we compute the survival probability, which is the churn that will

happen after the time period s_i, as shown in equation 5.

$$\Pr(T_i = s_i) = h_{is_i} \prod_{s=1}^{s_i-1} (1 - h_{is}) \tag{4}$$

$$S_i(t) = \Pr(T_i \geq s_i) = \prod_{s=1}^{s_i} (1 - h_{is}) \tag{5}$$

Based on the assumption of independent customers, the likelihood function can be written in the form of equation 6, where $c_i = 0$ if data is censored and $c_i = 1$, when a customer churns. The log-likelihood can be rewritten as the logistic regression (see (7), derivation in Appendix), where y_{is} is the churn indicator, a dummy indicating if a customer churns or not. Therefore, logistic regression will be applied to estimate the parameters of survival analysis using the person-period dataset. [34] (Table 5 in Appendix). The dependent variable is the churn indicator. The explanatory variables include the time indicators $Year_{is}$ and characteristic variables X_i and Z_{is}.

$$L = \prod_{i=1}^{n} \Pr(T_i = s_i)^{c_i} \Pr(T_i \geq s_i)^{1-c_i}. \tag{6}$$

$$l = \sum_{i=1}^{n} \sum_{s=1}^{s_i} \left[y_{is} \ln \left(\frac{h_{is}}{1 - h_{is}} \right) + \ln(1 - h_{is}) \right] \tag{7}$$

Model Specification

In this section, we will present our model specification to illustrate the connections between churn and the explanatory variables under the assumption that the baseline hazard remains consistent for each time interval across individuals. In this specification, we introduce the hazard rate function as equation 8, wherein we assume that customers receiving a welcome discount exert both direct and indirect effects on churn. This assumption originates from the understanding that discounts not only directly affect customer premiums but also influence the dynamics of the customer-company relationship, potentially altering other parameters [51]. To model this, we introduce a dummy variable w_i, indicating whether a customer received a welcome discount (equal to 1 if received). This setup allows us to measure both the direct impact of the discount value on churn probability and the indirect effects produced through other parameters. This approach enhances model flexibility compared to scenarios where discounts are assumed to shift the hazard rate solely. Moreover, we incorporate optimal clusters and discount eligibility as time-independent variables to capture observed heterogeneity.

$$h_{is} = \frac{1}{(1 + exp(-[\alpha_w^T Year_{is} + w_i(\beta_w^T Z_{is} + \gamma_w^T X_i) + (1 - w_i)(\beta^T Z_{is} + \gamma^T X_i)]))} \tag{8}$$

It is noticeable that the current specifications are based on other important assumptions mentioned in the model setting section. We assume independence among customers as their locations are dispersed. For simplicity of calculation, no time-varying coefficient is assumed. In addition, the common baseline hazard rate assumption is made, as if the dataset is split based on the known groups, models will not be nested as most observations in our model

did not receive a welcome discount, which makes different specifications difficult to compare. Moreover, the non-unobserved heterogeneity assumption is relatively strict, even though we add some group-based variables. In this case, considering the length of the data collection period and the large data size, we impose the assumption satisfied. However, violating this assumption can influence the fitted hazard rate and generate biased estimators [52]. Another important assumption to mention is the linearity relationship of the log-odds and the variables. The assumption can be validated by incorporating the generalized additive model (GAM) with logistic regression [53] or other statistical inference methods. We explored the GAM model (see Appendix). However, we found no significant improvement in AIC, suggesting that the linear assumption is acceptable, particularly given the interpretability requirement [34]. Additionally, in the next section, we will introduce the random survival forest algorithm to capture potentially more intricate relationships.

4.3.2 Random survival forest analysis

In this section, the random survival forest (RSF), a non-parametric ensemble method, is introduced for its machine learning approach. The RSF algorithm is suitable for the dataset at hand and is able to handle right-censored data and complex variable interactions. It effectively identifies significant variables influencing survival time. Moreover, unlike the discrete-time model, the RSF does not require many of the strict assumptions included in semi-parametric models.

Random survival forest algorithm

The random survival forest (RSF) algorithm is an extension of the random forest technique for survival analysis. It is good in handling right-censored survival data, taking both censored and uncensored observations into account during model training. By recursive partitioning, the RSF constructs survival trees from the bootstrap samples extracted from the dataset. This intentional randomness enhances model robustness and enables precise predictions.

Before implementing the RSF algorithm, the data must undergo formatting for survival analysis. This involves structuring the survival outcomes to indicate the event and the time until the event, or censoring. The time to event or censoring measures the duration until the event transpires for both uncensored and censored cases. Hence, the survival time (t) represents a continuous variable, where $t \geq 0$, intending how long each individual remains until the event occurs or the observation period ends. Until the observation period ends, the event indicator remains a binary variable denoting whether the event of interest, in this case churn, has already occurred or not.

In this framework, predictor variables are utilized for both estimation and prediction. When implementing the algorithm, the necessary hyperparameters include the number of survival trees to be grown (n estimators), the subset of features considered at each node (max features), the maximum depth of the trees, the minimum sample split, and the minimum sample leaves for depth and complexity adjustment. Optimal hyperparameters are determined through a grid search and k-fold cross-validation to enhance the model's accuracy. Initially, the data is split into training and testing sets, followed by repeating the algorithm on 1,...,k equal subsets of the training dataset. The combination of optimal parameters is then selected based on the

estimation with the lowest error rate. The error rate, calculated as 1 - C, where C represents Harrell's concordance index [38], guides the parameter selection process. The subsequent steps outline how survival trees are constructed in the training set for each k-fold:

1. Draw B (`n_estimators`) same-size subsamples from the training dataset.
2. For each bootstrap sample $b = 1, \ldots, B$, a survival tree is constructed:
   a) Select a random subset of features at each node, which will be candidates for splitting (`max_features`).
   b) Determine the optimal split using the log-rank test which compares the survival curves between two groups by looking at the observed and expected number of events at each time point.
   c) Repeat steps a) and b) recursively to grow the tree until the stopping criteria are met, which are determined by the `min_samples_split` and `min_samples_leaf`.
3. Calculate the aggregated survival function for each individual using the B survival trees and use the aggregated information from the trees to get a risk prediction ensemble.
4. Predict the error rate for each validation set.

Implementing this algorithm allows for a survival function estimation for each individual in the subset. The survival function, denoted as $S(t) = P(T > t)$, forecasts the survival probability of an individual at a given time $t \geq 0$, where the time to event (T) exceeds a specific point. Moreover, the survival function is interconnected with the hazard function through $S(t) = exp(-H(t))$. This function projects the survival probabilities beyond each time point within the observation period.

Additionally, another predictive outcome includes risk prediction, which involves assigning risk scores to individuals representing the expected likelihood of the event occurring. These scores are subsequently employed to rank individuals and identify those with a higher risk of experiencing the event.

The final outcome to be used is permutation importance, a metric employed to assess feature importance by measuring the model's performance when the value of a feature is randomly shuffled. The importance feature represents the importance of a feature to the split at different nodes and, therefore, also in the decision to assign the different survival probabilities per individual.

Lastly, running the RSF algorithm on the test set will determine the final results. A final model assessment will be performed by predicting the feature importance, risk scores, and survival function on the test set, as well as using accuracy metrics to evaluate the model's performance.

Model Setting

This section offers insight into the implementation of the RSF algorithm on the AEGON dataset, aiming to establish comparability with the Discrete-time survival model. The variables used for the RSF model are those used in the other survival model. However, unlike the discrete-time

model, the RSF model is based on continuous time, as it predicts the time to event. The dataset is transformed by taking the last observation from each individual and making this into a single observation per individual during five years (from 2019 until 2023), where $t = 0$ for the year 2019. The predictor variables include continuous time variables and descriptive or categorical variables from the insurance policy and customer.

For the model setting, there needs to be a time and event dependent variable (y) and predictor variables (x) that explain the time to event. In this case, the time is the variable that states the years since the policy started, and the event variable is the churn variable. The end of the observation period is in 2023, therefore the observations span over 5 years. The independent variables are the same as those in the discrete model, with time now being continuous rather than discrete. Subsequently, the dataset is partitioned into training data, consisting of 70%, and testing data, consisting of 30%, such that the model can be trained on the train set. The RSF algorithm is then applied to the training set for hyperparameter optimization using a grid search and 5-fold cross-validation.

After hyperparameter optimization, the optimal parameters are applied to the test set to get the final survival function over a 5-year period for each individual. The survival probability of the individual not churning is predicted for 5 years (t) after the year 2023 (T). The final risk prediction and importance features are predicted using the test set.

4.4 Model comparison

Although the three models are structurally very different, each has the capability to predict whether an individual will churn. To evaluate which of these models performs best, we split the data into a train dataset with 70% of the data points and a test dataset of 30% of the data points, where the final predictions are performed on the test set. The survival models and causal model aim to estimate the effect of having discounts (or specifically the welcome discount) on the customer churn rate. The prediction performance of the machine learning and econometric models will be assessed using the Brier score (BS) and the area under the ROC curve (AUC). The Brier score is a commonly used metric that mainly measures calibration [54], while the AUC exclusively measures discrimination of the models.

The Brier score is time-dependent, and the function is $BS(t, s)$. This score is defined as the mean square error between the binary true event indicator and the predicted survival probability of churning at each time point in the prediction window, measuring the accuracy of the probabilistic predictions. The values range between 0 and 1, with lower values indicating better predictive performance in this context. We use the survival probabilities from the survival function of the models to predict the BS for each year over the five-year window. For the causal model, we transform the predicted churn probabilities to survival probabilities and perform the same computations.

