Data Mining for Business

Automating Data Mining Solutions & Model Monitoring

Dr. Shipra Maurya

Department of Management Studies

IIT (ISM) Dhandbad

Email: shipra@iitism.ac.in

Automating Data Mining Solutions

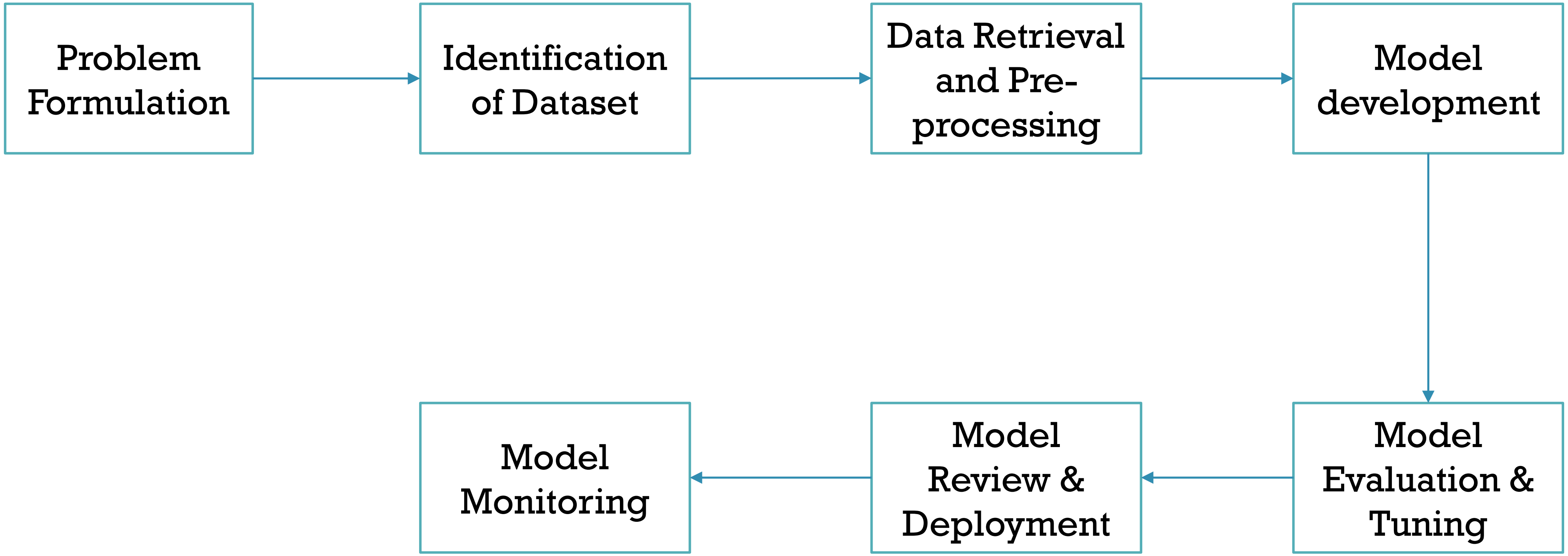
- Automation is a process in which very least amount of manual intervention is required in running a process/workflow

- In data mining applications, we focus on models that can be used on an ongoing basis to predict or classify new records. One time models (static models) are used for ad-hoc studies.

- The initial analysis will be in prototype mode, while we explore and define the problem and test different models. At this stage, all steps in data mining pipeline are followed

- Once the model is finalized, it has to be deployed in an automated fashion.

Data Mining Pipeline/ Process



Automating Data Mining Solutions

Data streams → Data Pipeline for scoring purpose → Scoring the new records using the deployed model → Displaying the model output on the user interface or share the output using any other mode

Iterative process (the model scores the new records on scheduled time which is decided by the analytics team)

Model Monitoring

- It is an operational stage that comes post model deployment in data mining pipeline

- It entails monitoring the ML models for any model degradation and data drift etc. to ensure that the model is maintaining a particular level of performance (MAPE/Accuracy/Precision & Recall/Rank ordering etc.)

- Earlier model performance was measured looking at the usage level and cost metrics

- At present, organizations are looking forward to automated model monitoring systems which consider model quality, data quality etc.

Why Model Monitoring is required?

