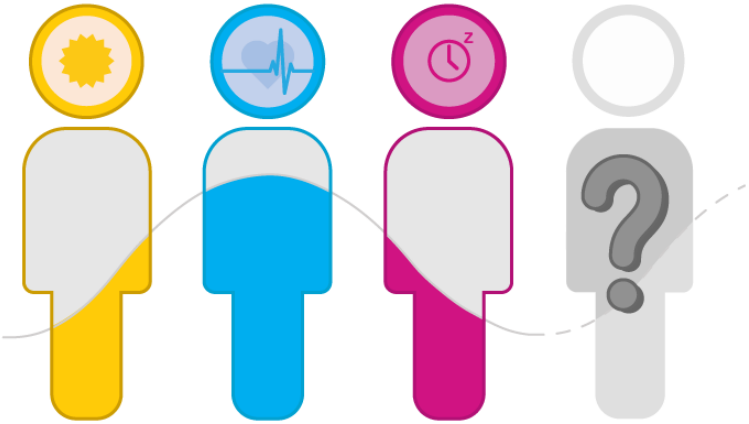
University of Oxford

Mathematical Institute

InFoMM
Industrially Focused Mathematical Modelling

User Cancellation Modelling:
on Clustering of Customer Behaviours



Victor Wang

Supervisors: Andrew Mellor, Junaid Mubeen

Agenda

- Motivation and Goal
- Background – Business Settings at Whizz
- Model
  - Customer Journey
  - Methodology
- Result
  - Clusters
  - Feature Analysis
- Conclusion

Motivation and Goal

Motivation

Retained customers create ***higher revenues***

Making a sell to a new customer ***costs up to 5 times more***.

Customer retention strategies targeting on ***high risk customers***

Motivation and Goal

InFoMM
Industrially Focused
Mathematical Modelling

University of Oxford

Mathematical Institute

Motivation

Retained customers create ***higher revenues***

Making a sell to a new customer ***costs up to 5 times more***.

Customer retention strategies targeting on ***high risk customers***

Goal – ***Churn*** (Cancellation) modelling

Who?

- ✓ Identify customers most prone to cancel subscription
- ✓ Assess likelihood

Why?

- ✓ Analyse critical reason triggering cancellation
- ✓ Make bespoke retention policy

University of Oxford

Mathematical Institute

Business Settings at Whizz

InFoMM
Industrially Focused Mathematical Modelling

Cancellation Mechanism

Contractual

- Pupils subscribe to access Whizz products on a 1-month or 1-year contract

Voluntary

- Subscribers make the choice to leave the service at the end of the subscription; otherwise auto-rolled

University of Oxford

Mathematical Institute

Business Settings at Whizz

InFoMM
Industrially Focused Mathematical Modelling

Cancellation Mechanism

Contractual

- Pupils subscribe to access Whizz products on a 1-month or 1-year contract

Voluntary

- Subscribers make the choice to leave the service at the end of the subscription; otherwise auto-rolled

Data Records

Time period: 2014-Jan-01 – 2018-Apr-20

Pupil

- Pupil ID
- Date of birth

✓ 5685 pupils

Subscription

- Pupil ID
- Start date
- End date

✓ 28025 subscriptions

Lesson History

- Lesson ID
- Pupil ID
- Time
- Mark
- …

✓ ~1.6 million lesson activities

Business Settings at Whizz

InFoMM
Industrially Focused Mathematical Modelling

Cancellation Mechanism

Contractual

- Pupils subscribe to access Whizz products on a 1-month or 1-year contract

Voluntary

- Subscribers make the choice to leave the service at the end of the subscription; otherwise auto-rolled

Data Records

Time period: 2014-Jan-01 – 2018-Apr-20

Pupil

- Pupil ID
- Date of birth

✓ 5685 pupils

Subscription

- Pupil ID
- Start date
- End date

✓ 28025 subscriptions

Lesson History

- Lesson ID
- Pupil ID
- Time
- Mark
- …

✓ ~1.6 million lesson activities

Features

$$X = \begin{bmatrix} 1.2 & 20 & \cdots & 80\% \\ 52.2 & 34 & \cdots & 31\% \\ \vdots & & \ddots & \vdots \\ 25.1 & 99 & \cdots & 43\% \end{bmatrix}$$

Usage time (hours)

Number of visits

Pass rate

Churn outcome (renew / cancel)

$$\begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix}$$

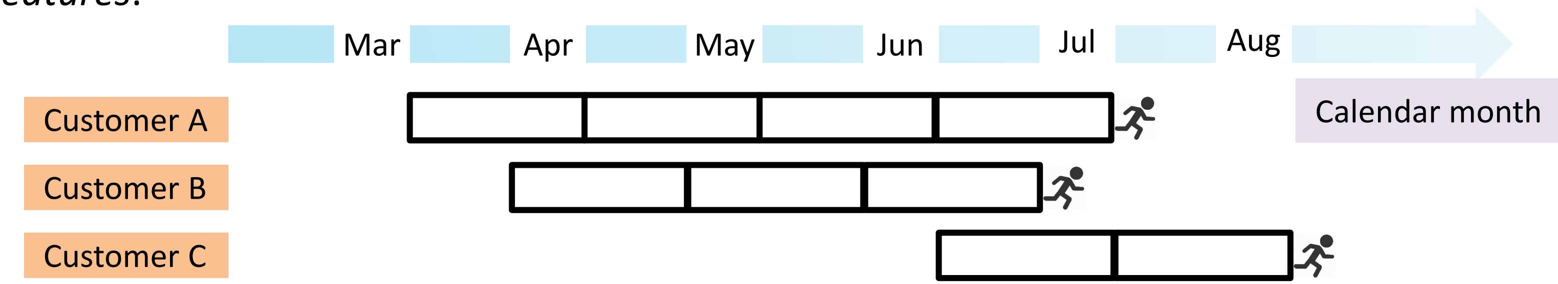
Renew

Cancel

Customer Journey

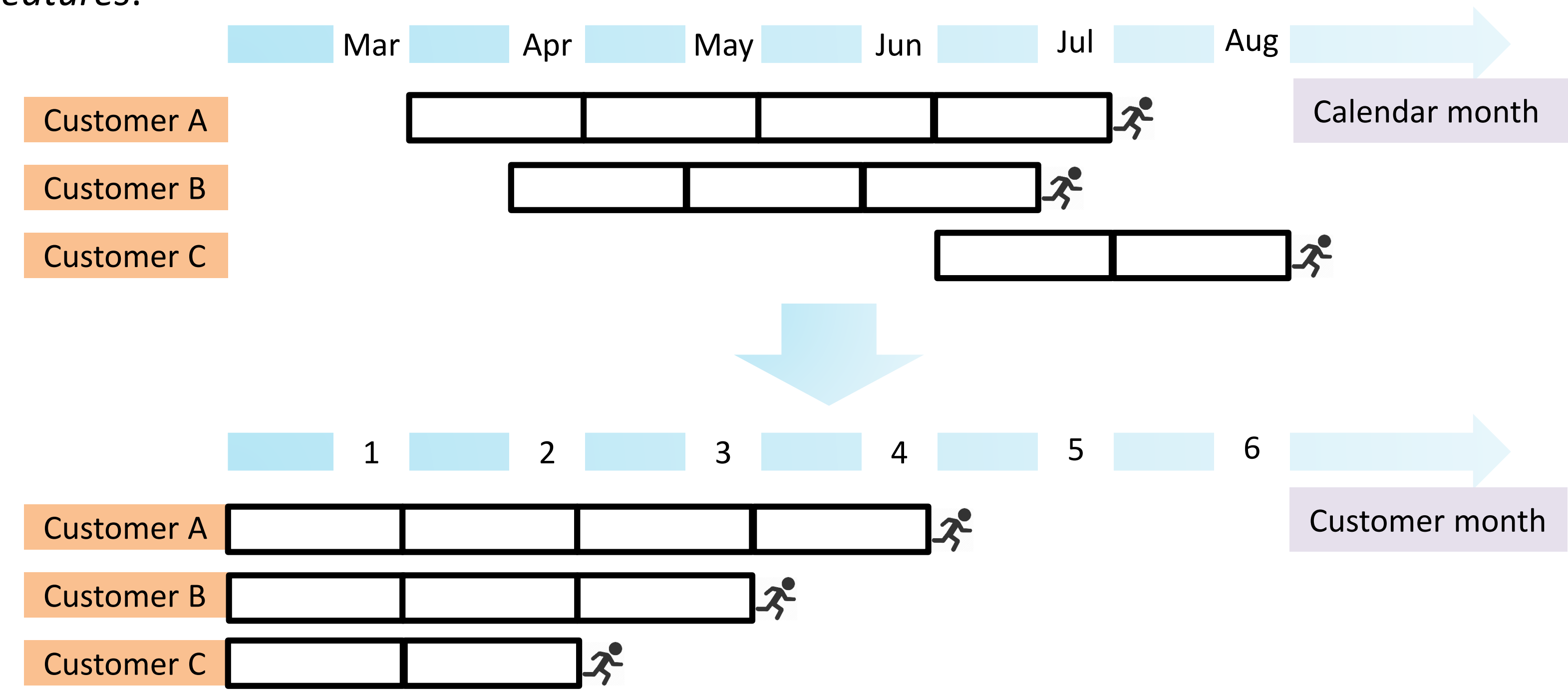
InFoMM
Industrially Focused Mathematical Modelling

Pupils' activities are split into monthly time periods with each **customer-month** represented by *features*.



Customer Journey

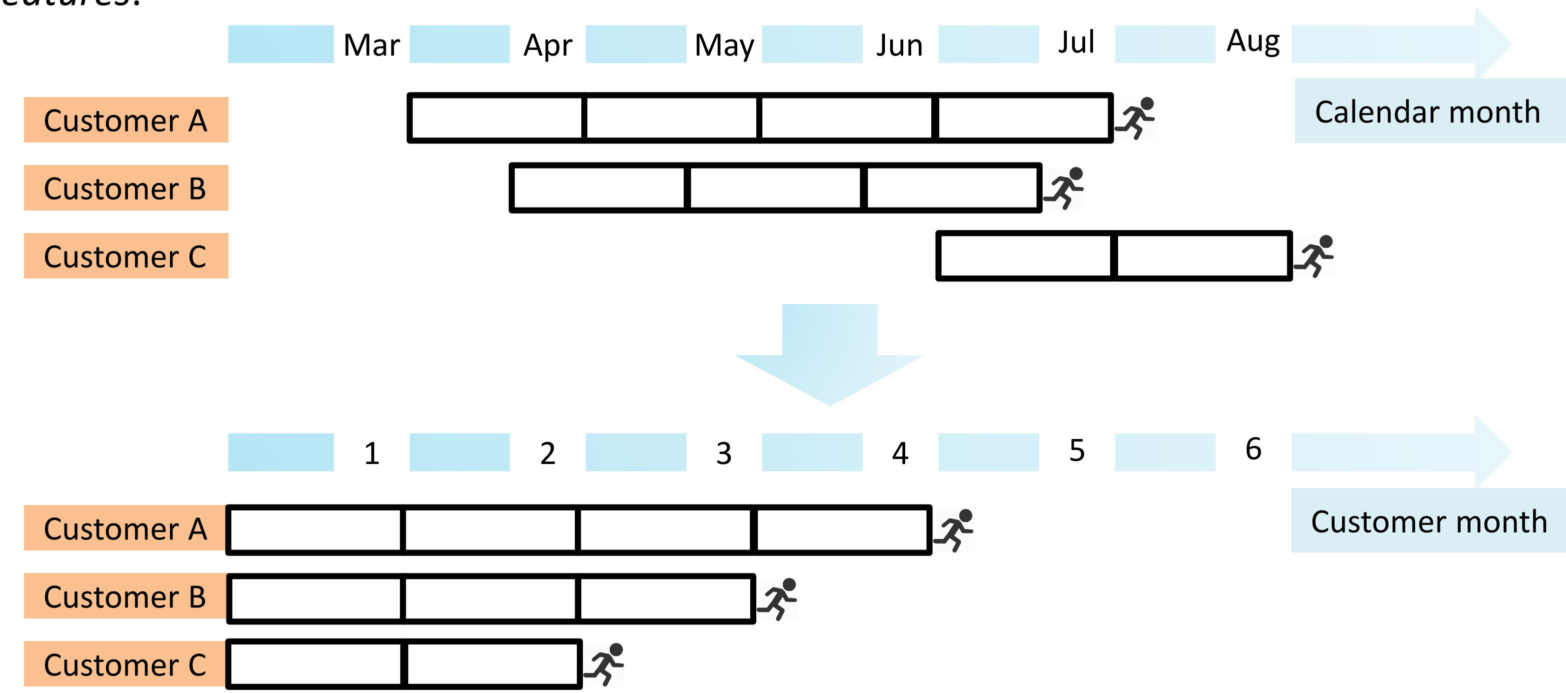
Pupils' activities are split into monthly time periods with each **customer-month** represented by *features*.