The time-dependent $AUC(t, s)$ will be used in this case, where t is the time indicator and s is the prediction window indicator, based on the time-dependent ROC curve [55]. This metric determines how well the predicted probabilities discriminate between customers likely to churn or not along each time point in the prediction window. The values range between 0 and 1, with higher values indicating better discrimination. The benchmark score of AUC is 0.5, the

score generated by random guessing. The random survival model applies the risk predictions to determine the AUC at each time point over the 5-year window, whereas the causal model evaluates it over a 3-year window, as its dataset spans from 2021 to 2023.

Finally, to enable a comprehensive comparison of the overall model output, we will also determine the survival probability for both the discount and non-discount groups of customers across each model. This will be dependent on whether they received a discount or not. While all models are expected to present values close to the unconditional probability, the different model structures might generate interesting differences when the groups are analysed separately.

5 Results

5.1 Clustering

All cluster setups are assessed based on the Silhouette Score, Davies-Bouldin Index and Calinski-Harabasz Index. The exact results can be found in Table 8 in the Appendix.

The silhouette score is ranging from -1 to 1, a score close to 1 resembles clearly distinguishable clusters that are well separated, a score of 0 means that the clusters are not significantly distinguishable and not significantly separated. -1 is related to wrongly assigned clusters. All cluster setups are close to zero, therefore all cluster setups are not providing great clusters according to the silhouette score. For the Davies-Bouldin Index, a lower value is preferred and for the Calinski-Harabasz Index a higher value is preferred.

Since the 3 evaluation methods have different value ranges, all values of the 3 scores are individually normalized, making it more convenient to compare the methods. Besides normalizing the values, the normalized values of the Davies-Bouldin score are mirrored along the vertical 0.5 line (since lower Davies-Bouldin values are indicating better clustering performance), making it easier to compare the Davies-Bouldin scores with the 2 other methods.

Moving to the evaluation, plot 3 reveals that the K-Prototypes clustering with 2 clusters performs best on all evaluation metrics. Besides that, the smallest cluster of the K-Prototypes clustering contains 38008 unique customers. Therefore, not making cluster size a constraint. Hereby, we conclude that K-Prototypes with 2 clusters is the preferred clustering method for this report. An overview of non-normalized cluster evaluation outcomes can be found in 8 in the Appendix.

To conclude, a more general note is that based on the clustering elevation criteria can be observed that generating very distinct clusters is not established (as is indicated by the silhouette score). Figure 4 gives a graphical description of observations of accident free years vs car value and clearly reveals overlapping clusters. This can be related to the methods, but also to the data itself.

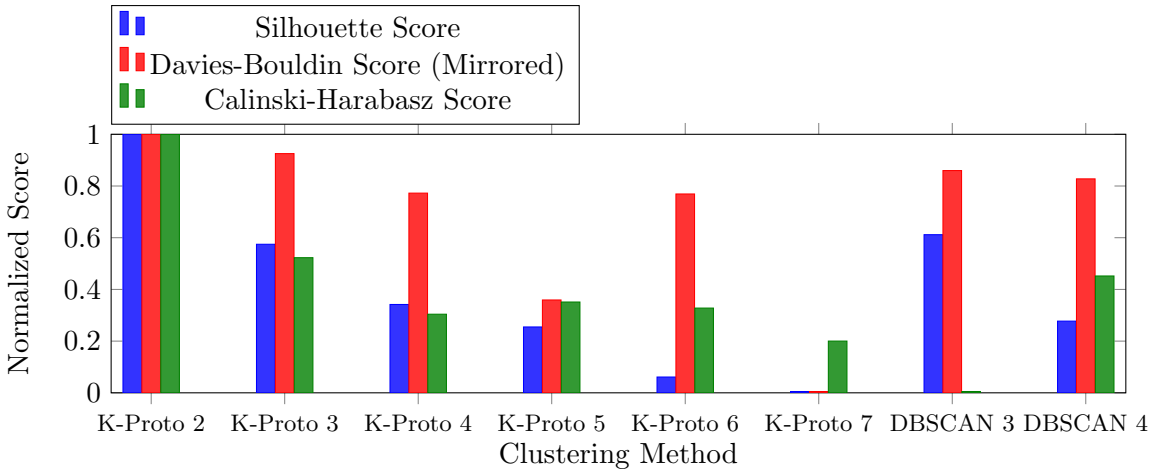


Figure 3: Normalized Clustering Validation Scores

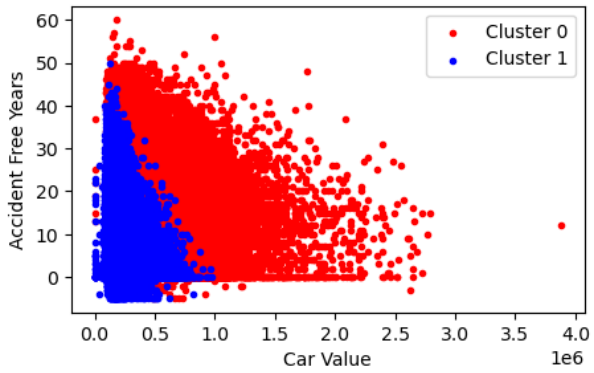


Figure 4: Plot of observations of Accident free years vs Car value per cluster

5.1.1 Cluster interpretation

For cluster interpretation, we employ a plot depicting the mean of numerical cluster variables per cluster relative to the overall mean of those variables to explain the numerical characteristics of each cluster. Given the overabundance of categories within categorical variables and the impracticality of describing clusters through categorical labels (with fuel types being the exception), we exclusively consider numerical variables plus fuel types for cluster interpretation, with the fuel type distribution detailed in Table 7. The 2 cluster are characterized as follows.

Cluster 0 consists of older customers characterized by an increased average of accident-free years and ownership of newer, higher-value vehicles, often insured comprehensively. Notably, most electric car drivers are present in this cluster.

Conversely, Cluster 1 is characterized by relatively younger customers with fewer accident-free years on average, and possession of older, less valuable vehicles, frequently insured without comprehensive coverage. Additionally, Cluster 1 represents a minimal presence of electric car drivers compared to Cluster 0.

The exact mean values of the numerical variables for each cluster can be found in Table 9.

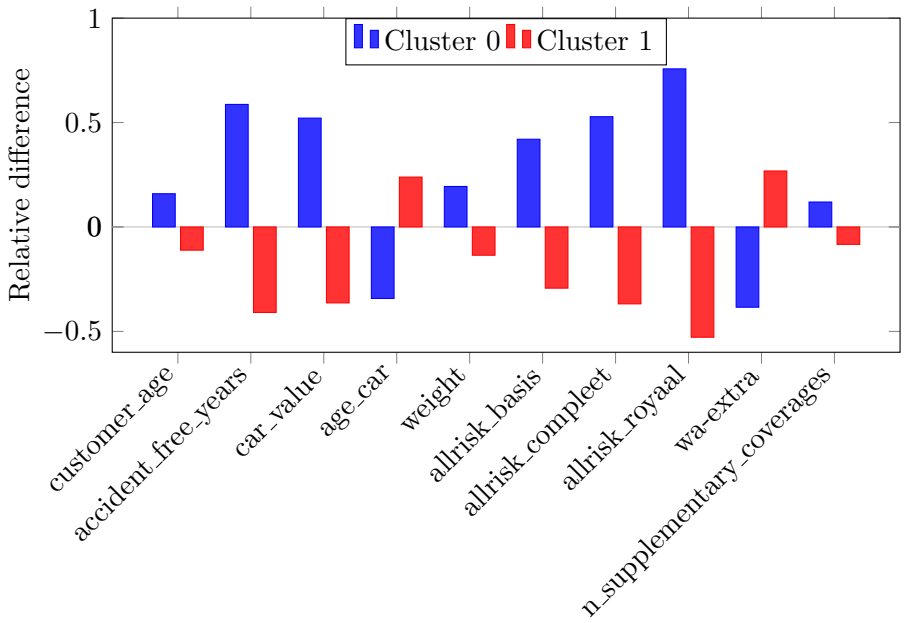
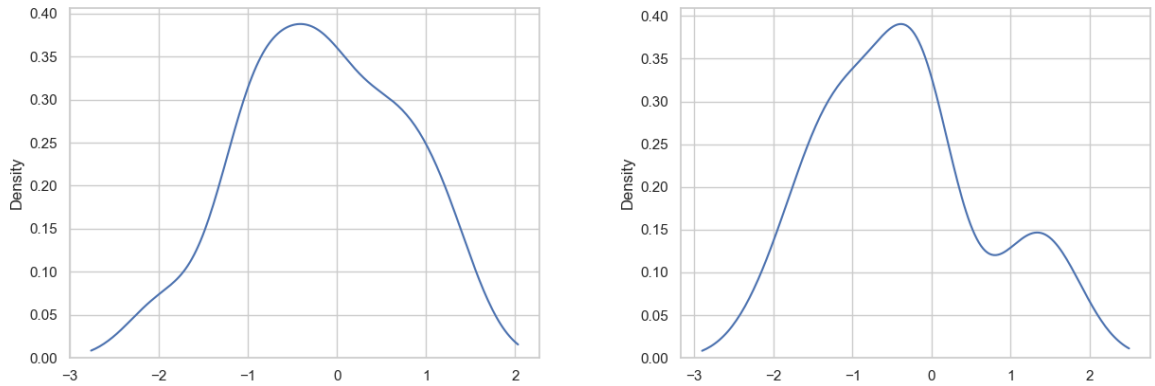


