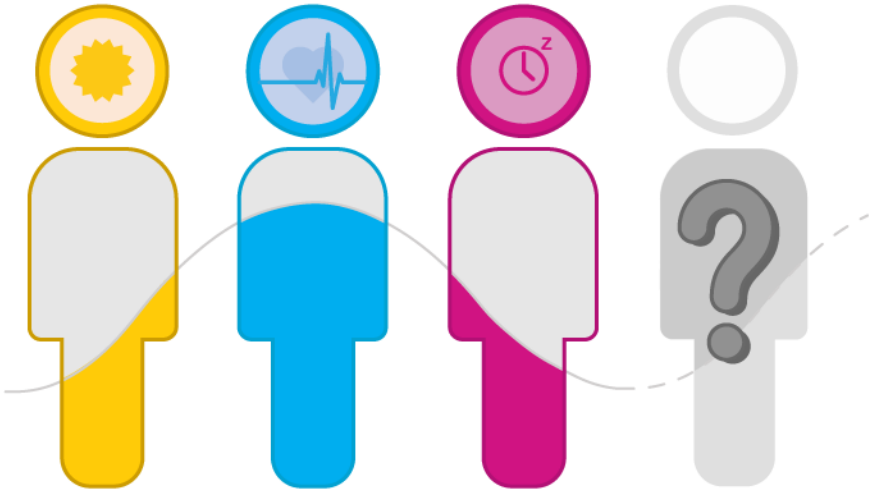
EPSRC
Engineering and Physical Sciences Research Council

InFoMM
Industrially Focused Mathematical Modelling

EPSRC Centre for Doctoral Training in Industrially Focused Mathematical Modelling



User Cancellation Modelling: on Clustering of Customer Behaviours

Victor (Sheng) Wang

UNIVERSITY OF OXFORD

Maths-Whizz by whizz EDUCATION

Contents

1. Introduction

Motivation and Goal

Retained customers in general create higher revenues than new customers do, and making a sell to a new customer can cost up to 5 times more depending on the business. Therefore, many companies form the Customer Relationship Management (CRM) team with a focus on customer retention strategies. A crucial step is then to identify high risk customers who are intending to discontinue their usage of the services. This assessment is better known as *churn prediction*.

Churn prediction helps the business to detect customers who are likely to cancel a subscription.

Our project aims to perform the churn prediction task based on making probabilistic clustering assignment of customers' behaviours, and formulate the sequential processes into a scalable pipeline which can be easily reused, updated and extended for many applications. In particular, we apply the pipeline to analyse pupil subscribers' data for Whizz Education (referred to as "Whizz"). Whizz provides online virtual tutorial service, Math-Whizz, which pupils can access by purchasing monthly subscriptions. At the end of live subscription, pupils can make the choice to cancel or do nothing to renew a new 1-month subscription.

Train and Prediction Workflow

Whizz can use the 2-step workflow proposed in Figure 1.1 to predict pupils' probability of churn, or churn risk, which provides the downstream CRM team with target subscribers to apply retention strategies more efficiently.

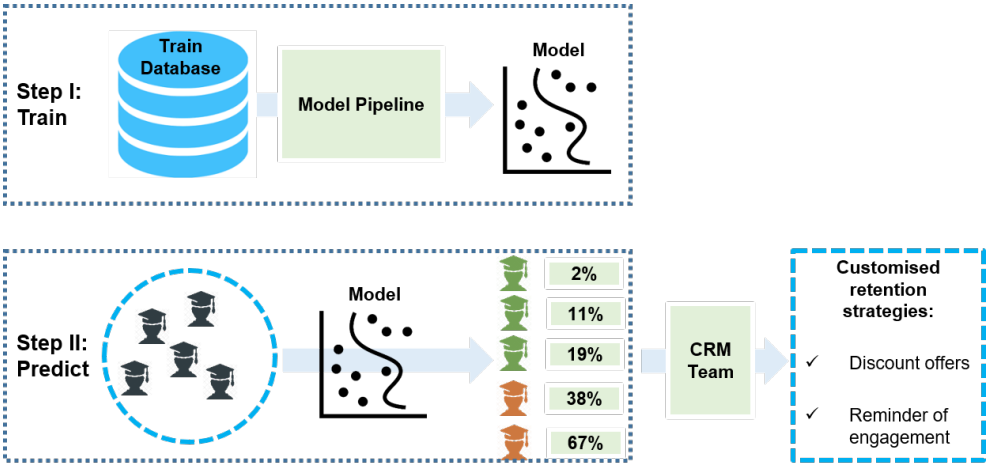


Figure 1.1 – Train and prediction workflow, and its downstream process.