Customer Journey

InFoMM
Industrially Focused Mathematical Modelling

Pupils' activities are split into monthly time periods with each **customer-month** represented by *features*.



We assume the cancellation to be **ONLY** dependent on the current month.

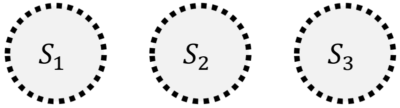
- ✓ The assumption holds since we observe statistically constant churn rate over customer month.
- ✓ The assumption enables us to treat activities in different customer months indifferently.

Methodology

Key Assumptions

✓ Churners and non-churners exhibit different behaviours generated by some **states**

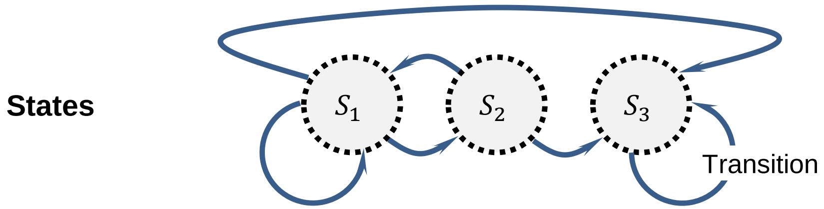
States



Methodology

Key Assumptions

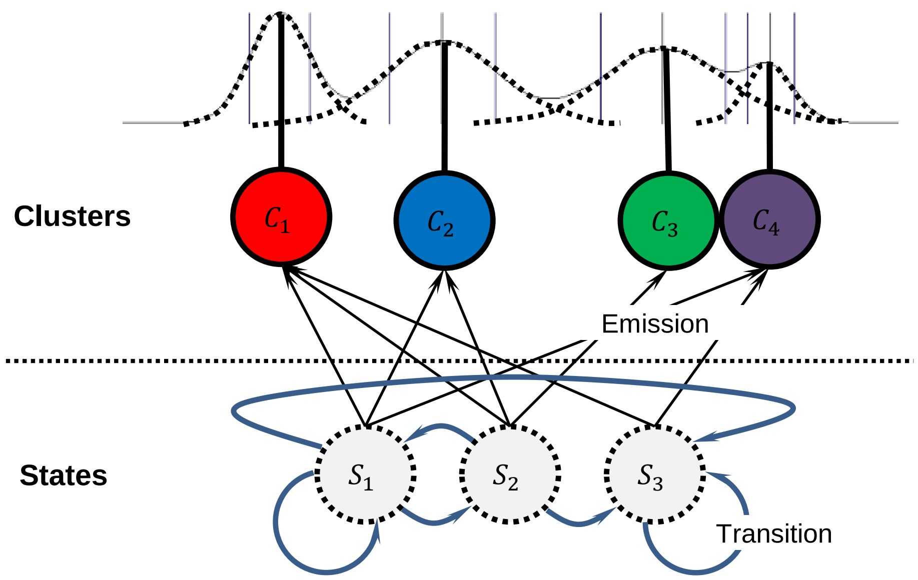
✓ Churners and non-churners exhibit different behaviours generated by some **states**



Methodology

Key Assumptions

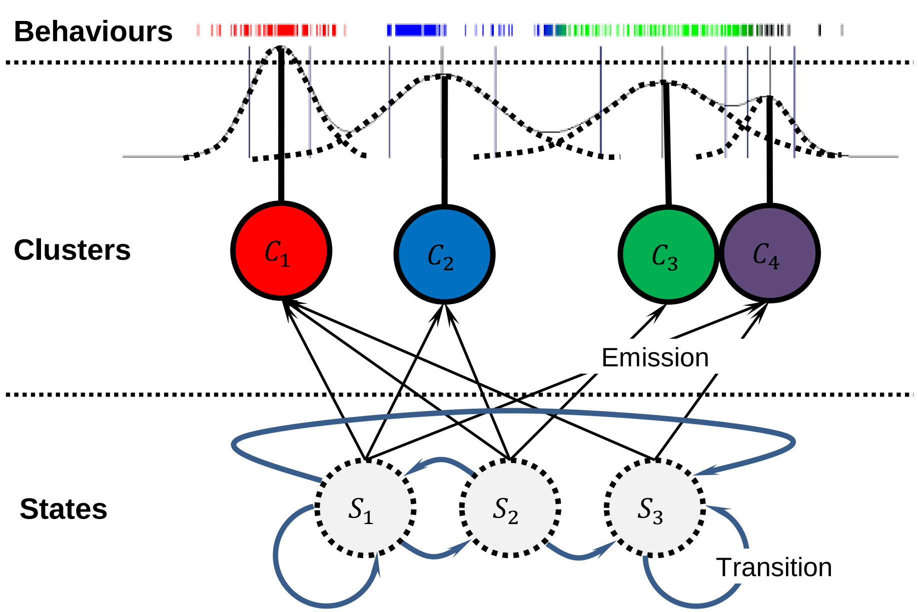
- ✓ Churners and non-churners exhibit different behaviours generated by some **states**
- ✓ Generative process: State → Cluster



Methodology

Key Assumptions

- ✓ Churners and non-churners exhibit different behaviours generated by some **states**
- ✓ Generative process: State → Cluster → Behaviour observations

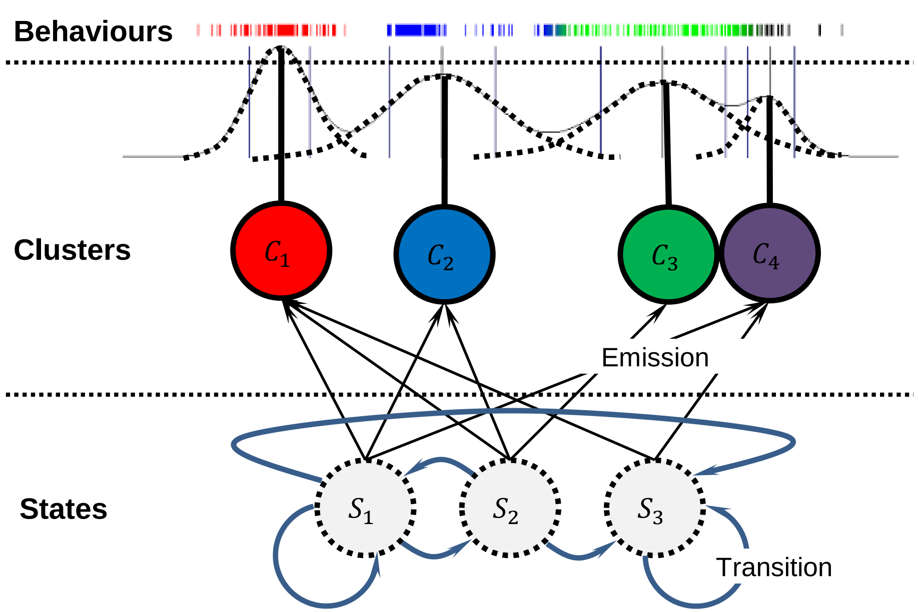


Methodology

InFoMM
Industrially Focused Mathematical Modelling

Key Assumptions

- ✓ Churners and non-churners exhibit different behaviours generated by some **states**
- ✓ Generative process: State → Cluster → Behaviour observations



Mixture Model

$$f(\boldsymbol{X}) = \sum_{k=1}^{K} \pi_k f_k(\boldsymbol{X} \mid \boldsymbol{\theta}_k)$$

Markov State Transition probability

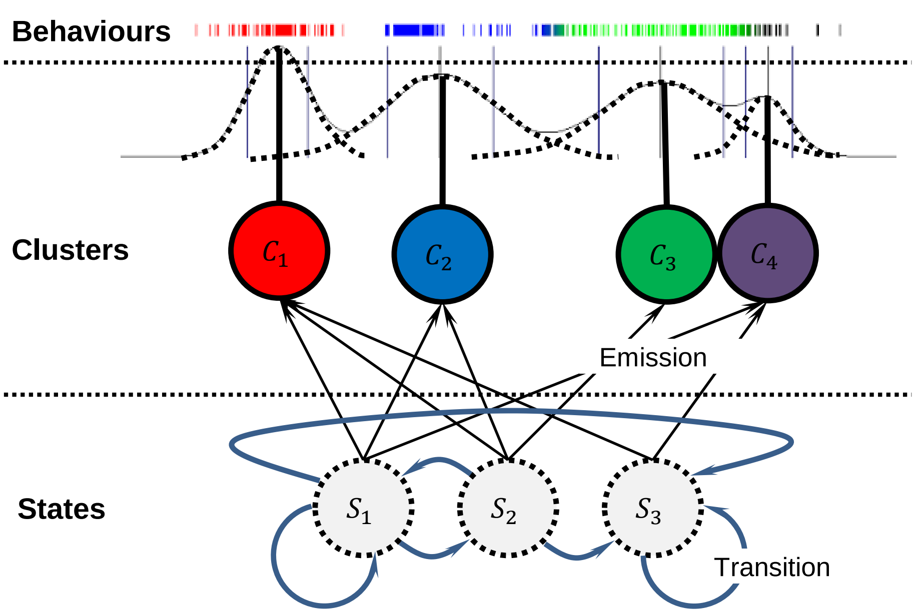
$$A = (a_{ij})$$
$$a_{ij} = P(s_{t+1} = S_i | s_t = S_j)$$

Methodology

InFoMM Industrially Focused Mathematical Modelling

Key Assumptions

- ✓ Churners and non-churners exhibit different behaviours generated by some **states**
- ✓ Generative process: State → Cluster → Behaviour observations
- ✓ Customers' behaviours are independent from others'



Mixture Model

$$f(\boldsymbol{X}) = \sum_{k=1}^{K} \pi_k f_k(\boldsymbol{X} \mid \boldsymbol{\theta_k})$$

Markov State Transition probability

$$A = (a_{ij})$$

$$a_{ij} = P(s_{t+1} = S_i | s_t = S_j)$$

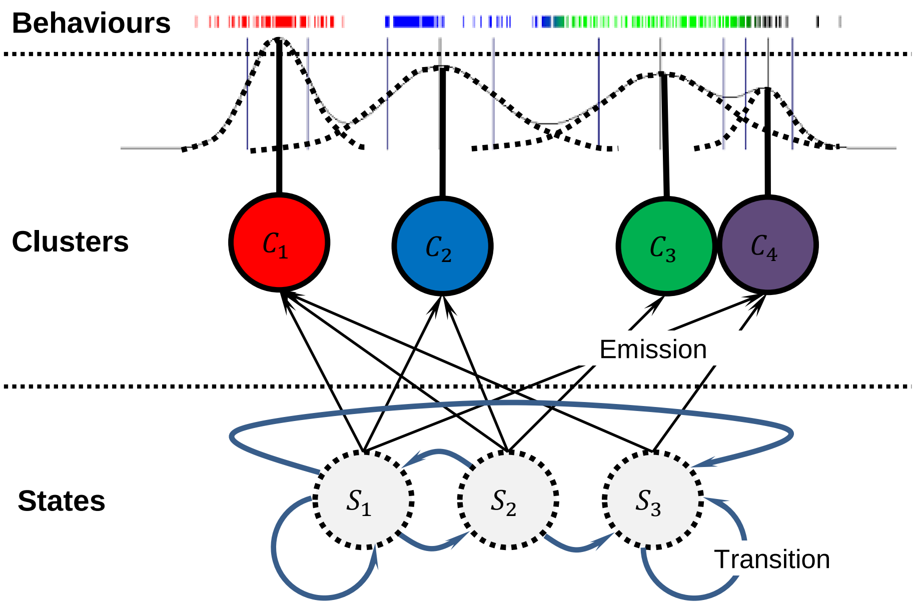
Methodology

InFoMM Industrially Focused Mathematical Modelling

Key Assumptions

- ✓ Churners and non-churners exhibit different behaviours generated by some **states**
- ✓ Generative process: State → Cluster → Behaviour observations
- ✓ Customers' behaviours are independent from others'

Make inference!