Figure 5: Bar Plot of the relative difference of the mean of the clusters for each numerical variable

5.2 Hierarchical Bayes

To estimate the causal model, the Hamiltonian Monte Carlo (HMC) algorithm with No U-Turn sampling is used. For this, the programming language Stan is used, implemented in Python though the pystan package [56]. We generate a total of 16 chains of 1000 parameter draws, of which 250 are used as warm-up draws and subsequently 750 are stored. Ideally this number is set as large as possible, but given a roughly 90-dimensional parameter space, computational complexity is a limiting factor. Given a posterior parameter draw, we can generate the estimated treatment effect as well as the posterior probability for churning for each individual. These values can then be aggregated across all draws to derive a point estimate of the churn probability for each individual.



a) Welcome discount, first year

b) Welcome discount, second year

Figure 6: Posterior density of treatment-specific intercept $\alpha_{E_i,t}$

Figure 6 shows the posterior distribution of two $\alpha_{E_i,t}$ coefficients as an example. It can be

seen that this distribution is somewhat skewed and may even display two local modes. Though this behaviour is not uncommon behaviour given that our target variable is binary, it is strictly speaking not possible to identify if the model assigns a positive or negative coefficient to this value at this point. The 95%-highest posterior density-interval of these parameters (a Bayesian analogue for the confidence interval) all contain the value zero, so we cannot reject the idea there are no significant differences over time in general churning behaviour after taking into account all variables. In fact, this is the case for nearly all estimated parameters, which, with 90 parameters to estimate, would be expected in the case of complete absence of any effect. In short, the Bayesian model fails to identify any clear individual variable and/or treatment category-specific effects, both in terms of levels (by the intercept) and slopes.

There are two possible reasons for this. Firstly, it could be the case that the prior distributions are set too wide, causing the model's initial draws to be unlikely in the first place, causing the entire HMC-chain to consist of unlikely parameter draws. Running the model again with half-Cauchy parameters $s_{(\cdot)}/10$ yielded similar inconclusive results, however. Second, the model could suffer from identifiability problems due to the number of parameters that are estimated. Given our causal model, however, the only assumption that would be permissible to impose is that the coefficients are invariant over time. Since it can be theorised that churning behaviour is different over time, and in our case it is one of the key questions at hand, valuable information would be lost in this case.

Although the estimated parameters are not significant, we can compare the relative average size of the treatment effect $\phi_{E_i,t}$ in our different models to investigate whether customer heterogeneity can be identified. Table 1 outlines the mean $\phi_{E_i,t}$ values across the different parameter draws and individuals that fall under each specific category. All these values are negative, with more negative values indicating a decreased likelihood to churn. These strongly negative values are compensated though the positive contributions of the general β and γ parameters, as defined in equation 1.

Table 1: Estimated treatment effects on log-odds ratio

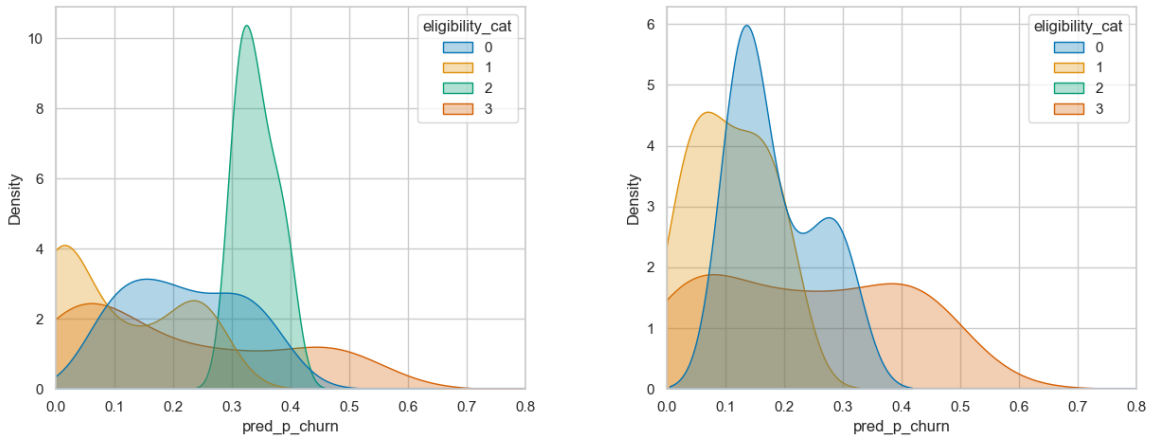
ϕ		Mean parameter value			
E_i	t	Full model	Dummy model	Cluster 0	Cluster 1
1	0	-65259	-41755	-23395	-7734
	1	-239221	-259384	-295987	-113601
	2	-302754	-364315	-394101	-144376
2	0	-68333	-86922	-53547	-18845
	1	-203729	-201430	-279449	-103279
	2	-119442	-221710	-253965	-93677
3	0	-16899	-177	-2074	-49
	1	-57139	-97635	-90996	-20331
	2	-417703	-529705	-605259	-217311
Accuracy		0.7806	0.7438	0.7630	0.7071
F1-score		0.1328	0.1666	0.1301	0.1612
η			-0.1412		

Examining Table 1, one can observe that the magnitude of the treatment effects in the full model is comparable over time for category 1 and 2, so receiving either one of the discounts. Category 3 displays lower values overall for $t \leq 1$, indicating more proneness to churn. The values for $t = 2$ do not exactly fit this pattern, though are expected to be more extreme because of the relative lack of data; with relatively flat priors, the posterior density can be swayed either way more easily when there are few data points.

Unfortunately, though perhaps expectedly so, the standard errors of all these mean treatment effects are usually more than half the (absolute) value reported here, indicating that they would not be significant in the frequentist sense. A key problem in our dataset that may lead to this result is that the discount-eligible groups are inherently not very different from the ineligible group if the majority of the eligible customers are not given a discount. Having data on whether a list price adjustment was applied would allow for a much more distinctive separation of the data.

Comparing the results of the full model with those catered to our clustering method, we see that for both the dummy model and the separate cluster models, their relative magnitude largely stays about the same, displaying an increasing magnitude over time in categories 1 and 3, and $t = 1$ displaying the largest magnitude in category 2. The dummy coefficient η that is added to the full model displays a slightly negative value, but this is again insignificant. We therefore conclude that applying clustering beforehand to identify segments did not yield any considerably different results in our causal estimation procedure.

Similar to R-squared for in-sample fit, we compute each individual's estimated probability of churn by applying each parameter draw to derive a singular probability. These probabilities are then averaged, and subsequently used to randomly generate a binary outcome. The accuracy and F1-score can then be determined and is also reported in Table 1. The out-of-sample fit was computed by making predictions for 2023 without using this data to train the model. For new customers in 2023, so $t = 0$, the model obtained an accuracy of 67.6% and an F1-score of 14.9%. For customers that have been with AEGON for a year already, so $t = 1$, the model obtained an accuracy of 71.9% and an F1-score of 12.3%.



a) First year of policy

b) Second year of policy

Figure 7: Churn probabilities per discount eligibility type

Despite the instability of the model, we can still visualise the distribution of estimated churn probabilities for the individuals in our data set. Figure 7 displays the posterior predictive density of the churn variable for each category in two time periods. Treating category zero, receiving no discounts by design, as the baseline, we can see that generally, in the first year, a lower likelihood to churn is predicted for those that are eligible for a welcome discount (category 1) and a higher likelihood to churn is predicted for those eligible for a list price adjustment. The result of being eligible for both seems to be that the probability of churn is even harder to predict, with its distribution being relatively flat across the domain. This is echoed in the plot for the second year, though the density for category two is omitted, as it is very concentrated around 0.35. This has likely happened because each chain in the simulation has always assigned probabilities of either zero or one, and the chain of parameter draws never generated enough movement to cause this to change within a chain. Nonetheless, the model seems to hint at welcome discounts being a useful tool to retain customers, given the current price optimisation algorithm AEGON uses. However, with the posterior densities of the parameters not displaying sufficient evidence, this conclusion should be used with caution.

5.3 Survival analysis

5.3.1 Discrete-time survival model

We applied a discrete-time survival analysis to describe the influence of the discounts and groups of customers. Table 2 (Table 10 containing all variables, is included in the Appendix) presents the parameter estimators, the corresponding standard errors, and the marginal effect. The estimator's sign indicates the relationship between the variables and the customer's hazard rate, while the average marginal effect represents the average change in predicted probability for a one-unit change in a variable across all variable values.

According to Table 2, the baseline hazard rate decreases over the years as the parameters are negative, ceteris paribus. Given the other variables fixed, a higher welcome discount shifts down the hazard rate, promising a lower churn rate in the initial year, which is implied by the steeper slope between year zero and year one in Figure 8. The observed heterogeneity determines the difference levels between the two curves. Car age and car values positively correlate with customers' hazard rate, while a customer with a higher accident-free years record has a lower probability of churn. However, customers' ages, in this case, are not linearly significant in both groups. Yet, there might exist an unknown non-linear relationship [53].

In addition, customers who are eligible to receive a welcome discount and are eligible to receive the list price adjustment are less likely to churn. Clusters determined in the first part are not significant in the model specification, likely because variables contributing to the clusters, such as the number of supplementary coverage, are already accounted for in the model.

An interesting result is that welcome discount eligibility within the welcome received group positively correlates with the churn hazard rate. Given all values in this variable are one, it is difficult to distinguish whether the observed positive correlation is due to a significant increase in total premium between the first and second years or the selection bias of the two different groups.