- **Loss of brand reputation** - Amazon's AI powered Recruiting tool
- **Life risk** - Uber's self-driving car fatality
- **Financial loss** - HonKong real estate tycoon Li sues Tyndaris Investments in 2017 after an AI's automated trade cost him USD 20 MN
- **Information loss** - Face ID hacked using a 3D printed mask

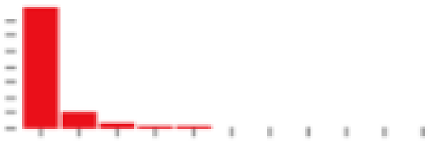
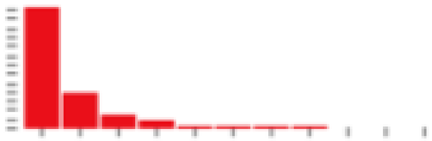
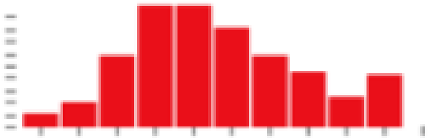
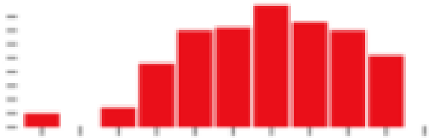
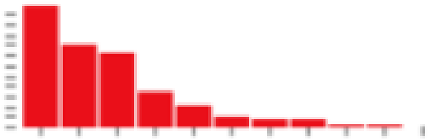
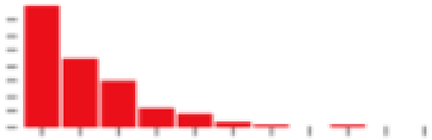
Why good models go bad? (1/6)

- Models are probabilistic and trained on historical data. This means models deployed into production carry forward characteristics of the data used to train them, including any hidden biases.

- It also means their output will change if the relationship between the incoming data and the predicted target drift apart

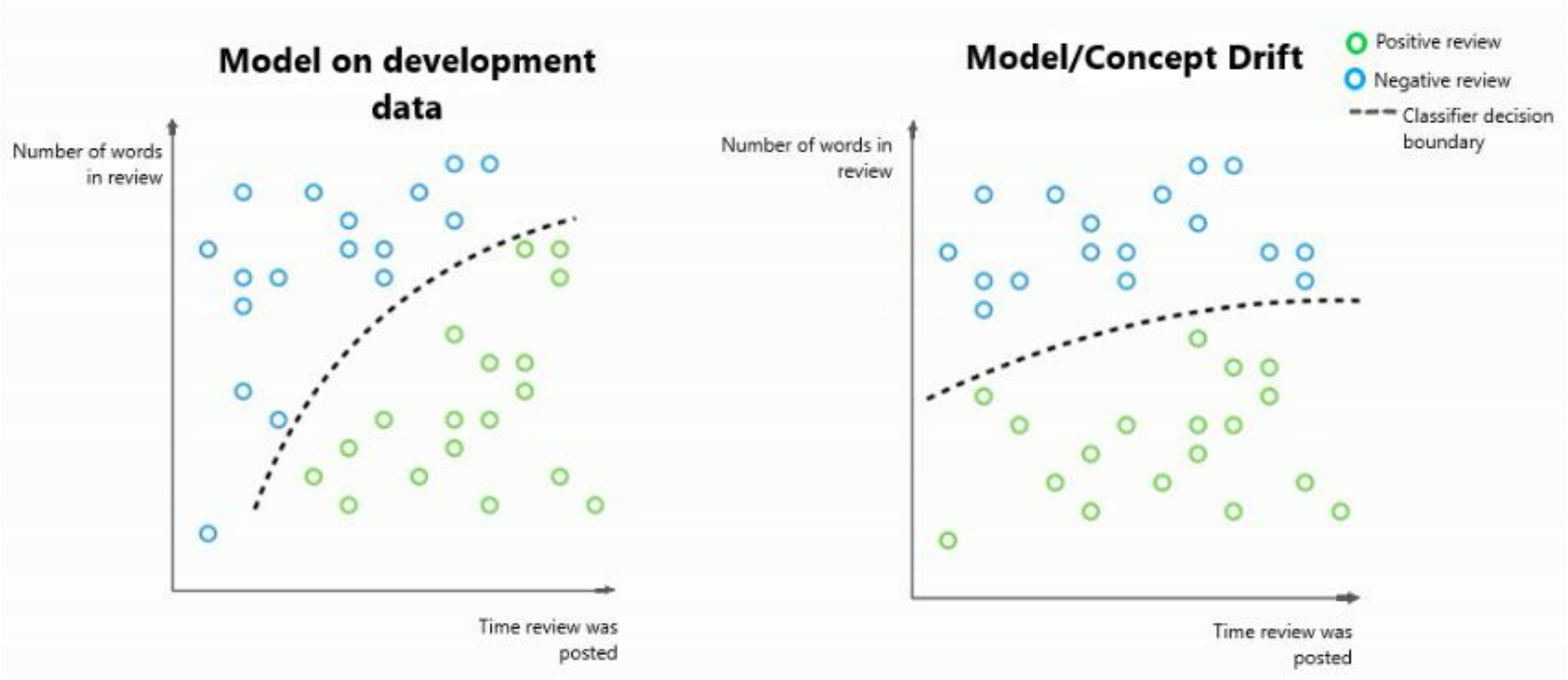
- **Data Drift** – The patterns in production data that a deployed model uses for predictions gradually diverge from the patterns in the model's original training data, which lowers predictive power of the model.