Mixture Model

$$f(\boldsymbol{X}) = \sum_{k=1}^{K} \pi_k f_k(\boldsymbol{X} \mid \boldsymbol{\theta_k})$$

Markov State Transition probability

$$A = (a_{ij})$$

$$a_{ij} = P(s_{t+1} = S_i | s_t = S_j)$$

Methodology

Industrially Focused Mathematical Modelling

Key Assumptions

- ✓ Churners and non-churners exhibit different behaviours generated by some **states**
- ✓ Generative process: State → Cluster → Behaviour observations
- ✓ Customers’ behaviours are independent from others’

Pipeline

Methodology

Key Assumptions

- ✓ Churners and non-churners exhibit different behaviours generated by some **states**
- ✓ Generative process: State → Cluster → Behaviour observations
- ✓ Customers' behaviours are independent from others'

Pipeline

Feature Extraction

Extract features to represent customers within a specific customer month.

Feature Distributional Modelling

Assess distributions, correlations, etc.

Clustering

Fit mixture model;

Define states

Analytics and Prediction

Interpret clusters

University of Oxford

Mathematical Institute

Methodology

InFoMM
Industrially Focused Mathematical Modelling

Key Assumptions

✓ Churners and non-churners exhibit different behaviours generated by some **states**

✓ Generative process: State → Cluster → Behaviour observations

✓ Customers’ behaviours are independent from others’

Pipeline

Feature Extraction

Extract features to represent customers within a specific customer month.

Feature Distributional Modelling

Assess distributions, correlations, etc.

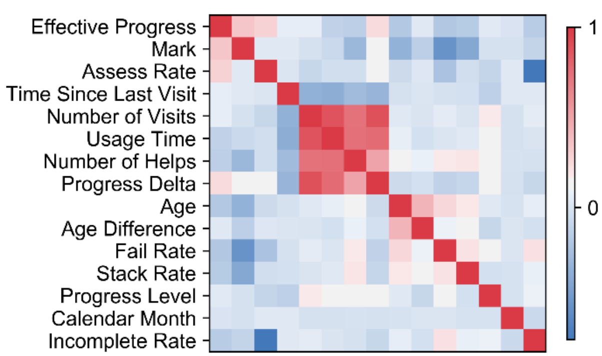
Clustering

Fit mixture model;

Define states

Analytics and Prediction

Interpret clusters



University of Oxford

Methodology

InFoMM
Industrially Focused Mathematical Modelling

Key Assumptions

- ✓ Churners and non-churners exhibit different behaviours generated by some **states**
- ✓ Generative process: State → Cluster → Behaviour observations
- ✓ Customers’ behaviours are independent from others’

Pipeline

Feature Extraction

Extract features to represent customers within a specific customer month.

Feature Distributional Modelling

Assess distributions, correlations, etc.

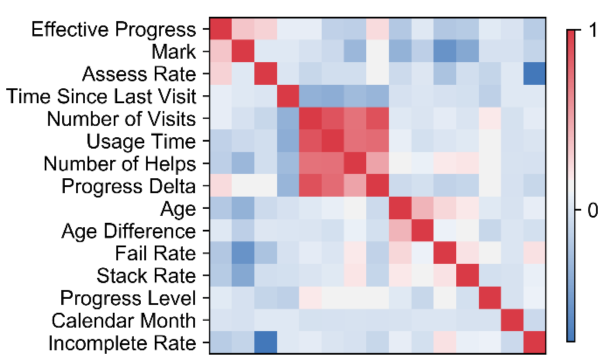
Clustering

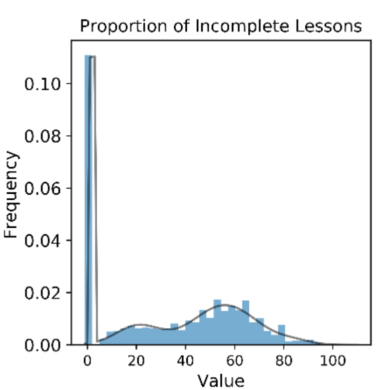
Fit mixture model;

Define states

Analytics and Prediction

Interpret clusters





Mathematical Institute

EPSRC Centre for Doctoral Training in Industrially Focused Mathematical Modelling

University of Oxford

Mathematical Institute

Methodology

InFoMM
Industrially Focused Mathematical Modelling

Key Assumptions

- ✓ Churners and non-churners exhibit different behaviours generated by some **states**
- ✓ Generative process: State → Cluster → Behaviour observations
- ✓ Customers’ behaviours are independent from others’

Pipeline

Feature Extraction

Extract features to represent customers within a specific customer month.

Feature Distributional Modelling

Assess distributions, correlations, etc.

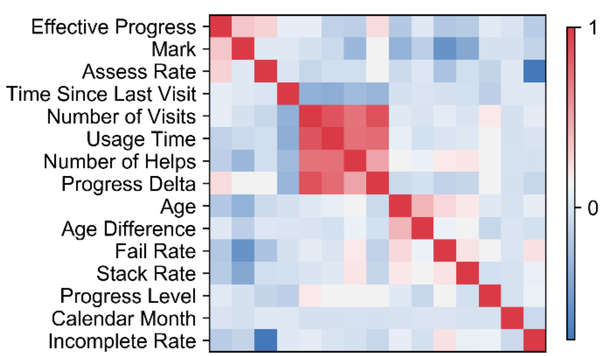
Clustering

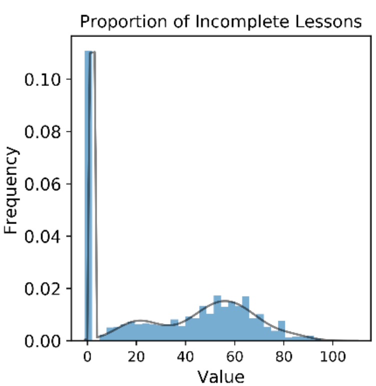
Fit mixture model;

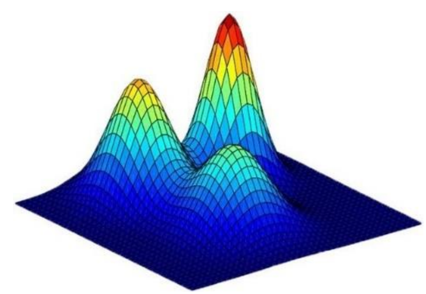
Define states

Analytics and Prediction

Interpret clusters







EPSRC Centre for Doctoral Training in Industrially Focused Mathematical Modelling

University of Oxford

Mathematical Institute

Methodology

Key Assumptions

- ✓ Churners and non-churners exhibit different behaviours generated by some **states**
- ✓ Generative process: State → Cluster → Behaviour observations
- ✓ Customers' behaviours are independent from others'

Pipeline

Feature Extraction

Extract features to represent customers within a specific customer month.

Feature Distributional Modelling

Assess distributions, correlations, etc.

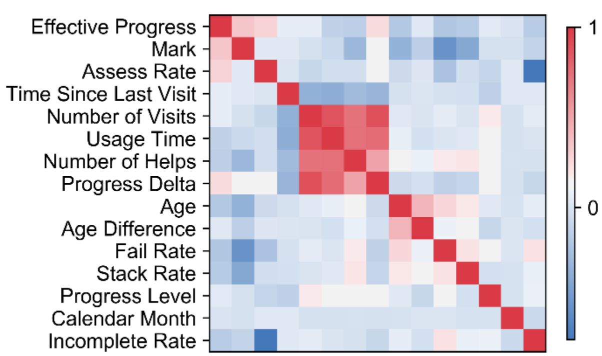
Clustering

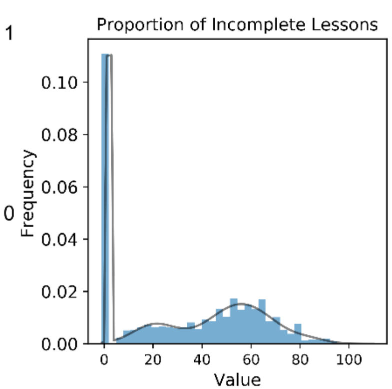
Fit mixture model;

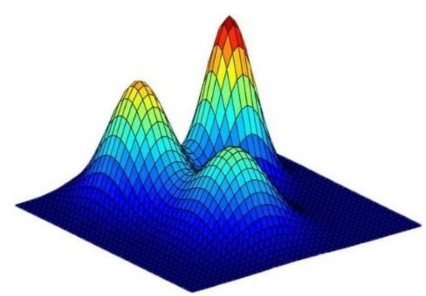
Define states

Analytics and Prediction

Interpret clusters







EPSRC Centre for Doctoral Training in Industrially Focused Mathematical Modelling

A Closer Look at Clustering…

Different users have different **data available**. For example, inactive users will have no records in features like marks, pass rates, etc.

Rather than interpolating missing information, we divide customers by activity level.

A Closer Look at Clustering…

Different users have different **data available**. For example, inactive users will have no records in features like marks, pass rates, etc.
Rather than interpolating missing information, we divide customers by activity level.