Table 2: Discrete-Time Survival Analysis estimators

Variables	Estimator(SE)	Average Marginal Effect
year 0	-5.943 (0.498)	-0.668 (0.056)
year 1	-5.628 (0.498)	-0.632 (0.056)
year 2	-5.782 (0.499)	-0.650 (0.056)
year 3	-5.753 (0.500)	-0.646 (0.056)
year 4	-5.898 (0.503)	-0.662 (0.056)
year 5	-8.045 (0.871)	-0.904 (0.098)
size of WD	-0.479 (0.166)	-0.054 (0.019)
accident free years	-0.018 (0.001)	-0.002 (0.000)
accident free years(with WD)	-0.016 (0.003)	-0.002 (0.000)
customer age	0.001 (0.001)	$8.587e^{-05}$ ($9.08e^{-05}$)
customer age(with WD)	-0.002 (0.001)	-0.0002 (0.000)
car value	0.092 (0.013)	0.010 (0.001)
car value(with WD)	0.061 (0.025)	0.007 (0.003)
car age	0.028 (0.002)	0.003 (0.000)
car age(with WD)	0.036 (0.003)	0.004 (0.000)
n supplementary coverages	-0.091(0.018)	-0.010(0.002)
n supplementary coverages (with WD)	-0.118(0.022)	-0.013(0.003)
cluster 1	-0.033 (0.033)	-0.004 (0.004)
WD eligible	-0.105 (0.028)	-0.012 (0.003)
LPA eligible	-0.369 (0.030)	-0.042 (0.003)
cluster 1(with WD)	0.007 (0.05)	-0.001 (0.006)
WD eligible(with WD)	1.046 (0.37)	0.117 (0.042)
LPA eligible(with WD)	-0.062 (0.04)	-0.008 (0.004)

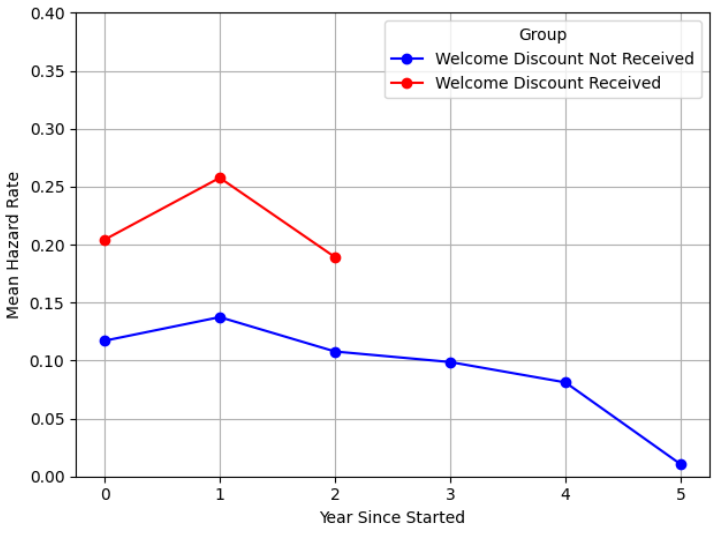


Figure 8: Mean Hazard rate between groups

5.3.2 Random Survival Forest

In modeling the random survival forest, a grid search and cross-validation were conducted on the test set to determine the optimal parameters. The lowest error rate obtained was 0.219, achieved with the following hyperparameters: max depth of 30, min sample leaf of 5, max features set to 'sqrt', and min sample nodes of 10. In the training set, the out-of-bag error rate was 0.217, which was also the lowest observed. The model's error rate stabilizes with 100 survival trees, as described in the Appendix in Figure 13. These optimized parameters were then applied to the test set for prediction and evaluation, resulting in an error rate of 0.221.

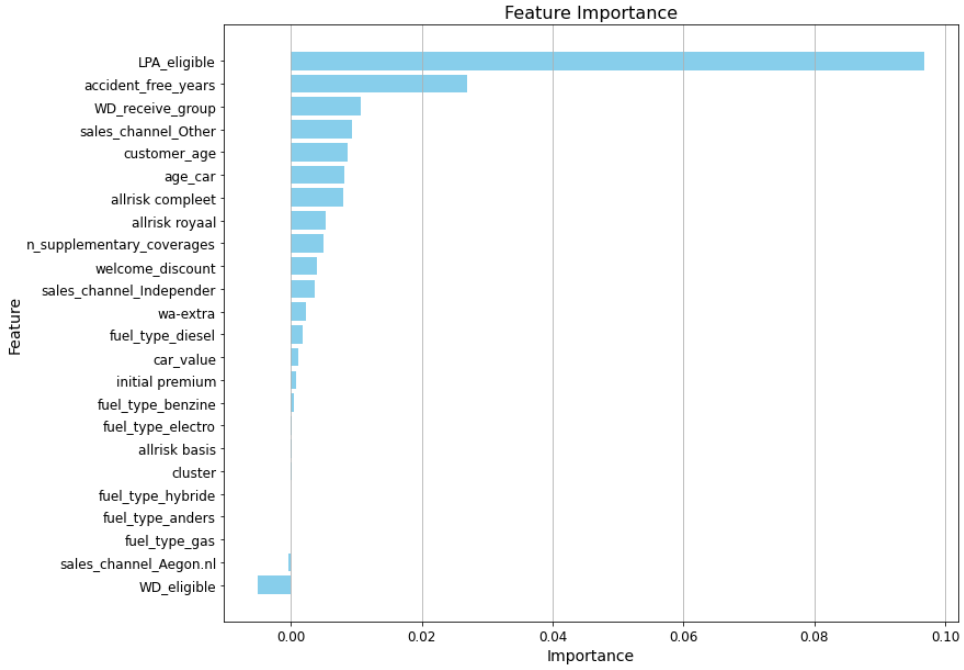


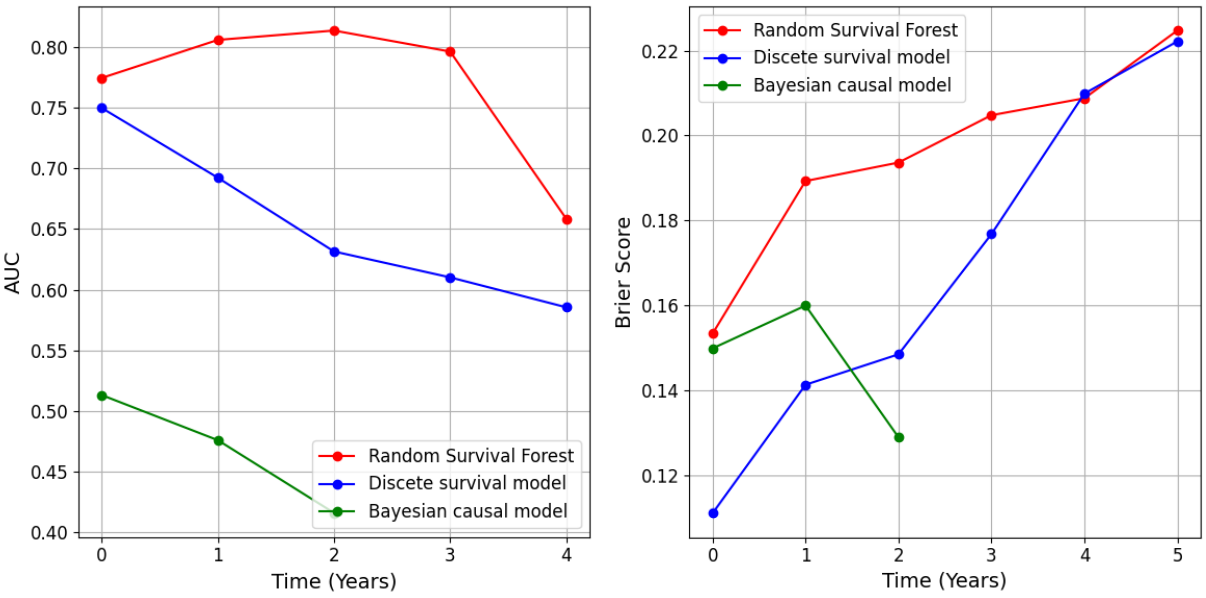
Figure 9: Feature importance variables

The importance value for each predictor is shown in Figure 9. For the feature importance, large positive values indicate informative variables. While values approaching zero or being negative indicate less informative variables [38]. From Figure 9, the variables that are clearly predictive and are significantly larger than other predictors include being eligible for a list price adjustment and the number of accident-free years. The other variables that are informative but have relatively smaller values include customer receiving welcome discount, customer age, car age, size of welcome discount, and different sales channels or coverages. These variables are also significant in the discrete model. The initial premium, car value, clusters, welcome discount eligibility, and fuel types were very small or negative and therefore unlikely to be informative.

5.4 Model comparison

In this section, the survival probability and performance of prediction accuracy will be compared among the causal model, the discrete-time survival model, and the random survival forest model. AUC- and Brier-scores are applied, and shown in Figure 10.