Glossary of Terms

- **Clustering:** The task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar to each other than to those in other groups.
- **Mixture model:** A probabilistic model for representing the presence of a mixture of subpopulations within an overall population.
- **Distribution:** A mathematical function that provides the probability of occurrence of different possible values of a variable.
- **Markov chain:** A model of random event (called a state) that happens over time. The probability of each event depends only on the previous event (Markov property).

2. A Generative Process for Customer Behaviours

Customers' behaviour evolves over time as they respond to business offering and adjust to their own demand. We employ the Markov chain to describe the generative process of customers' behaviours, where the. In addition, we use the mixture model to find Markov states, each of which defines a partition of the behaviours within a single time interval. Splitting pupils' behaviours into monthly time periods makes most sense provided the business settings at Whizz. Pupils subscribe to Whizz products on a 1-month contract, and make the choice to cancel at the end of current subscription. Otherwise a renewal will be made by default.

Customer Journey

Customer journeys reflect the dynamics of their monthly-behaviours over time. There is inconsistency present in journeys of different customers. The inconsistency refers to the problem that the time intervals for different customers being alive in the services are not aligned, so that their behaviours are not comparable. To resolve the inconsistency, we align and aggregate customers' behaviours by switching the reference from calendar month to customer month. This is illustrated by an example in Figure 2.1.

Customer month is a time period between the start and the end of a monthly subscription.

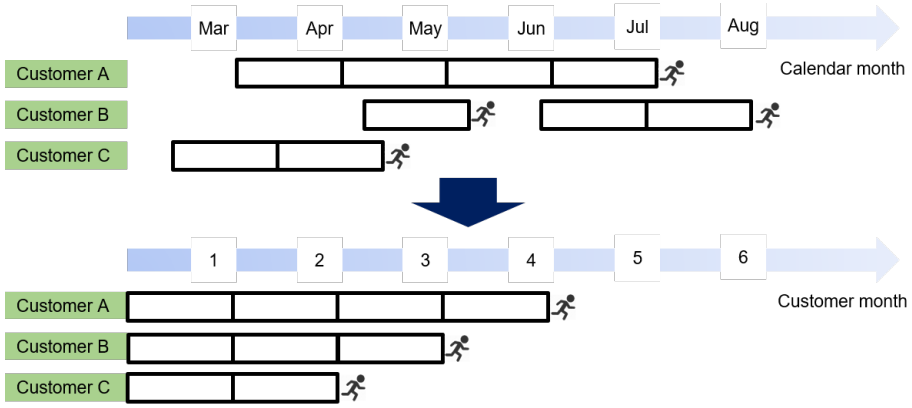


Figure 2.1 – Change reference from calendar month to customer month. Customer A, B and C have very different journeys in the sense of subscription start and end dates. Each block represents customer's monthly behaviours. Under calendar month reference, we have to choose studying months from March to August to cover all activities. This choice results in irregular temporal distribution of missing information for all 3 customers. After changing the reference to customer month, behaviours are aligned by customer month and therefore comparable. Moreover, the missing information only occurs after the customer churns. It can also handle discontinuous subscriptions like the case of customer B.

Behaviour, Cluster and State

We assume that customers with intentions to churn exhibit different behaviours than others do. Behaviours are different distributionally, and generated from a finite number of states. Then the behaviour dynamics of each customer results from a chain of states over time. This formulates into a discrete-time Markov chain with states *transiting* over time, where each state *emits* distinguishable behaviour distribution. An example is given in Figure 2.2. In brief, a Markov chain is a random process characterised by transition and emission probabilities.

An observation is generated as: state → cluster → behaviour.

The distribution of the emitted behaviours from a state is assumed to be a mixture of component distributions such as uniform, Gaussian, etc. Each component represents a cluster. Therefore, the sequence of observed behaviours are generated in the following way:

1. At each time step, the system generates a state according to the state-to-state transition.
2. Once the state has been generated, the system generates a cluster according to the state-to-cluster emission probability.
3. Once the cluster has been determined, a behaviour observation is produced probabilistically according to the cluster-parametrised distribution.