Inactive

G1
Inactive

No activity

G2
No-assess

There's activity but no assessment

G3
Fine

There's activity and assessment taken

Most Active

A Closer Look at Clustering…

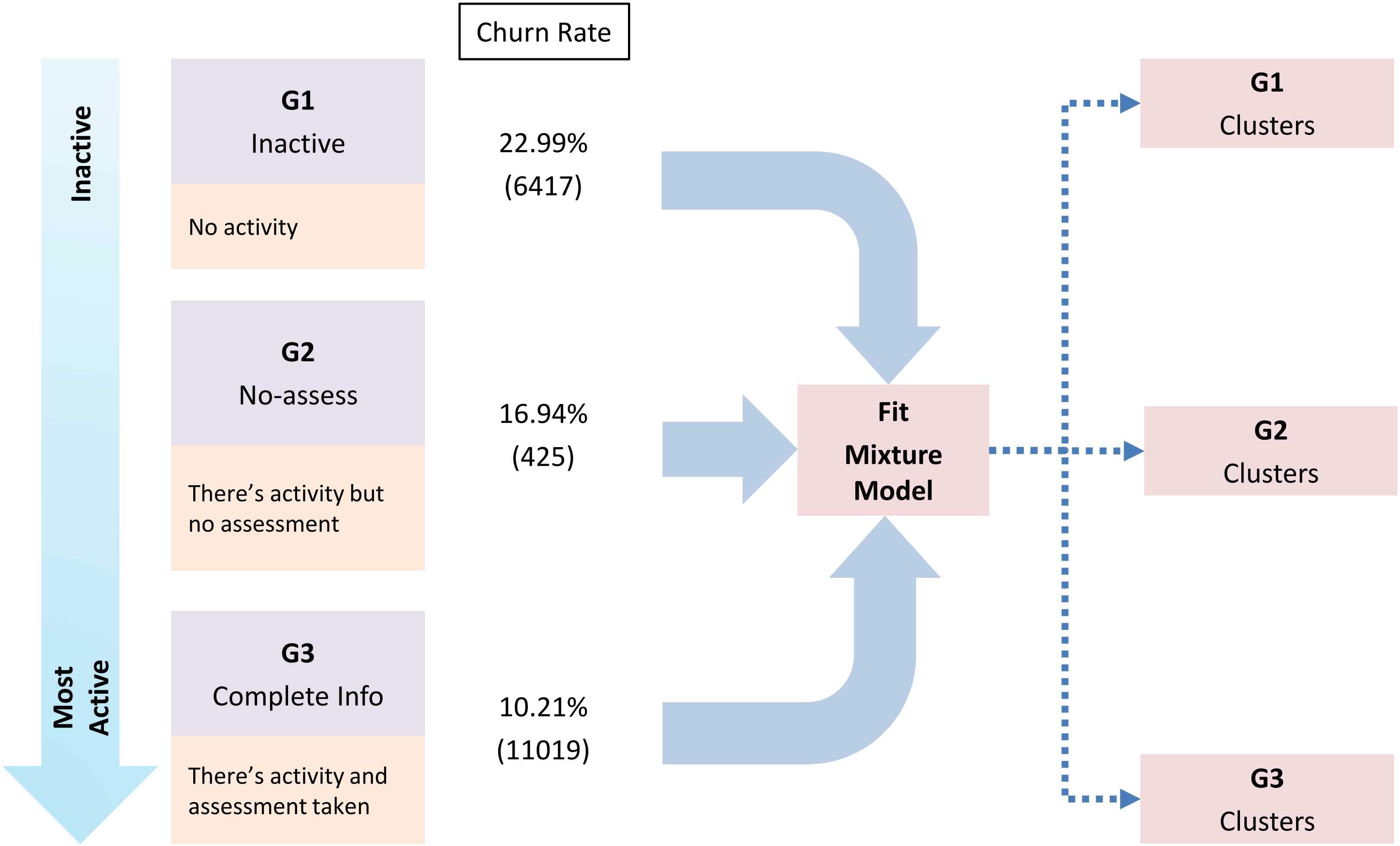
Different users have different **data available**. For example, inactive users will have no records in features like marks, pass rates, etc.
Rather than interpolating missing information, we divide customers by activity level.

Inactive → **Most Active**

		Churn Rate
G1 Inactive	No activity	22.99% (6417)
G2 No-assess	There's activity but no assessment	16.94% (425)
G3 Fine	There's activity and assessment taken	10.21% (11019)

A Closer Look at Clustering…

Different users have different **data available**. For example, inactive users will have no records in features like marks, pass rates, etc.
Rather than interpolating missing information, we divide customers by activity level.

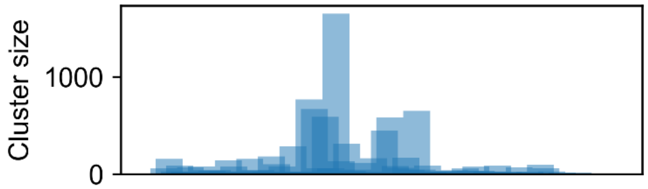


University of Oxford

Mathematical Institute

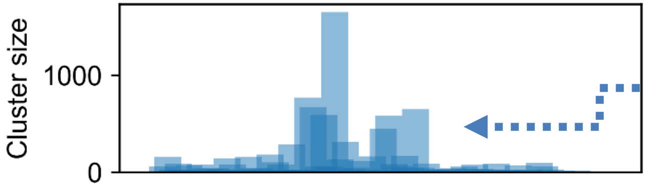
Cluster Churn Rates (Who)

Identified Clusters: Size



Cluster Churn Rates (Who)

Identified Clusters: Size



Features (X)

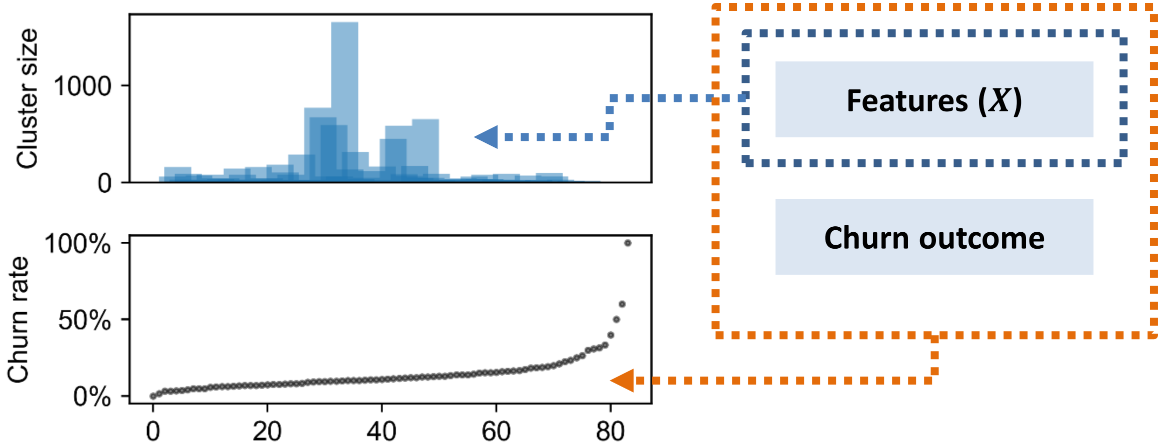
Churn outcome

Cluster Churn Rates (Who)

University of Oxford

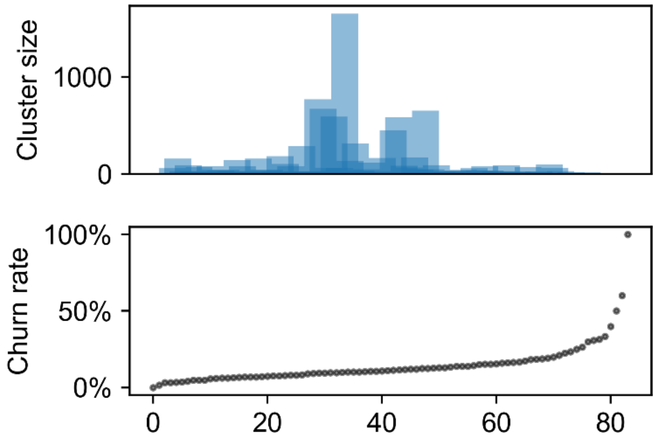
Mathematical Institute

Identified Clusters: Size and Churn Rate



Cluster Churn Rates (Who)

Identified Clusters: Size and Churn Rate

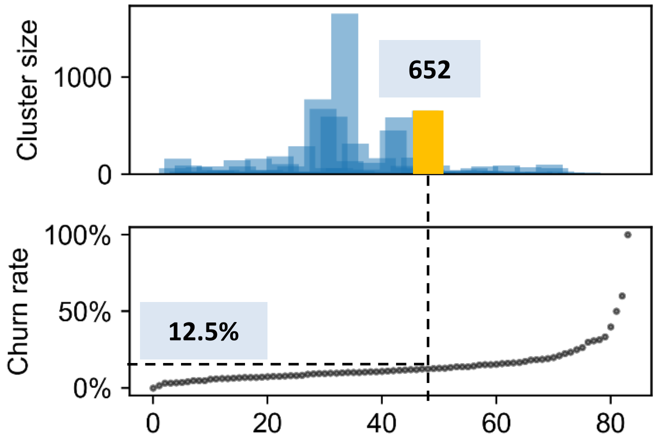


Observations:

- ✓ The purely behaviour based model infer many clusters:
  - ❑ the corresponding churn rate ranges from 0% to 100%.
- ✓ It is more challenging to identify users of extreme higher/low churn probability:
  - ❑ the cluster size tends to be much smaller for those extreme high/low churn rate.

Cluster Churn Rates (Who)

Identified Clusters: Size and Churn Rate



Observations:

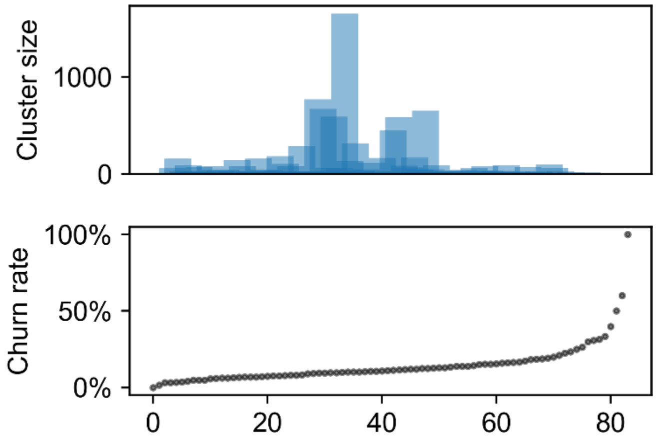
- ✓ The purely behaviour based model infer many clusters:
  - ❑ the corresponding churn rate ranges from 0% to 100%.
- ✓ It is more challenging to identify users of extreme higher/low churn probability:
  - ❑ the cluster size tends to be much smaller for those extreme high/low churn rate.

From **Cluster Churn Rate** to **Churn Probability of Individual Pupil**

Cluster Churn Rates (Who)

InFoMM Industrially Focused Mathematical Modelling

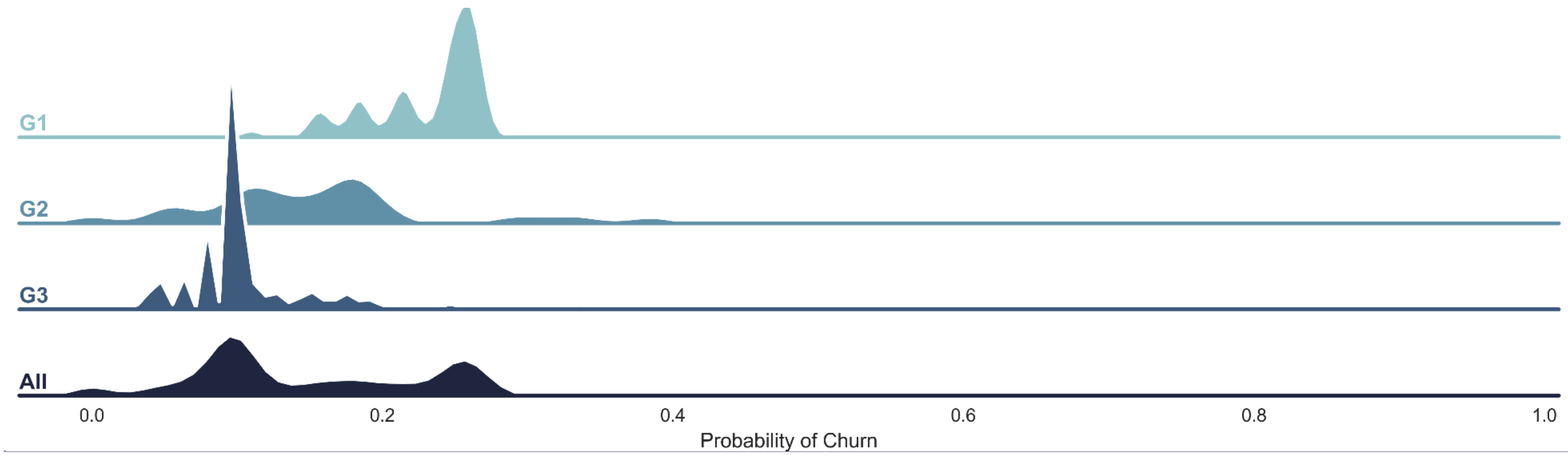
Identified Clusters: Size and Churn Rate



Observations:

- ✓ The purely behaviour based model infer many clusters:
  - ❑ the corresponding churn rate ranges from 0% to 100%.
- ✓ It is more challenging to identify users of extreme higher/low churn probability:
  - ❑ the cluster size tends to be much smaller for those extreme high/low churn rate.