For the AUC-score shown in Figure 10, the RSF model gives the best prediction in every



a) AUC-score over time

b) Brier-score over time

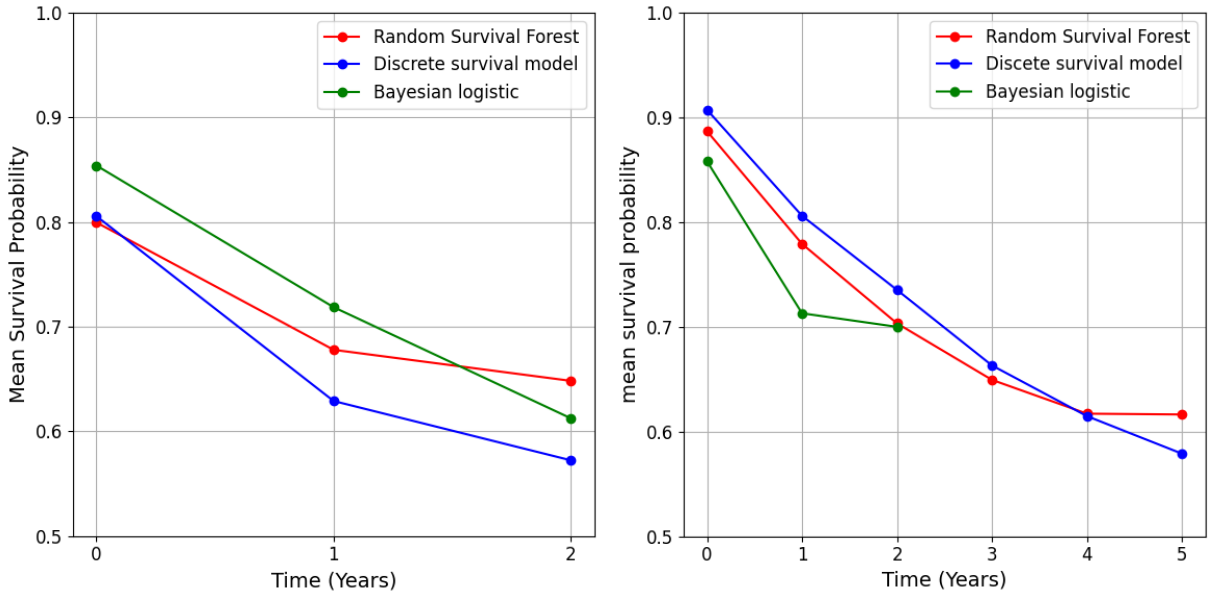
Figure 10: Comparison of AUC and Brier Score

year. For RSF, the AUC increases up until year 2 with value of 0.81 and then decreases up until year 4 with value of 0.66 (with a steep decrease from year 3 to 4). The AUC of the discrete model decreases from 0.75 at year zero to 0.58 by the fourth year. Both models decrease over time since these models are trained on smaller data sets at later prediction times. The scores can only be computed for three periods for the causal model, since only the data from 2021 onwards was available. In Figure 10, the AUC for the first two years is significantly lower for the causal model compared to the other models, even below the benchmark of 0.5 for $t \geq 1$. On average, the RSF gives a more accurate classification at each time point than the discrete-time survival and causal model. The models score differently in terms of the Brier score, where the discrete-time survival model performs the best. It has a value of 0.11 in the initial year, increasing to a value of 0.22 in the last year. RSF gives a slightly higher Brier score on average, starting with a value of 0.15 and ending with a value of 0.22 for year 5. The two models have similar scores in the last two periods, which are rather high but still better than the benchmark score of 0.25. For the causal model, the Brier score is in the middle for year 0 (0.149) and 1 (0.159) and lower for year 2 (0.128).

The different conclusions drawn from the two different metrics are related to AUC and Brier measuring different aspects of prediction performances. Where the AUC measures the accuracy of classification, the Brier-score assesses the accuracy of probability prediction. RSF, as a machine learning algorithm, is optimized to make correct predictions and can capture complexity between the variables. Our result is consistent with this property, as it performs the best when giving classifications. In contrast, the causal and discrete-time survival models are designed to illustrate the relationship between churn and explanatory variables and directly estimate the churn probability at each time point, allowing these models to provide well-calibrated probability estimates, especially when time is involved. On the contrary, the RSF model is trained on the person-level data, which is aggregated over time.

To address the research question regarding which model provides the most accurate prediction, AUC should be favoured. AUC serves as a measure of discrimination, making RSF the preferred predictive model due to its significantly higher AUC score compared to the others. The other two models perform better in predicting probabilities of each year when comparing the Brier score. This is because this metric is a measure of calibration. In addition, the causal model does not perform well in prediction accuracy, probably due to the limited period of data available. The discrete-time survival model is preferred in the case of probability prediction.

Figure 11 illustrates the survival probabilities generated by the causal model, the discrete model, and the RSF, segmented into two groups: those who received a welcome discount at $t = 0$ and those who did not. Survival predictions extend up to 5 years for the survival models, corresponding to the maximum duration observed among customers. However, for the discount group and the causal model overall, predictions only span up to 2 years due to the introduction of welcome discounts in 2021.



a) Welcome discount received group b) No welcome discount received group

Figure 11: Comparison of different predicted survival probabilities

All three models show comparable survival probabilities over time for the non-discount group. The discrete model has on average a slightly higher survival probability from the initial year until year 3 compared to the RSF and casual models. For years 4 and 5, the survival probability remains relatively the same for the RSF, whereas it is continually decreasing for the discrete model. For the discount group, the survival probabilities decrease over 2 years for all three models. For year 0 and 1, the causal model has a slightly higher survival probability compared to the RSF and discrete model, while in year 2 the survival probability is higher for the RSF. Comparing the results from the discount group and the non-discount group, from year 0 up until year 2, the non-discount group had a higher probability of surviving compared to the discount group for all models. Although this difference cannot causally be attributed to the discount itself, it does mean that AEGON should be cautious in applying these discounts to more customers.

6 Conclusion

The insurance company AEGON wishes to investigate what causes customers to terminate their car insurance policy, and wishes to know what discount mechanisms are effective in getting customers to stay. In this paper, multiple approaches were taken to use AEGON's dataset to find the effect of a welcome discount and/or a list price adjustment on churn. First, clustering was implemented to find different segments of customers. After this, 3 models were developed: a causal model and two survival models. The causal model is fully parametric and used Bayesian logistic techniques. The survival models comprise a semi-parametric model, which is a discrete survival model, and a non-parametric model, which is a random survival forest. These models were then compared and evaluated based on accuracy performance.

To see if different segments of customers have different behaviour in churning, a clustering method was implemented. The clusters were created using characteristics of customers and policy information, and different clustering methods were used to evaluate the best method. The best-performing method was k-prototypes clustering with 2 clusters. These clusters were then used in the three different models, to see if there was a difference in behaviour. When the cluster label was included as an additional variable, it did not add significant information for any of the models. The causal model and discrete survival model were also estimated separately for both clusters, but the outcomes failed to show noticeable differences, showing that there appears to be little incentive for AEGON to segment their churn model.

The three developed models yielded similar outcomes in terms of churn probabilities aggregated over the customer base, but they result in different interpretations with respect to their finer details. The random survival forest showed that being eligible for a list price adjustment was clearly predictive, and the received discount and the type of discount were slightly predictive. Unfortunately, this model is unable to tell how these variables affect churn. The discrete model does show the average marginal effect of each variable, however. This model found welcome discounts to have a negative average marginal effect on customers churning in the initial year, though overall they will churn more frequently due to other indirect effects. In addition, car value and car age were found to be positively related to the likelihood to churn, while the number of accident-free years has the opposite effect. The developed causal model was unable to find reliable effects with the current dataset, and reported a negative effect of being eligible for welcome discounts on likelihood to churn.

For all three models, predicting future churn could be done by predicting survival probabilities per customer over time. Using a test set, the predictive accuracy of the probabilities of each model were evaluated using the Brier- and AUC-score. The best-performing model in terms of prediction accuracy was the random survival forest model, which showed the highest AUC score. Overall, the survival curves for all three models showed the same decreasing pattern. The probability of survival over two years was higher for the non-discount group than the discount group for all three models.

Ultimately, the choice of which model to use depends on what aspect of the modelling of churn is considered most important. If causal inference is a must, then a Bayesian approach can be taken, though we have shown that with the current dataset it is difficult to obtain useful results. If predicting churn accurately while taking into account customer lifetimes is the most important,

for instance for making good internal risk assessments, then the random survival forest can be tuned to serve this purpose, at the expense of interpretability. If model interpretability is desired and causal estimates are not a requirement, then a discrete-time survival model may provide the desired balance between determining relations between the explanatory and outcome variables and accurate churn probability estimates.

7 Discussion

This case study encountered several limitations and shortcomings that originate from factors such as the available data, underlying assumptions, and methodological approaches. These challenges are addressed sequentially, beginning with limitations related to clustering, followed by shortcomings with respect to causality, then issues in relation to prediction, and concluding with recommendations for future research.

Starting with clustering, a more efficient approach could involve segregating the data based on time periods required for different analyses. Instead of clustering the entire dataset and subsequently filtering out irrelevant observations, we could directly cluster the subsets of data corresponding to the time periods needed for each analysis. This may lead to more accurate and meaningful clusters tailored to the specific contexts of each analysis.

Shifting our focus to the Bayesian model, an alternative approach could involve utilizing the actual data on the treatment variable D to estimate the coefficients, thereby constructing a multivariate hierarchical model. This approach was not explored in this study due to the lack of data regarding the size of the list price adjustment. It is worth noting that we are already working with panel data, which inherently incorporates two layers of correlation matrices. Introducing multivariate regression would further complicate the model by adding another layer of correlation between the outcome variables, thus increasing the complexity of the analysis. Furthermore, it is worth considering non-linear effects for specific customer and vehicle characteristics included in the variable set I from equation 1, such as age, as evidenced by a variety of studies highlighting their influence [53, 34]. In this paper, we have assumed linearity, which may potentially violate the underlying relationships and impact the results of both the causal and discrete survival models.