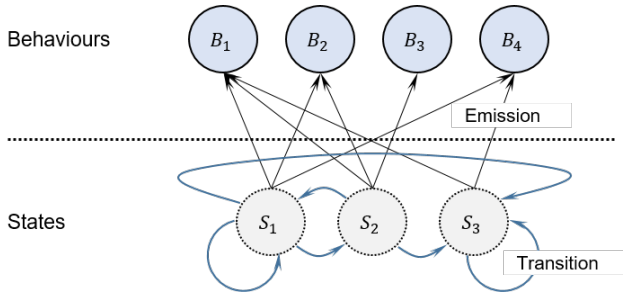


Figure 2.2 – Behaviours emitted from Markov states. States S_1, S_2 and S_3 generate differently distributed behaviours. For example, S_1 generates $\{B_1, B_2, B_4\}$ while S_3 produces $\{B_1, B_4\}$. Even if the two states can generate the same set of behaviours, the emission probabilities can be different, thus still resulting in different behaviour distributions. States transit between each other over time randomly.

3. Probabilistic Clustering Using Mixture Model

A critical step of the state-cluster-observation Markov chain is the mixture model that describes the probabilistic assignment of observations to clusters. We choose specifically the *Dirichlet Process Mixture* setting which has two important features:

1. It is Bayesian and treats component parameters as random variables, of which posterior distributions will be updated from a prior as observed data coming in.
2. The Dirichlet process (DP) is used as a nonparametric prior resulting in that the number of clusters is random and grows as new data are observed.

Dirichelet process mixture is Bayesian and can infer the number of clusters from data.

The benefits of this model choice are massive. It does not view the observed data as infinitely available as independent replicates like frequentists, so that it can be updated with new data coming in. Moreover, it infers the number of clusters from observed data, and opens the opportunities of finding new clusters as more data are observed.

4. Modelling Pipeline

Scope of Data and Features

Whizz stores and maintains data generated from business activities in a database consisting of several relational tables. The most relevant tables we use are listed in Table 4.1. The study period is from January 1st, 2014 to April 20th, 2018. The number of records within this study period in each data table is also indicated.

Data Table	Description	Number of Records
Account Information	Pupils' ID and personal information such as date-of-birth.	2,672
Subscription history	Start date and end date of each new subscription or renewed associated with each pupil.	17,861
Lesson history	Details of each visit activity for each pupil during his subscription period. The visit activity includes the date of visit, time spent, score achieved, lesson outcome, etc.	450,548

Table 4.1 – Description of data tables. The number of records within the study period 2014-01-01 ~ 2018-04-20 in each data table is also indicated.

A *feature* is an individual measurable property of a behaviour being observed, and choosing informative features is crucial for effective clustering. For example, to measure how often the pupil uses Whizz online tutorial, we can define the time spent or number of visits within a month as the feature. For each customer, we define a collection of features to capture his behaviours of usage, progress, etc. within a customer month.

Modelling Pipeline

We summarise the processes of our modelling framework as a pipeline displayed in Table 4.2. We will apply this pipeline to build a churn prediction model for Whizz.

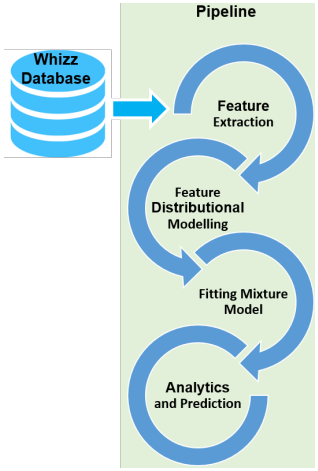


Figure 4.1 – Modelling Pipeline.

No.	Process	Input	Output
1	**Feature extraction**: extract informative feature data from historical records of pupils' activity, and represent them in the suitable data structure that can be fed into the behavioral model.	Raw data from Whizz database that records pupils' ID, subscription history, activity history, etc.	Feature data.
2	**Feature distributional modelling**: choose the most suitable distributions for each feature to fit. Independent features can be modelled separately, while correlated features shall be modelled as a part of a multivariate distribution.	Feature data.	Distributional form for each feature.
3.1	**Fitting mixture model**: by assuming behaviours are generated from a Dirichelet process mixtures, make inference on parameters based on observed features.	Feature data, feature distributional form, mixture model.	A collection of clusters.
3.2	**Fitting Markov chain**: use the identified clusters as well as churn outcome information to define states; uncover their transition probabilities.	Identified probabilistic clusters with distributional form; churn outcome.	A finite set of states and state-to-state transition probabilities.
4.1	**Analytics on behaviours**: study the properties of pupils' behaviours such as the temporal transition probabilities, how each feature impact the level of churn risk, etc.	States, transition, cluster churn rate etc.	Temporal state transition analysis, feature analysis, etc.
4.2	**Prediction on new pupils**: predict the level of churn risk for new pupils based on their behaviours with the observation-cluster-state probabilistic assignment trained from previous steps.	States, clusters, cluster-state assignment.	State assignment for new pupils as well as their associated level of churn risk.