From **Cluster Churn Rate** to **Churn Probability of Individual Pupil**

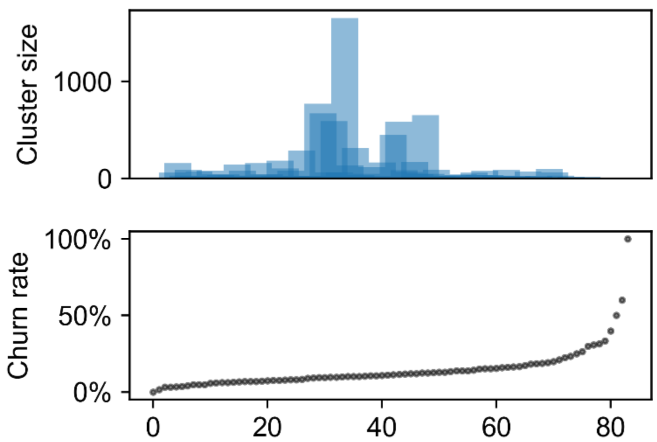


Cluster Churn Rates (Who)

InFoMM
Industrially Focused Mathematical Modelling

University of Oxford
Mathematical Institute

Identified Clusters: Size and Churn Rate

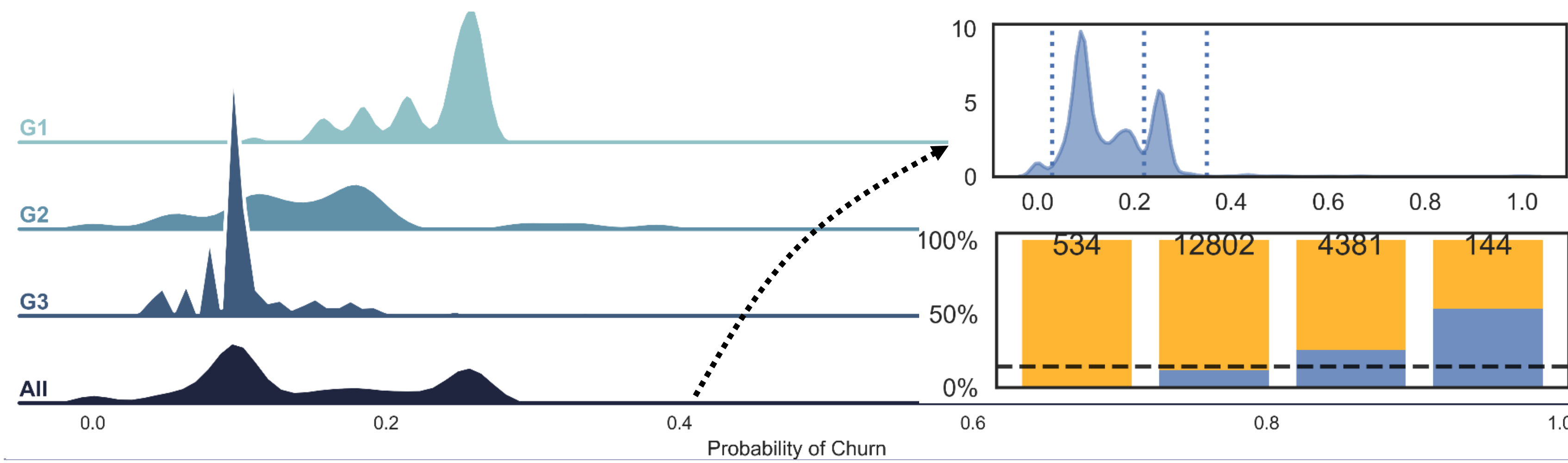


Observations:

✓ The purely behaviour based model infer many clusters:
- ❑ the corresponding churn rate ranges from 0% to 100%.

✓ It is more challenging to identify users of extreme higher/low churn probability:
- ❑ the cluster size tends to be much smaller for those extreme high/low churn rate.

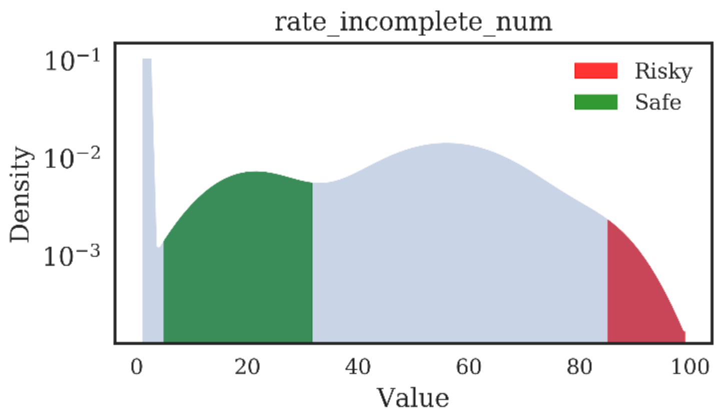
From **Cluster Churn Rate** to **Churn Probability of Individual Pupil**

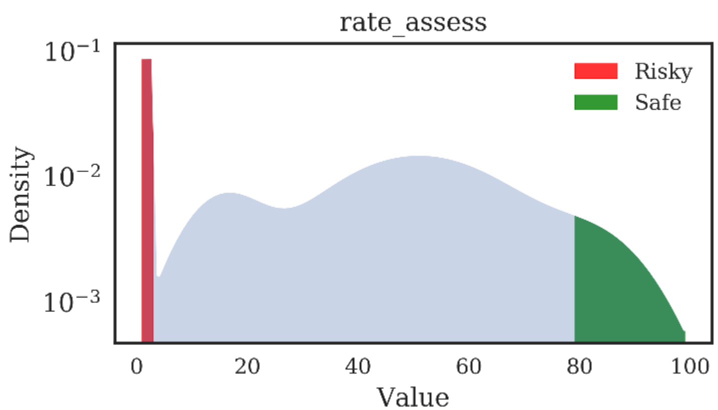


Feature Analysis (Why)

InFoMM Industrially Focused Mathematical Modelling

The mixture model methodology, by its nature, is designed for identifying clusters, but not explicitly for analysing feature importance.
However, by counting frequencies of components of **risky** and **safe** clusters, we can see how feature impacts churn outcome.





Risky cluster: churn rate > 50%

Safe cluster: churn rate < 5%

We can look at the component each cluster falls in for features.

Observations:

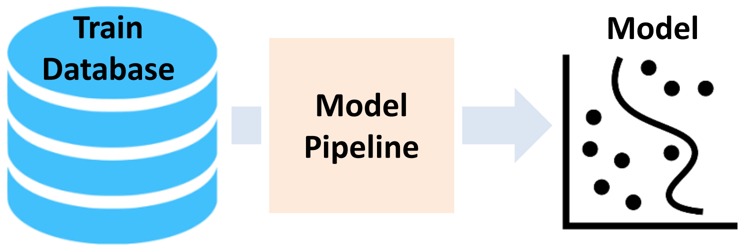
- ✓ Subscribers having many incomplete lesson records are most likely to churn.
- ✓ Subscribers taking few assessment are most likely to churn

Prediction Workflow

InFoMM
Industrially Focused Mathematical Modelling

How does Whizz use this model to predict the risk of churn for a set of subscribers?

Step I: Train

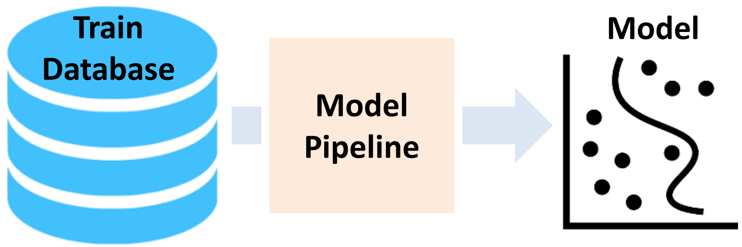


Prediction Workflow

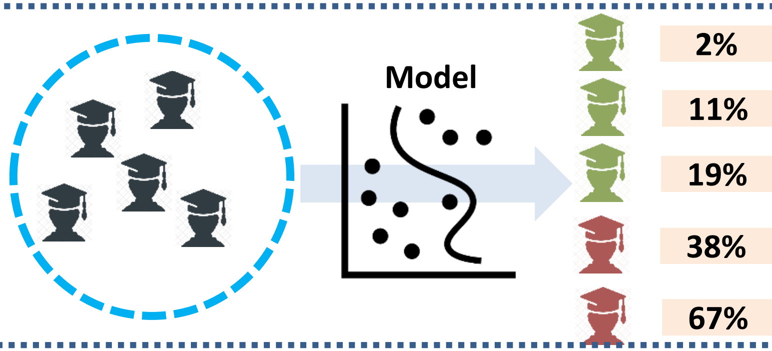
InFoMM Industrially Focused Mathematical Modelling

How does Whizz use this model to predict the risk of churn for a set of subscribers?

Step I: Train



Step II: Predict



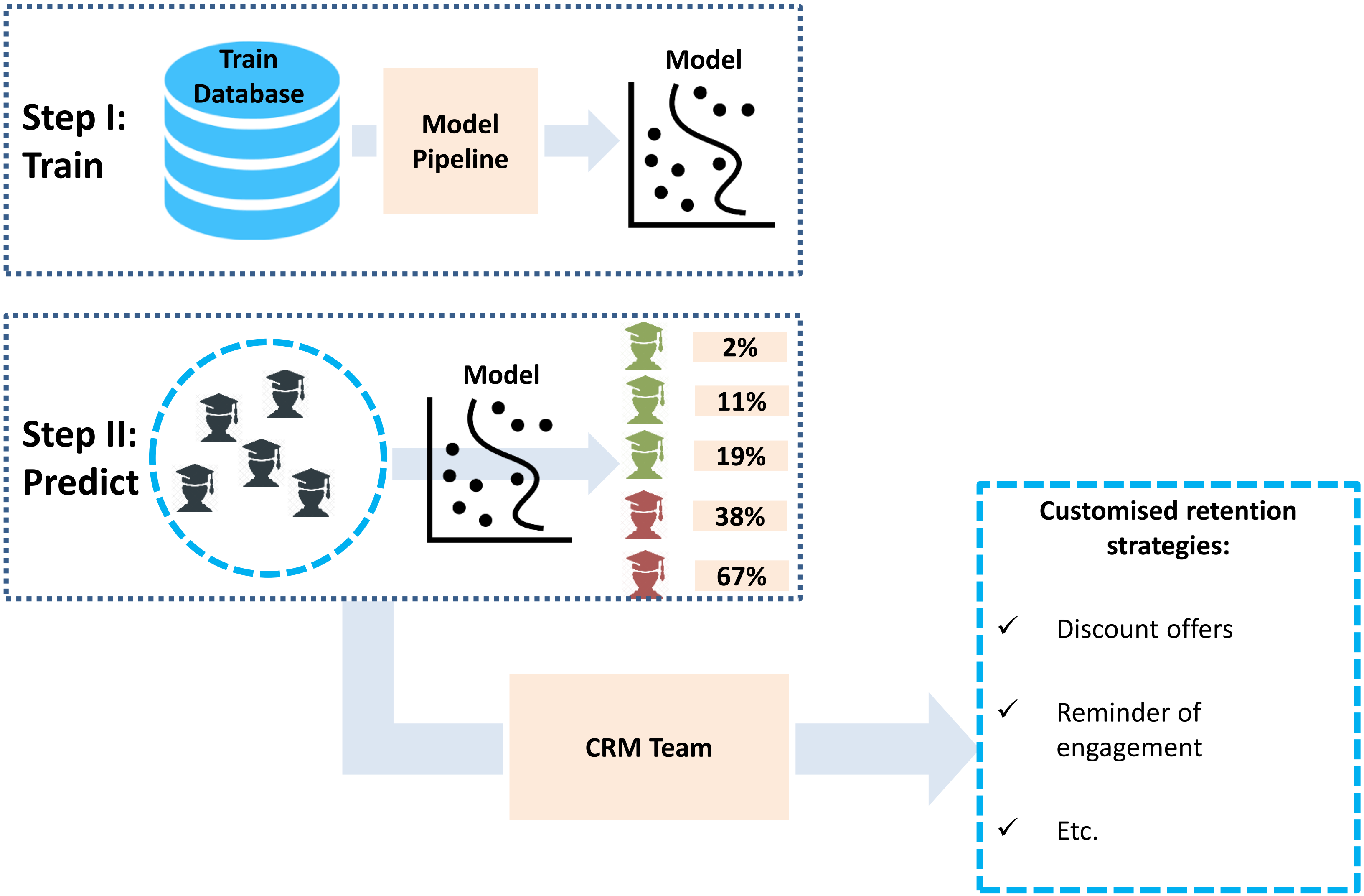
University of Oxford

Mathematical Institute

Prediction Workflow

InFoMM
Industrially Focused Mathematical Modelling

How does Whizz use this model to predict the risk of churn for a set of subscribers?