When developing our discrete survival model, we imposed the assumption of non-unobserved heterogeneity. This assumption has been examined in specific literature [34, 50]. To mitigate this assumption, frailty models have been proposed, incorporating random effects into the survival model [57]. Additionally, the performance of prediction in discrete-time survival analysis is influenced by the size of intervals [27]. AEGON could potentially enhance model performance by dividing the intervals into smaller (equally-sized) ones.

In performing our random survival forest analysis, a simplification was made due to time constraints where clusters were treated as dummies. However, exploring the option of running separate models on individual clusters could potentially provide more nuanced insights into the survival patterns of different customer segments. Furthermore, in the RSF, the assumption of continuous time, despite the panel data structure, may lead to inaccurate estimations. This issue arises because certain variables, like accident-free years, may fluctuate over time due to factors like accidents, whereas the model assumes a continuous increase. Similarly, premium

variables may exhibit varying changes annually, which may not be accurately captured under the continuous time assumption. However, descriptive variables related to insurance policies or customers align with this assumption and can provide valuable insights into factors affecting customer churn. The random survival forest does give valuable insights into survival probabilities and overall risk factors but still results in a complex interpretation of variables due to its ensemble nature and absence of feature coefficients which the discrete model does have. Also, having long processing time and problems related to over fitting is an issue when implementing this model and therefore needs to be taken into account.

Future research could explore extending survival models to detect causal effects, thereby expanding our understanding of causal survival analysis. Braun and Schweidel illustrate how this could be done for a case study in the telecommunications industry, where more data was available [23]. By combining the two approaches considered separately in this paper, more intricate insights into the drivers of customer churn can be obtained, which can inform more effective customer retention strategies in the insurance industry.

References

[1] Herbert Castéran, Lars Meyer-Waarden, and Werner Reinartz. Modeling customer lifetime value, retention, and churn. In *Handbook of market research*, pages 1001–1033. Springer, 2021.

[2] Ozer Çelik and Usame O Osmanoglu. Comparing to techniques used in customer churn analysis. *Journal of Multidisciplinary Developments*, 4(1):30–38, 2019.

[3] Matthias Templ, Alexander Kowarik, and Peter Filzmoser. Iterative stepwise regression imputation using standard and robust methods. *Computational Statistics Data Analysis*, 55(10):2793–2806, 2011.

[4] Jean Mundahl Engels and Paula Diehr. Imputation of missing longitudinal data: a comparison of methods. *Journal of Clinical Epidemiology*, 56(10):968–976, 2003.

[5] Joshua James A Bravante and Rex Aurelius C Robielos. Game over: An application of customer churn prediction using survival analysis modelling in automobile insurance. *IEOM Society International*, 2022.

[6] Indranil Bose and Xi Chen. Hybrid models using unsupervised clustering for prediction of customer churn. *Journal of Organizational Computing and Electronic Commerce*, 19(2):133–151, 2009.

[7] Kamil Matuszelański and Katarzyna Kopczewska. Customer churn in retail e-commerce business: Spatial and machine learning approach. *Journal of Theoretical and Applied Electronic Commerce Research*, 17(1):165–198, 2022.

[8] Farshid Abdi and Shaghayegh Abolmakarem. Customer behavior mining framework (cbmf) using clustering and classification techniques. *Journal of Industrial Engineering International*, 15:1–18, 2019.

[9] Mariusz Łapczyński and Bartłomiej Jefmański. Number of clusters and the quality of hybrid predictive models in analytical crm. *Studies in Logic, Grammar and Rhetoric*, 37(1):141–157, 2014.

[10] Shuli Wu, Wei-Chuen Yau, Thian-Song Ong, and Siew-Chin Chong. Integrated churn prediction and customer segmentation framework for telco business. *IEEE Access*, 9:62118–62136, 2021.

[11] Mark Junjie Li, Michael Ng, Yiu-ming Cheung, and Joshua Huang. Agglomerative fuzzy k-means clustering algorithm with selection of number of clusters. *Knowledge and Data Engineering, IEEE Transactions on*, 20:1519 – 1534, 12 2008.

[12] Joshua D. Angrist and Jorn-Steffen Pischke. *Mostly harmless econometrics: An empiricists companion*. Cram101 Publishing, 2013.

[13] Jeffrey M. Wooldridge. *Econometric analysis of cross section and panel data*. MIT Press, 2011.

[14] Frank Kleibergen and Eric Zivot. Bayesian and classical approaches to instrumental variable regression. *Journal of Econometrics*, 114(1):29–72, 2003.

[15] Guido W. Imbens and Donald B. Rubin. Bayesian inference for causal effects in randomized experiments with noncompliance. *The Annals of Statistics*, 25(1):305–327, 1997.

[16] Paul S. Clarke and Frank Windmeijer. Instrumental variable estimators for binary outcomes. *Journal of the American Statistical Association*, 107(500):1638–1652, 2012.

[17] Austin Nichols. Causal inference for binary regression with observational data. CHI11 Stata Conference 6, Stata Users Group, 2011.

[18] Guido W. Imbens and Joshua D. Angrist. Identification and estimation of local average treatment effects. *Econometrica*, 62(2):467–475, 1994.

[19] Richard McElreath. *Statistical rethinking: A bayesian course with examples in R and Stan.* CRC Press, 2020.

[20] Rafael A. Moral, Zhi Chen, Shuai Zhang, Sally McClean, Gabriel R. Palma, Brahim Allan, and Ian Kegel. Profiling television watching behavior using bayesian hierarchical joint models for time-to-event and count data. *IEEE Access*, 10:113018–113027, 2022.

[21] Robert Sawyer, Jonathan Rowe, Roger Azevedo, and James Lester. Modeling player engagement with bayesian hierarchical models. *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, 14(1):257–263, Sep 2018.

[22] Wang Hai-wei, Jiang Ming-hui, and Wang Ya-lin. Adding risk in measuring customer value using bivariate hierarchical bayesian approach. In *International Conference on Management Science and Engineering*, pages 89–93, 2006.

[23] Michael Braun and David A. Schweidel. Modeling customer lifetimes with multiple causes of churn. *Marketing Science*, 30(5):881–902, 2011.

[24] Ping Chou and Howard Hao-Chun Chuang. Stochastic churn modeling with dynamic attribution and bayesian estimation. *City, Society, and Digital Transformation*, page 57–71, Dec 2022.

[25] Richard Paap. Lectures in bayesian econometrics, 2023.

[26] James H. Albert and Siddhartha Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669–679, 1993.

[27] Krithika Suresh, Cameron Severn, and Debashis Ghosh. Survival prediction models: an introduction to discrete-time modeling. *BMC medical research methodology*, 22(1):207, 2022.

[28] Michael W Kattan. Comparison of cox regression with other methods for determining prediction models and nomograms. *The Journal of urology*, 170(6):S6–S10, 2003.

[29] Stephen P Jenkins. Survival analysis. *Unpublished manuscript, Institute for Social and Economic Research, University of Essex, Colchester, UK*, 42:54–56, 2005.

[30] Yan Chen, Lei Zhang, Yulu Zhao, and Bing Xu. Implementation of penalized survival models in churn prediction of vehicle insurance. *Journal of Business Research*, 153:162–171, 2022.

[31] Chester KM To, Kwok Pui Chau, and Chi Wai Kan. The logic of innovative value proposition: A schema for characterizing and predicting business model evolution. *Journal of Business Research*, 112:502–520, 2020.

[32] David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

[33] David Roxbee Cox. *Analysis of survival data*. Chapman and Hall/CRC, 2018.

[34] Judith D Singer and John B Willett. It's about time: Using discrete-time survival analysis to study duration and the timing of events. *Journal of educational statistics*, 18(2):155–195, 1993.

[35] Chae Woo Nam, Tong Suk Kim, Nam Jung Park, and Hoe Kyung Lee. Bankruptcy prediction using a discrete-time duration model incorporating temporal and macroeconomic dependencies. *Journal of Forecasting*, 27(6):493–506, 2008.

[36] Kelly J Tiller, Shiferaw T Feleke, and Jane H Starnes. A discrete-time hazard analysis of the exit of burley tobacco growers in tennessee, north carolina, and virginia. *Agricultural Economics*, 41(5):397–408, 2010.

[37] Luming Zhang, Tao Huang, Fengshuo Xu, Shaojin Li, Shuai Zheng, Jun Lyu, and Haiyan Yin. Prediction of prognosis in elderly patients with sepsis based on machine learning (random survival forest). *BMC Emergency Medicine*, 22(26), 2022.

[38] Dean Fantazzini and Silvia Figini. Random survival forests models for sme credit risk measurement. *Methodology and Computing in Applied Probability*, 11(1):29–45, 2009.

[39] Sitti Nurhaliza, Kusman Sadik, and Asep Saefuddin. A comparison of cox proportional hazard and random survival forest models in predicting churn of the telecommunication industry customer. *BAREKENG: Journal of Mathematics and Its Application*, 16(4):1433–1440, 2022.

[40] John A Hartigan and Manchek A Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)*, 28(1):100–108, 1979.

[41] Zhexue Huang. Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data mining and knowledge discovery*, 2(3):283–304, 1998.

[42] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pages 226–231, 1996.

[43] scikit-learn contributors. scikit-learn: Dbscan clustering algorithm. `https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html`. Accessed: February 29, 2024.