Table 4.2 – Modelling pipeline showing processes in sequence, along with the input and output of each modular process. The last two process 4.1 and 4.2 are independent from each other and are performed for different purposes.

5. Clustering Analytics

We assume that most features can be described by mixture of multivariate Gaussians. For those whose empirical distribution deviates too much from a Gaussian, we verify their independence from other features and model them as a general mixture of any suitable simple distribution such as uniform, exponential, etc. Examples are shown in Figure 5.1.

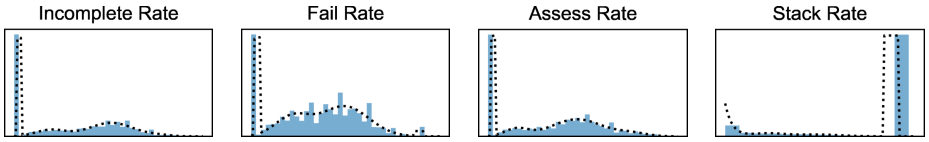


Figure 5.1 – Fitting bespoke mixture model for independent features.

Cluster Churn Rate

We combine the clusters inferred from independent and multivariate features together. For each identified cluster, because we know the churn outcome, we can calculate the proportion of churners, called the *cluster churn rate*. In Figure 5.2 we show the average cluster size and the associated churn rate. By "average", we calculate the average size for clusters of the same churn rate.

The model identifies a couple of clusters with 100% and 0% churn rate. However, the cluster size is really small for those extreme high/low churn rate. This implies that it is much easier to identify "normal" customers where the churn probability is around population average, but really challenging to distinguish "risky" or "safe" customers who have very high or low chance to churn.

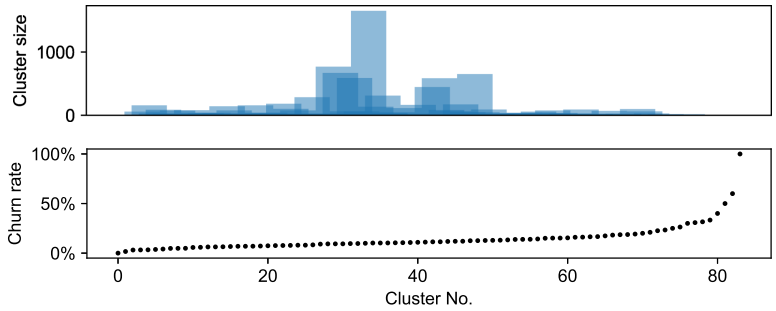


Figure 5.2 – Identified clusters, their average size and associated churn rate.

Churn Probability and Markov State

Because we have assumed that each pupil has independent behaviours from others, and that customer month are independent in leading to churn outcome, we can interpret the cluster churn rate as the probability of churn of each individual pupil in that cluster. This allows us to have the distribution of individual churn probabilities, shown in left of Figure 5.3.

Looking at the distribution of churn probabilities, we can observe a few concentrations of pupils centered at churn probabilities approximately 0%, 12%, 25%. If the behaviour based model does not work at all, the inferred churn probabilities will be purely random, which will converge to a normal distribution due to central limit theorem. Distributions shown in Figure 5.3 are definitely not normal. This has shown that the behaviour based model has the predictive power to distinguish pupils with very different churn probabilities. In other words, pupils with the intention to churn exhibit different behaviours than those otherwise.

The behaviour based model has the predictive power to distinguish pupils with very different churn probabilities!

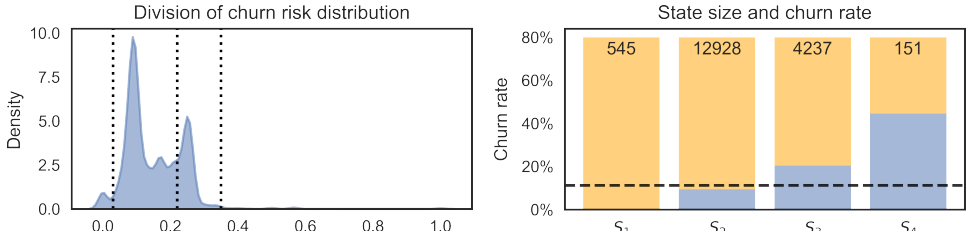


Figure 5.3 – Defining Markov states by grouping together pupils of similar levels of churn risk. The right plot shows the size of each state as well as churn rate. The dashed line indicates the level of population churn rate.