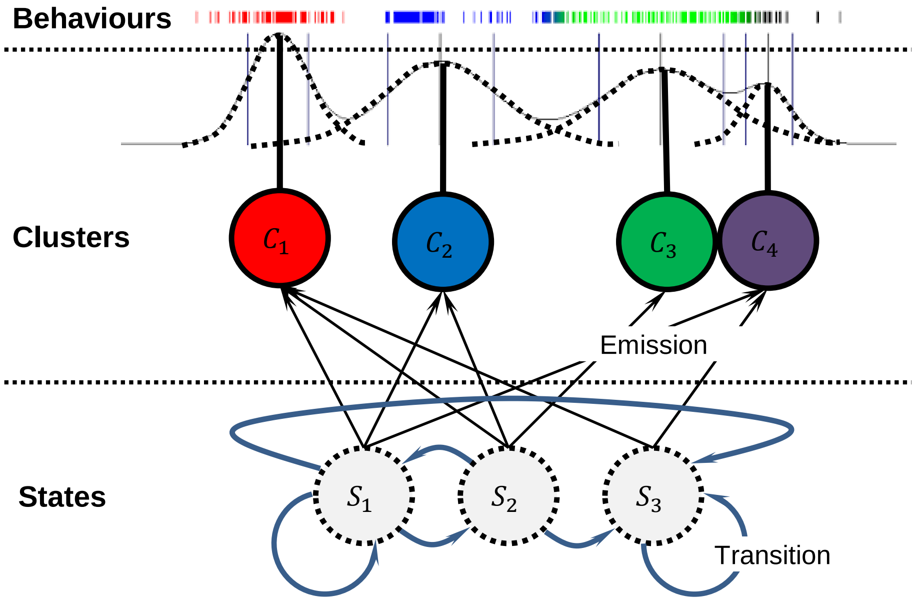


University of Oxford

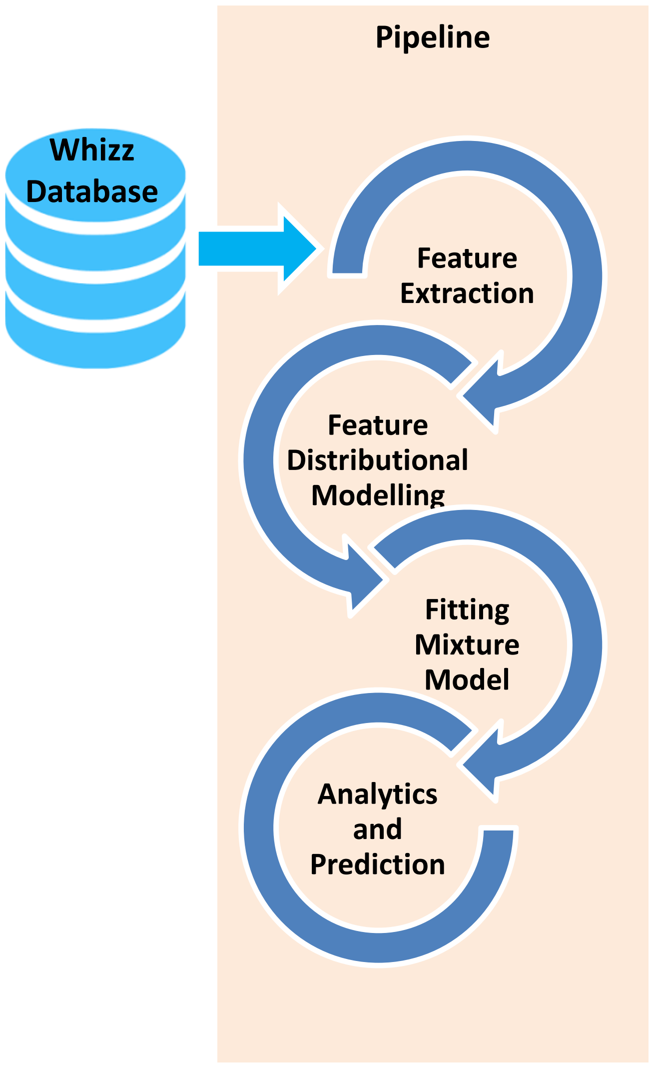
Mathematical Institute

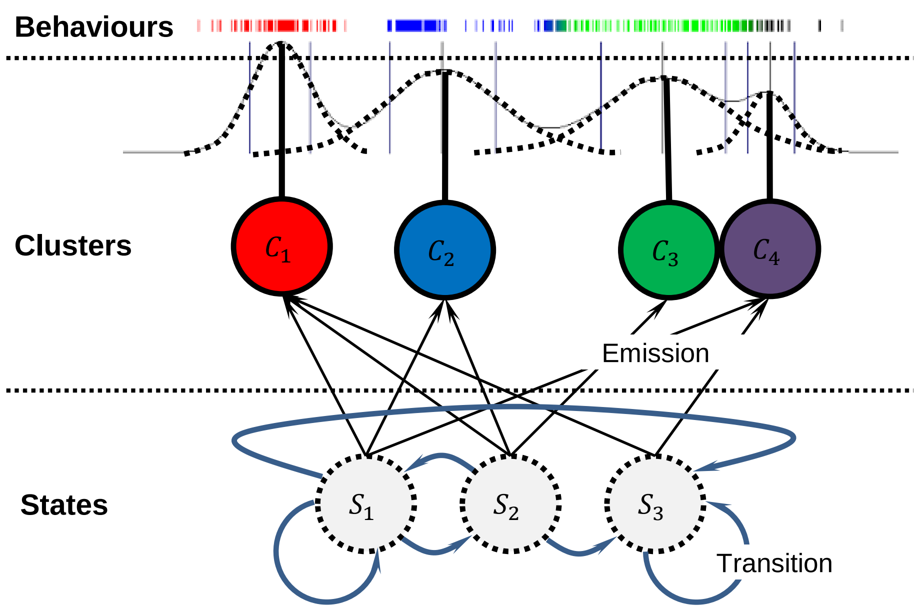
Conclusion

InFoMM
Industrially Focused Mathematical Modelling

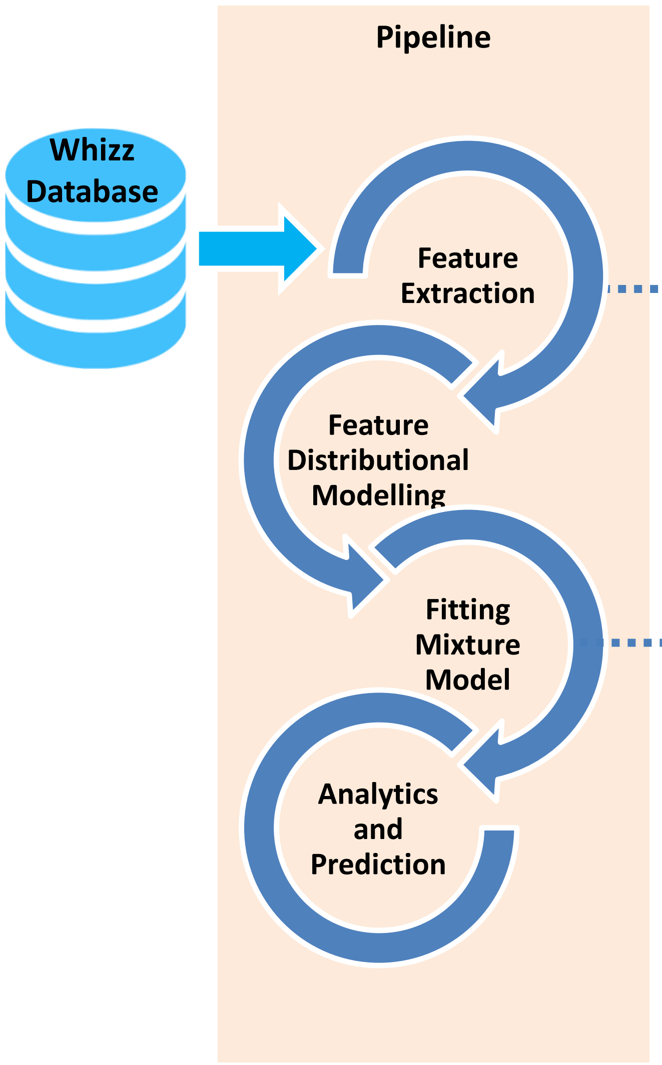


Conclusion


Pipeline
Whizz Database
Feature Extraction
Feature Distributional Modelling
Fitting Mixture Model
Analytics and Prediction


Behaviours
Clusters
C_1
C_2
C_3
C_4
Emission
States
S_1
S_2
S_3
Transition

Conclusion

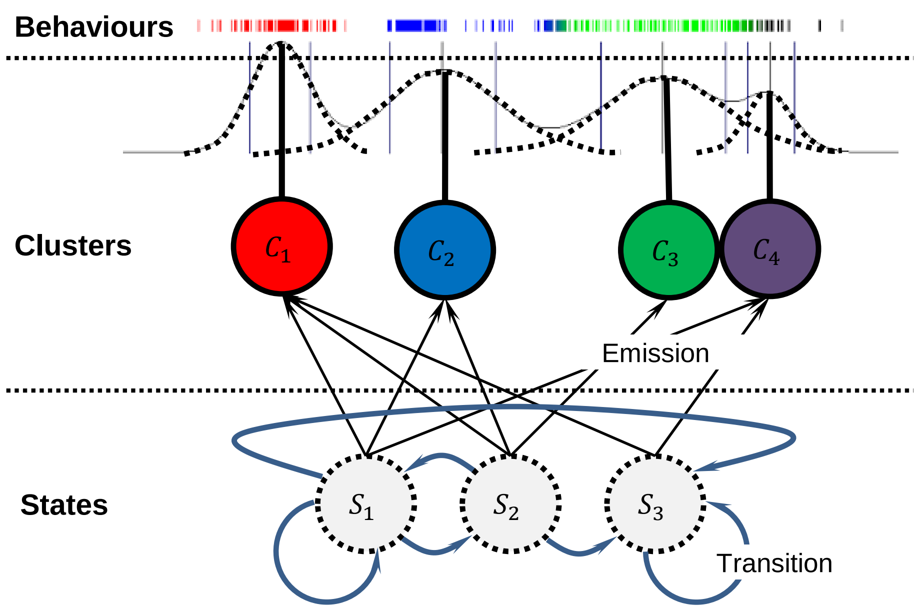


Feature selection:

- ✓ Select features by industrial expertise
- ✓ Add nonlinear transformation / interaction of existing features

Feature importance :

- ✓ Measure importance of dimension in a multivariate setting



Engineering and Physical Sciences Research Council

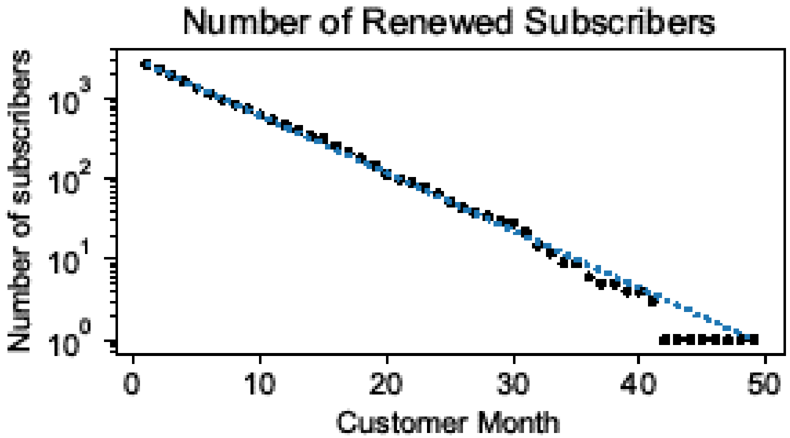
Industrially Focused Mathematical Modelling