[44] Peter Rousseeuw. Rousseeuw, p.j.: Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. comput. appl. math. 20, 53-65. *Journal of Computational and Applied Mathematics*, 20:53–65, 11 1987.

[45] David Davies and Don Bouldin. A cluster separation measure. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, PAMI-1:224 – 227, 05 1979.

[46] Tadeusz Caliński and Harabasz JA. A dendrite method for cluster analysis. *Communications in Statistics - Theory and Methods*, 3:1–27, 01 1974.

[47] Nicholas G. Polson and James G. Scott. On the half-cauchy prior for a global scale parameter. *Bayesian Analysis*, 7(4), 2012.

[48] Bob Carpenter. Prior choice recommendations, 2023.

[49] Andrew Gelman, John Carlin, Hal Stern, David Dunson, Aki Vehtari, and Donald Rubin. *Bayesian Data Analysis*. CRC Press, 3 edition, 2013.

[50] Paul D Allison. Discrete-time methods for the analysis of event histories. *Sociological methodology*, 13:61–98, 1982.

[51] Jung Eun Lee and Jessie H Chen-Yu. Effects of price discount on consumers' perceptions of savings, quality, and value for apparel products: mediating effect of price discount affect. *Fashion and Textiles*, 5:1–21, 2018.

[52] James W Vaupel and Anatoli I Yashin. Heterogeneity's ruses: some surprising effects of selection on population dynamics. *The American Statistician*, 39(3):176–185, 1985.

[53] Clara-Cecilie Günther, Ingunn Fride Tvete, Kjersti Aas, Geir Inge Sandnes, and Ørnulf Borgan. Modelling and predicting customer churn from an insurance company. *Scandinavian Actuarial Journal*, 2014(1):58–71, 2014.

[54] Humza Haider, Bret Hoehn, Sarah Davis, and Russell Greiner. Effective ways to build and evaluate individual survival distributions, 2018.

[55] Patrick J Heagerty, Thomas Lumley, and Margaret S Pepe. Time-dependent roc curves for censored survival data and a diagnostic marker. *Biometrics*, 56(2):337–344, 2000.

[56] Bob Carpenter, Andrew Gelman, Matthew D. Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1), 2017.

[57] Philip Hougaard. Frailty models for survival data. *Lifetime data analysis*, 1:255–273, 1995.

8 Appendix

8.1 Tables and Figures

Table 3: Description of Variables in the Dataset

Variable Name	**Data Type**	**Time Dependency**
policy_nr_hashed	Categorical Variable	Time-Independent
years_since_policy_started	Numerical Variable	Time-Independent
year_initiation_policy	Numerical Variable	Time-Varying
year_initiation_policy_version	Numerical Variable	Time-Varying
year_end_policy	Numerical Variable	Time-Varying
d_churn	Dummy Variable	Time-Varying
d_churn_cancellation	Dummy Variable	Time-Varying
d_churn_between_prolongations	Dummy Variable	Time-Varying
d_churn_around_prolongation	Dummy Variable	Time-Varying
premium_main_coverages	Numerical Variable	Time-Varying
premium_supplementary_coverages	Numerical Variable	Time-Varying
total_premium	Numerical Variable	Time-Varying
welcome_discount	Numerical Variable	Time-Varying
mutation	Numerical Variable	Time-Varying
premium_change_mutation	Numerical Variable	Time-Varying
allrisk basis	Numerical Variable	Time-Varying
allrisk compleet	Numerical Variable	Time-Varying
allrisk royaal	Numerical Variable	Time-Varying
wa-extra	Numerical Variable	Time-Varying
wettelijke aansprakelijkheid	Numerical Variable	Time-Varying
n_main_coverages	Numerical Variable	Time-Varying
n_supplementary_coverages	Numerical Variable	Time-Varying
n_coverages	Numerical Variable	Time-Varying
product	Categorical Variable	Time-Independent
sales_channel	Categorical Variable	Time-Independent
accident_free_years	Numerical Variable	Time-Varying
car_value	Numerical Variable	Time-Varying
age_car	Numerical Variable	Time-Varying
weight	Numerical Variable	Time-Varying
customer_age	Numerical Variable	Time-Varying
brand	Categorical Variable	Time-Independent
type	Categorical Variable	Time-Independent
fuel_type	Categorical Variable	Time-Independent
postcode	Categorical Variable	Time-Independent
welcome_discount_control_group	Categorical Variable	Time-Independent

Table 4: Variables Containing Missing Values

Variable	**Type**	**Number**
premium_main_coverages	Nan & 0	17282
premium_supplementary_coverages	Nan	3813
total_premium	0	4439
d_churn	inaccurate	6930

Table 5: will be in Appendix: example of person-period data set

Policy Number	D_1	D_2	D_3	D_4	D_5	D_6	x_{pi}	z_{kis}	y_{is}
01	1	0	0	0	0	0	0	1	1
02	1	0	0	0	0	0	1	0	0
02	0	1	0	0	0	0	1	0	0
02	0	0	1	0	0	0	1	1	1
03	1	0	0	0	0	0	1	0	0
03	0	1	0	0	0	0	1	0	0
03	0	0	1	0	0	0	1	0	0
03	0	0	0	1	0	0	1	0	0

Table 6: Considered variables for clustering

Considered Variables for Clustering
customer_age
accident_free_years
car_value
age_car
weight
allrisk_basis
allrisk_compleet
allrisk_royaal
wa-extra
n_supplementary_coverages
n_coverages
brand_label
type_label
fuel_type_label
product_label
sales_channel_label
postcode_label

Table 7: Fuel Type Distribution by Cluster

	Fuel Type Distribution					
	Diesel	Benzine	Electro	Gas	Hybrid	Other
Cluster 0	0.738	0.200	0.055	0.004	0.002	0.000
Cluster 1	0.903	0.092	0.003	0.002	0.000	0.000
All observations	0.835	0.136	0.024	0.002	0.002	0.000

Table 8: Clustering evaluation outcome

Clustering Method	Silhouette Score	Davies-Bouldin Index	Calinski-Harabasz Index
K-prototype (2 clusters)	0.014	11.585	522.616
K-prototype (3 clusters)	-0.021	19.119	350.239
K-prototype (4 clusters)	-0.040	34.544	271.261
K-prototype (5 clusters)	-0.048	76.250	288.296
K-prototype (6 clusters)	-0.064	34.876	279.835
K-prototype (7 clusters)	-0.069	112.485	233.727
DBSCAN (3 clusters)	-0.018	25.735	161.453
DBSCAN (4 clusters)	-0.046	28.982	324.678

Table 9: Mean numerical variable values per cluster

Variable	**Cluster 0**	**Cluster 1**
customer_age	55.63	42.67
accident_free_years	16.22	6.04
car_value	538896.15	225468.11
age_car	6.87	12.93
weight	1457.85	1055.82
allrisk_basis	0.08	0.04
allrisk_compleet	0.42	0.17
allrisk_royaal	0.22	0.06
wa-extra	0.20	0.40
n_supplementary_coverages	0.93	0.76