We define the Markov states by grouping together clusters of similar churn rates because we want the states to indicate churn risk level. In right of Figure 5.3 we show an example. We choose to form 4 states S_1, S_2, S_3 and S_4 by grouping clusters according to 3 anchor points: 0.3, 0.22, 0.35 (the three vertical dashed line shown on the left plot). The state churn rates are 0.0%, 11.7%, 25.4% and 55.6% respectively. Therefore, we can say S_1 is the safest state where pupils in this state are very unlikely to churn. In contrast, S_4 is a risky state in the sense that pupils in this state are 5.5 times more likely than average to cancel the subscription.

Transitional Analysis

Following the state definitions in the previous section, we can move to the transitional analysis. We assume stationarity of the Markov chain and estimate the transition probabilities by maximum likelihood. Note that we need to add a state S_{churn} that once entered cannot be left (called an absorbing state). Hence, the complete state set is $\{S_1, S_2, S_3, S_4, S_{\text{churn}}\}$.

The transition probabilities are visualised in Figure 5.4 as a Sankey diagram. Churned pupils (those in S_{churn}) can not transit to other states. S_2 is the largest destination state, which makes sense as S_2 represents a "normal" state in which the churn probability is approximately the

same as population average. Barely pupils transit from other states to S_1 (the safest state) or S_4 (the riskiest state). This seems imply two behavioural paths. First, if a pupil is very likely to stay live in the subscription (in S_1), then his intention of exit will be gradually accumulated since most likely he will transit to S_2 in the next period and S_2 again the next. Second, the reason for strong intention of churn might be purely external and irrelevant to customer experience at Whizz, since there is little chance of transiting to S_4.

State transition probabilities help us understand behaviour dynamics.

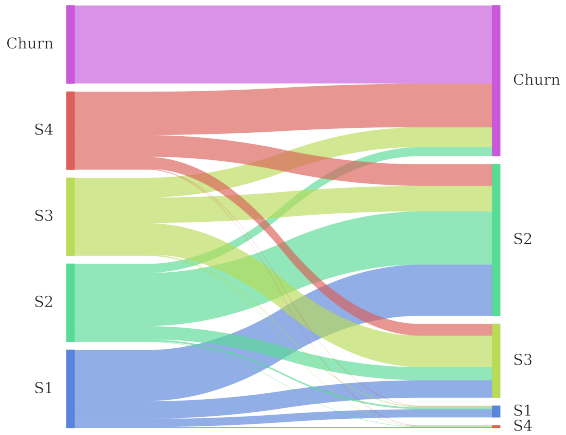


Figure 5.4 – Sankey diagram showing the transition probabilities between states. The thickness of each band represents the magnitude of the transition probability from the state on the left to the state on the right.

6. Discussion, conclusions & recommendations

To help predict churn risk of pupils from the subscription to Whizz online tutorial service, we have proposed a mixture based model which identifies behaviour clusters associated with distinguished levels of churn rate. Moreover, we fit this model into a Markov chain setting to allow us understand temporal customer behavioural path. The sequential modelling processes have been formulated into a scalable pipeline that can be easily reused, updated and extended.

We describe customer behaviours as a result of the state-cluster-observation generative process. The mixture based model can infer clusters from observed behaviours, and clusters will be grouped to form states by bespoke risk appetite. We have formed 4 states from Whizz's data where the "riskiest" state has a churn rate 5.5 time higher then the population average. The model trained from observed data can then be used to make prediction on the churn risk of active pupils, and also analyse the potential causes of such risk.

I'm very confident about this mixture based behavioural model approach in churn prediction. There's room for future work. First, we have not discussed in details about the feature selection and feature engineering, though the model itself can absorb any number of features. Selecting independent, informative features may greatly improve the clustering outcome and prediction accuracy. There is room to mine useful information from Whizz's pupils activity database. Second, it is possible to extend the analysis about how feature impacts churn risk. One direction is to find a way to compare feature importance, which translates into measuring importance of dimensions in a multivariate setting. This will be very useful to help the downstream CRM team to prioritise retention strategies.

Dr. Junaid Mubeen, Director of Education at Whizz Education said: "*We are very pleased with Victor's contribution. He has addressed our key requirement of developing a model that allows us to predict the likelihood of home users cancelling their subscription from one month to the next. Discussions are already underway with our Development and Marketing teams to understand how we can make use of Victor's model on a continual basis to turn around customers "at risk" of cancelling. Victor has also illuminated a number of Machine Learning strategies that we can see having applications in other business problems.*"