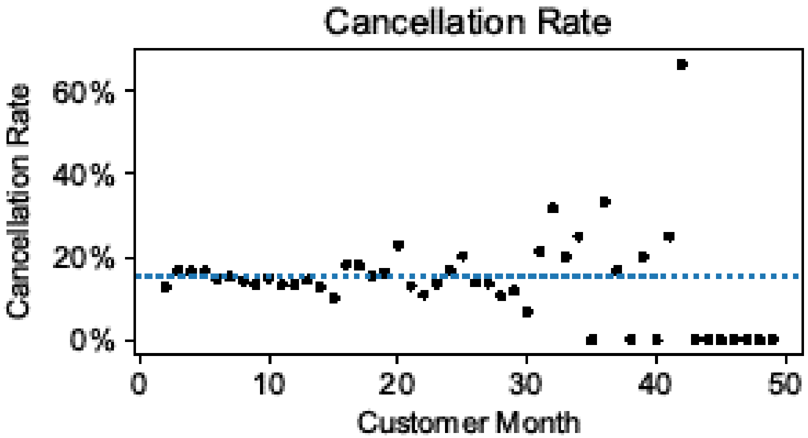
Appendix

Customer Month Independence

We assume the cancellation to be **ONLY** dependent on the current month.

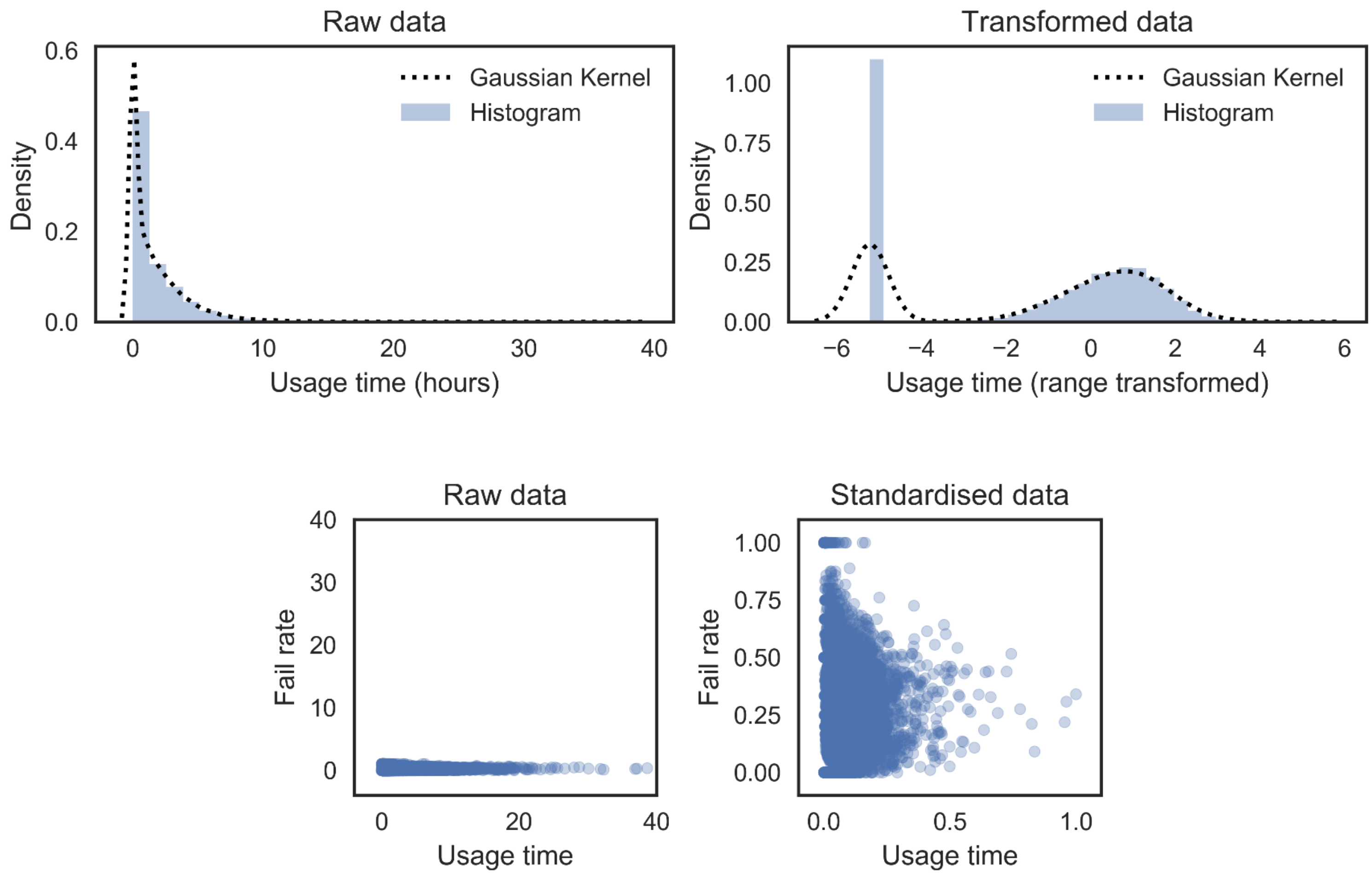
- ✓ The assumption holds since we observe statistically constant churn rate over customer month.
- ✓ The assumption enables us to treat activities in different customer months indifferently.



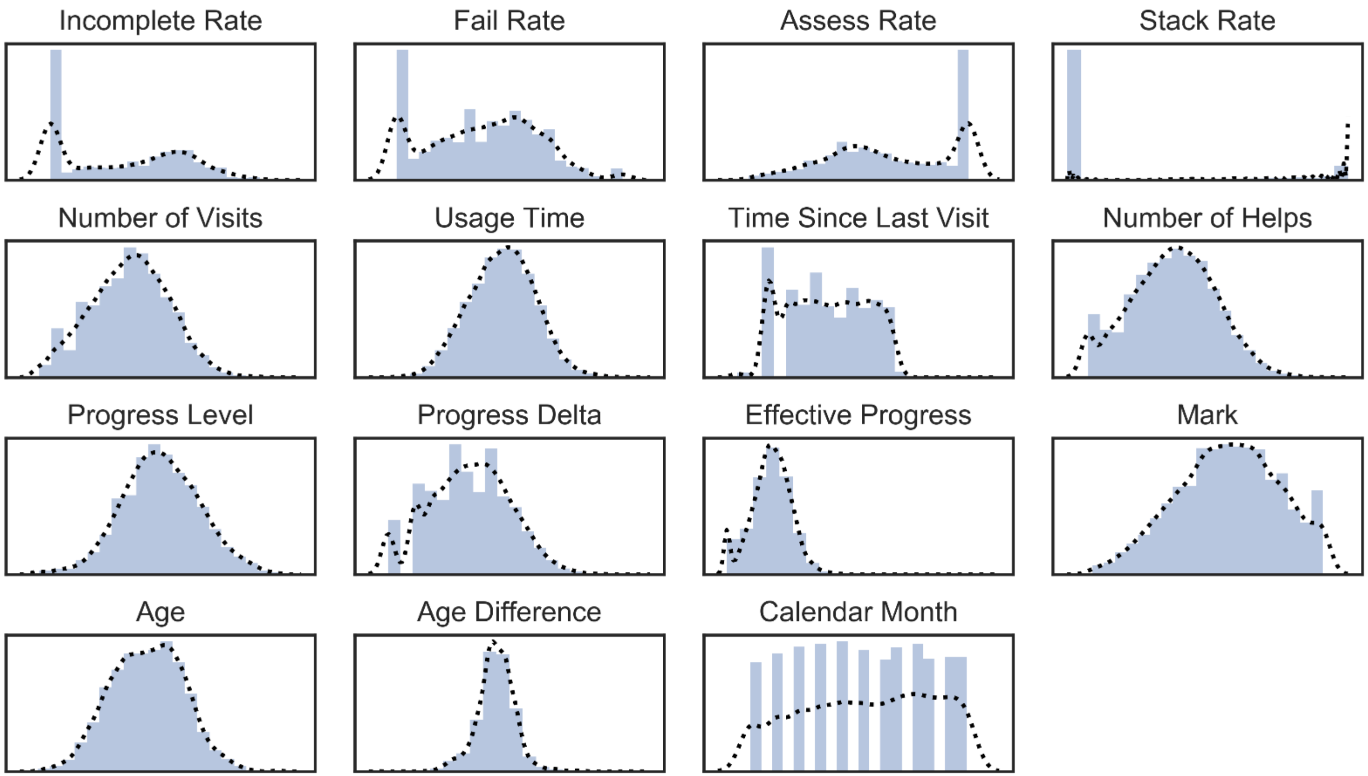


Data Transformation

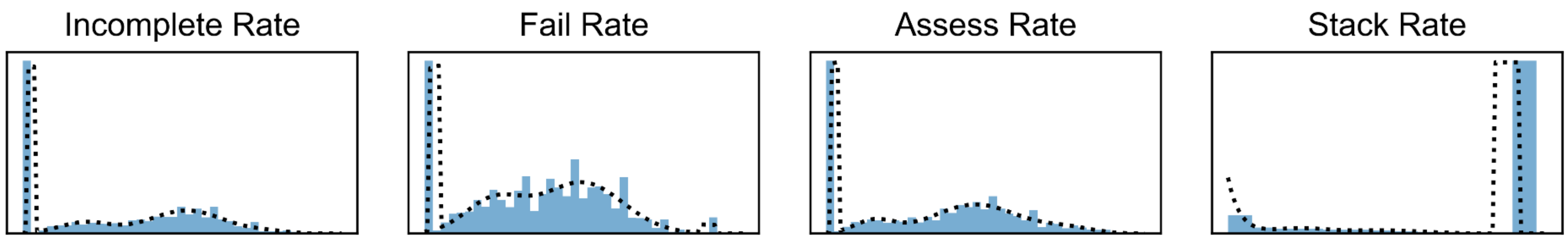
InFoMM
Industrially Focused Mathematical Modelling

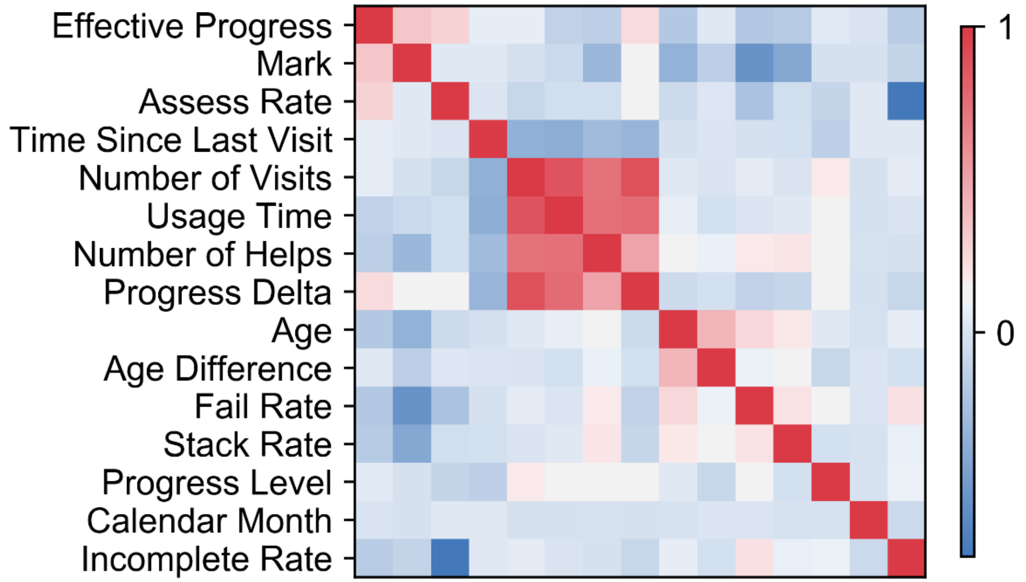


Distributional Modelling – I



Distributional Modelling – II





Dirichlet Process Mixture – I

DP mixture model

Use DP as prior for $\boldsymbol{\theta}$

$$\boldsymbol{x}|\theta \sim f(\cdot\,|\theta), \qquad \theta|G \sim G(\cdot), \qquad G \sim DP(\cdot\,|G_0, \alpha)$$