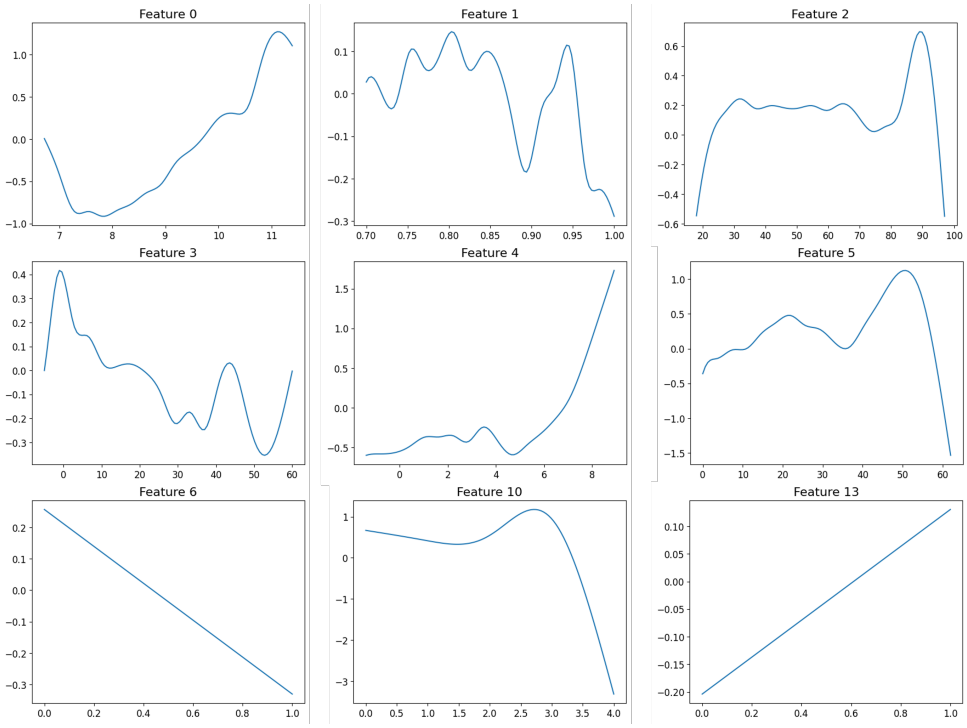


Figure 12: The marginal relationship shown by GAM model

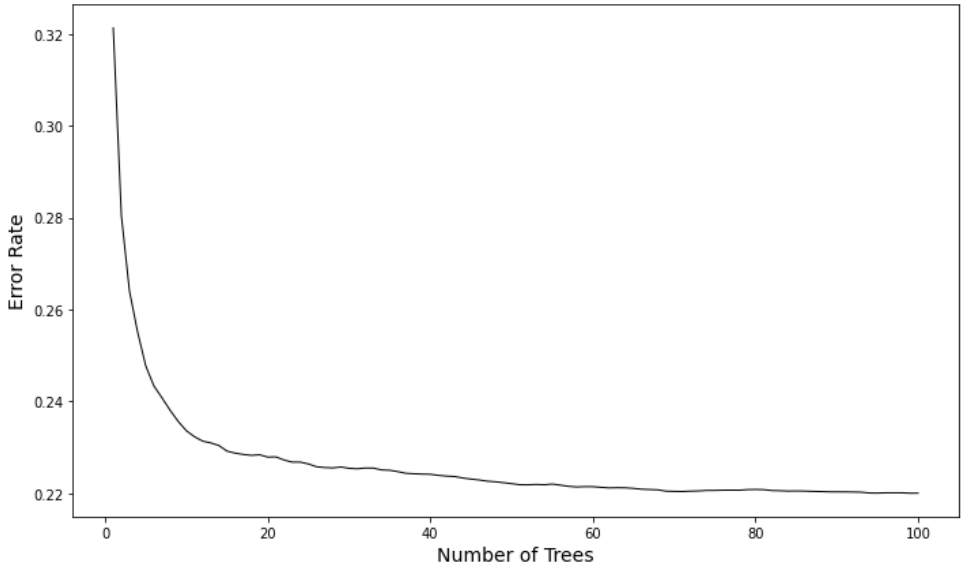


Figure 13: error rate vs number of trees RSF model

Table 10: Logistic Regression Coefficients

Variable	**Coefficient**	**Std. Err.**	**z**	**P> |z|**	**[0.025**	**0.975]**
total premium	0.5401	0.028	18.994	0.000	0.484	0.596
welcome discount	-0.4789	0.166	-2.890	0.004	-0.804	-0.154
customer age	0.0008	0.001	0.946	0.344	-0.001	0.002
accident free years	-0.0177	0.001	-12.048	0.000	-0.021	-0.015
car value	0.0920	0.013	7.245	0.000	0.067	0.117
age car	0.0282	0.002	12.556	0.000	0.024	0.033
allrisk basis	-0.6659	0.053	-12.582	0.000	-0.770	-0.562
allrisk compleet	-1.0469	0.036	-29.483	0.000	-1.117	-0.977
allrisk royaal	-1.2335	0.047	-26.369	0.000	-1.325	-1.142
wa-extra	-0.5685	0.027	-20.818	0.000	-0.622	-0.515
n supplementary coverages	-0.0914	0.018	-5.004	0.000	-0.127	-0.056
WD eligible	-0.1053	0.028	-3.811	0.000	-0.159	-0.051
LPA eligible	-0.3693	0.030	-12.496	0.000	-0.427	-0.311
years since policy started 0	-5.9432	0.498	-11.927	0.000	-6.920	-4.966
years since policy started 1	-5.6280	0.498	-11.293	0.000	-6.605	-4.651
years since policy started 2	-5.7822	0.499	-11.596	0.000	-6.759	-4.805
years since policy started 3	-5.7536	0.500	-11.518	0.000	-6.733	-4.775
years since policy started 4	-5.8984	0.503	-11.728	0.000	-6.884	-4.913
years since policy started 5	-8.0449	0.871	-9.240	0.000	-9.751	-6.338
cluster 1.0	-0.0330	0.033	-0.999	0.318	-0.098	0.032
fuel type benzine	-0.1602	0.445	-0.360	0.719	-1.033	0.713
fuel type diesel	0.1275	0.446	0.286	0.775	-0.747	1.002
fuel type electro	-0.1662	0.451	-0.369	0.712	-1.049	0.717
fuel type gas	0.2361	0.478	0.493	0.622	-0.702	1.174
fuel type hybride	-0.1029	0.489	-0.210	0.833	-1.061	0.856
sales channel AEGON.nl	0.3485	0.039	8.851	0.000	0.271	0.426
sales channel Independer	0.6954	0.069	10.032	0.000	0.560	0.831
WD total premium	0.2516	0.041	6.149	0.000	0.171	0.332
WD accident free years	-0.0158	0.003	-5.946	0.000	-0.021	-0.011
WD car value	0.0614	0.025	2.492	0.013	0.013	0.110
WD customer age	-0.0022	0.001	-1.867	0.062	-0.004	0.000
WD age car	0.0361	0.003	10.593	0.000	0.029	0.043
WD n supplementary coverages	-0.1182	0.022	-5.266	0.000	-0.162	-0.074
WD WD eligible	1.0413	0.370	2.813	0.005	0.316	1.767
WD LPA eligible	-0.0661	0.038	-1.719	0.086	-0.142	0.009
WD cluster 1.0	-0.0106	0.053	-0.200	0.842	-0.115	0.094
WD fuel type benzine	1.3689	0.640	2.140	0.032	0.115	2.622
WD fuel type diesel	1.7690	0.643	2.750	0.006	0.508	3.030
WD fuel type electro	1.6329	0.645	2.533	0.011	0.370	2.896
WD fuel type gas	0.7590	0.801	0.947	0.344	-0.811	2.330
WD fuel type hybride	1.4883	0.744	2.001	0.045	0.031	2.946
WD sales channel AEGON.nl	0.3759	0.040	9.399	0.000	0.298	0.454
WD sales channel Independer	0.5396	0.035	15.472	0.000	0.471	0.608
WD allrisk basis	-0.5454	0.062	-8.734	0.000	-0.668	-0.423
WD allrisk compleet	-0.5828	0.056	-10.353	0.000	-0.693	-0.472
WD allrisk royaal	-0.5862	0.076	-7.671	0.000	-0.736	-0.436
WD wa-extra	-0.3677	0.036	-10.118	0.000	-0.439	-0.296

8.2 Mathematical Details

Silhouette Score

The Silhouette Score is calculated as follows:

$$\text{Silhouette Score} = \frac{1}{N}\sum_{i=1}^{N}\frac{b(i)-a(i)}{\max\{a(i),b(i)\}}$$

Where:

Davies-Bouldin Index

The Davies-Bouldin Index (DBI) is calculated as follows:

$$DBI = \frac{1}{n}\sum_{i=1}^{C}\max_{j\neq i}\left(\frac{\text{avg\_d}_i + \text{avg\_d}_j}{d(i,j)}\right)$$

Where:

C is the number of clusters

$d(i,j)$ is the distance between cluster centers i and j

avg_d_i is the average distance of each point in cluster i to the centroid of cluster i

Calinski-Harabasz Index

The Calinski-Harabasz Index (CHI) is calculated as follows:

$$\text{Calinski-Harabasz Index} = \frac{Tr(B_k)}{Tr(W_k)} \times \frac{N-k}{k-1}$$

Where:

$Tr(B_k)$ is the trace of the between-cluster dispersion matrix

$Tr(W_k)$ is the trace of the within-cluster dispersion matrix

k is the number of clusters.

$$\begin{aligned} l &= \sum_{i=1}^{n} c_i \ln\left(\frac{h_{is}}{1-h_{is}}\right) + \sum_{s=1}^{s_i}\ln(1-h_{is}) \\ &= \sum_{i=1}^{n}\left[\sum_{s=1}^{s_i} y_{is}\ln\left(\frac{h_{is}}{1-h_{is}}\right) + \sum_{s=1}^{s_i}\ln(1-h_{is})\right] \\ &= \sum_{i=1}^{n}\sum_{s=1}^{s_i}\left[y_{is}\ln\left(\frac{h_{is}}{1-h_{is}}\right) + \ln(1-h_{is})\right] \end{aligned} \tag{9}$$

Algorithm 1 DBSCAN Algorithm

Input: Dataset D, ε, $MinPts$
Output: Clusters $C_1, C_2, ..., C_k$
function DBSCAN($D, \varepsilon, MinPts$)
    Initialize an empty set C
    Initialize all points as unvisited
    $k \leftarrow 0$
    for all points p in D **do**
        if p is visited **then**
            continue
        end if
        Mark p as visited
        $N \leftarrow$ getNeighbors(p, ε)
        if $|N| < MinPts$ **then**
            Mark p as noise
        else
            $C_k \leftarrow \{p\}$ expandCluster($p, N, C_k, \varepsilon, MinPts$)
            $k \leftarrow k + 1$
        end if
    end for
end function
function expandCluster($p, N, C_k, \varepsilon, MinPts$)
    for all points q in N **do**
        if q is not visited **then**
            Mark q as visited
            $N' \leftarrow$ getNeighbors(q, ε)
            if $|N'| \geq MinPts$ **then**
                $N \leftarrow N \cup N'$
            end if
        end if
        if q is not yet member of any cluster **then**
            Add q to C_k
        end if
    end for
end function
function getNeighbors(p, ε)
    Initialize an empty set N
    for all points q in D **do**
        if distance(p, q) $< \varepsilon$ **then**
            Add q to N
        end if
    end for
    return N
end function

Algorithm 2 K-Prototypes Algorithm

Input: Dataset D, number of clusters k, categorical weight λ
Output: Clusters $C_1, C_2, ..., C_k$
function KPROTOTYPES(D, k, λ)
    Initialize k random prototypes for both numerical and categorical attributes
    repeat
        Assign each data point to the nearest prototype
        Update the numerical prototypes by computing the mean of each cluster
        Update the categorical prototypes by selecting the mode of each cluster
    until Convergence
end function