Definitions A Dirichlet Process (DP) is a distribution of a random measure. Let G_0 be a base distribution (measure) for our cluster density parameter $\theta \in \Theta$, a measurable space, and let α be a positive, real-valued scalar. A random measure G is then distributed according to *Dirichlet Process* with scaling parameter α and base measure G_0:

$$G \sim \mathrm{DP}(\cdot|G_0, \alpha), \tag{16a}$$

if for all $K \in \mathbb{N}$, and all $\{\Theta_1, \ldots, \Theta_K\}$ finite partitions of Θ:

$$(G(\theta_1), \ldots, G(\theta_K)) \sim \mathrm{Dir}\left(\alpha G_0(\theta_1), \ldots, \alpha G_0(\theta_K)\right), \tag{16b}$$

where $\mathrm{Dir}(\cdot)$ denotes the *Dirichlet distribution*. The Dirichlet distribution is a distribution of the standard $K-1$ simplex. Let $\boldsymbol{\pi} = \{\pi_k\}_{k=1}^K$ with $\sum_{k=1}^K \pi_k = 1$ and $\forall k : \pi_k \geq 0$, and let $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$ with $\alpha_1, \ldots, \alpha_K \geq 0$. Then

$$\mathbb{P}(\boldsymbol{\pi}|\boldsymbol{\alpha}) = \mathrm{Dir}(\alpha_1, \ldots, \alpha_K) = \frac{1}{\mathrm{Beta}(\boldsymbol{\alpha})} \prod_{k=1}^K \pi_k^{\alpha_k - 1} = \frac{\Gamma\left(\sum_k \alpha_k\right)}{\prod_k \Gamma(\alpha_k)} \prod_{k=1}^K \pi_k^{\alpha_k - 1}, \tag{16c}$$

where $\mathrm{Beta}(\cdot)$ is the beta function, $\Gamma(\cdot)$ is the gamma function.

Dirichlet Process Mixture – II

Clustering Effect We use DP as a prior to distribution of cluster parameter θ:

$$\theta|G \sim G(\cdot) \quad \text{and} \quad G \sim \mathrm{DP}(\cdot|G_0, \alpha). \tag{17}$$

This model exhibits a "clustering effect" which enables us to infer number of clusters from data rather than pre-defining it. Suppose we independently draw n random values $\theta^{(j)}$ from G under the model (17), then Blackwell and MacQueen's urn representation theorem [1] states that, marginalising out the random measure G, the joint distribution of the collection of variables $\{\theta^{(1)}, \dots, \theta^{(n)}\}$ exhibits a clustering effect:

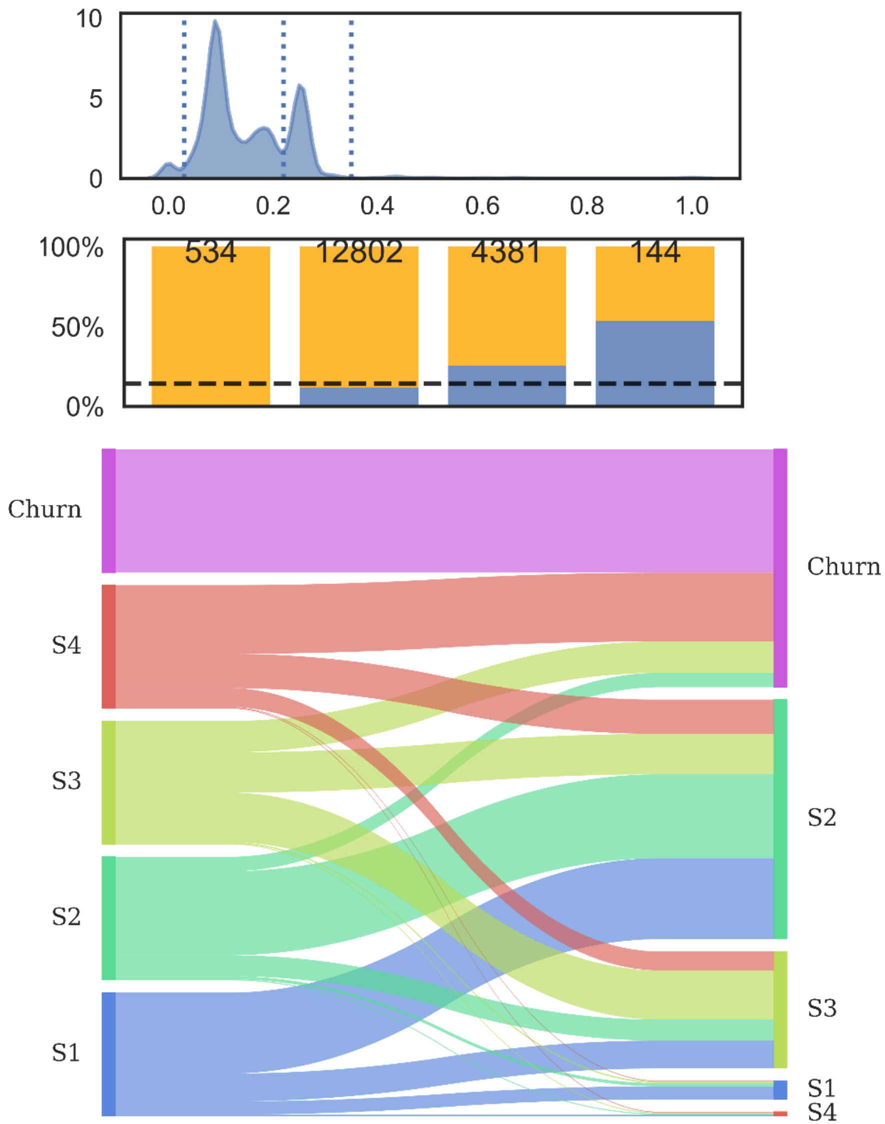
$$\mathbb{P}\left(\theta^{(j)}|\theta^{(1)}, \dots, \theta^{(j-1)}\right) \propto \alpha G_0(\theta^{(j)}) + \sum_{l=1}^{j-1} \delta_{\theta^{(l)}}(\theta^{(j)}), \tag{18}$$

where $\delta_{\theta^{(l)}}(\cdot)$ is a Dirac delta at $\theta^{(l)}$. Thus the variables $\{\theta^{(1)}, \dots, \theta^{(n)}\}$ are randomly partitioned according to which variables are equal to the same value. Moreover, let $\{\theta_1, \dots, \theta_K\}$ denote the distinct values of the drawn samples $\{\theta^{(1)}, \dots, \theta^{(j-1)}\}$, let $\{\kappa_1, \dots, \kappa_{j-1}\}$ be the assignment variables such that $\theta^{(l)} = \theta_{\kappa_l}$. Then,

$$\mathbb{P}\left(\theta^{(j)}|\theta^{(1)}, \dots, \theta^{(j-1)}\right) \propto \frac{\alpha}{j-1+\alpha} G_0(\theta^{(j)}) + \sum_{k=1}^{K} \frac{|\{l : \kappa_l = k\}|}{j-1+\alpha} \delta_{\theta^{(l)}}(\theta^{(j)}). \tag{19}$$

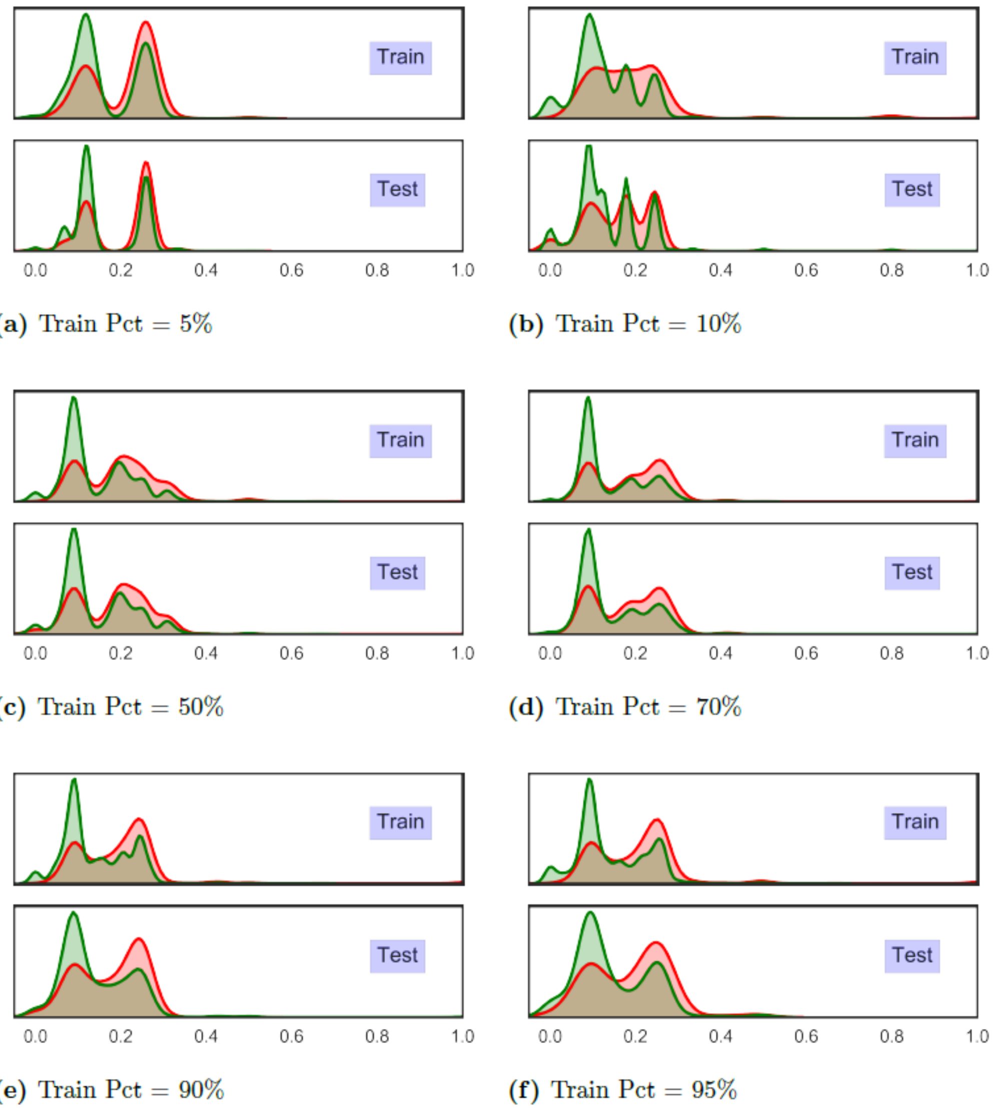
Temporal Transition of States

InFoMM
Industrially Focused Mathematical Modelling



- ✓ S2 is the largest destination state, which makes sense as S2 represents a "normal" state in which the churn probability is approximately the same as population average.
- ✓ Barely pupils transit from other states to S4 (the riskiest state). This seems imply that 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 S4.

Overfitting



(a) Train Pct = 5%
(b) Train Pct = 10%
(c) Train Pct = 50%
(d) Train Pct = 70%
(e) Train Pct = 90%
(f) Train Pct = 95%