Why good models go bad? (2/6)

Feature	Type	Reference Distribution	Production Distribution
casual	num		
humidity	num		
season	num		
registered	num		

Why good models go bad? (3/6)

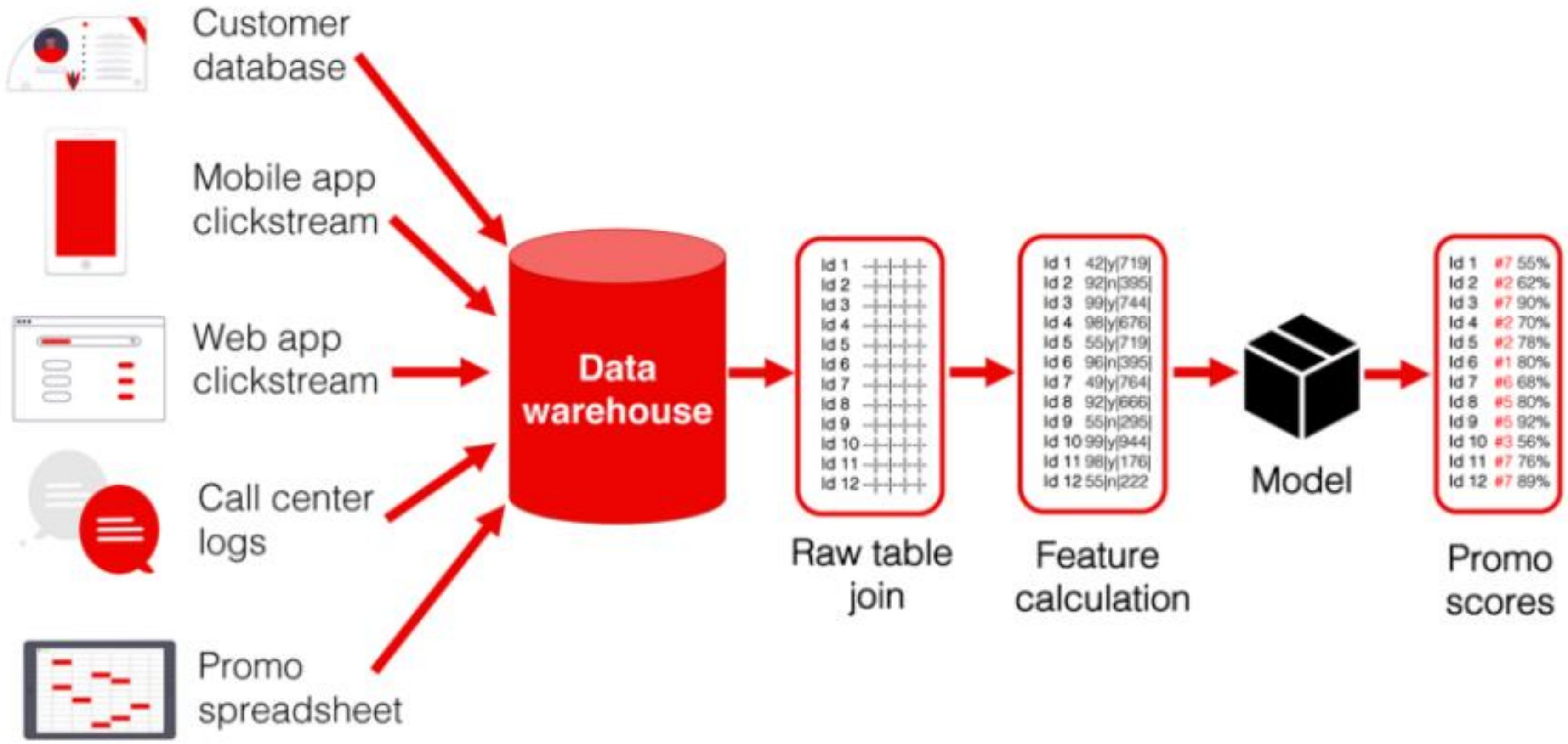
- **Model/concept Drift** – happens when the relationship between features and/or labels no longer holds because the learned relationship/patterns have changed over time.



Source: Internet

Why good models go bad? (4/6)

- Data pipeline issues



Source: Internet

Why good models go bad? (5/6)

- Data schema change

BEFORE

Cl_ID	Name	Type	Length	Status
#1229	######	card	2:27	solved
#1203	######	card	12:12	solved
#5661	######	account	8:06	solved
#8791	######	account	1:01	solved

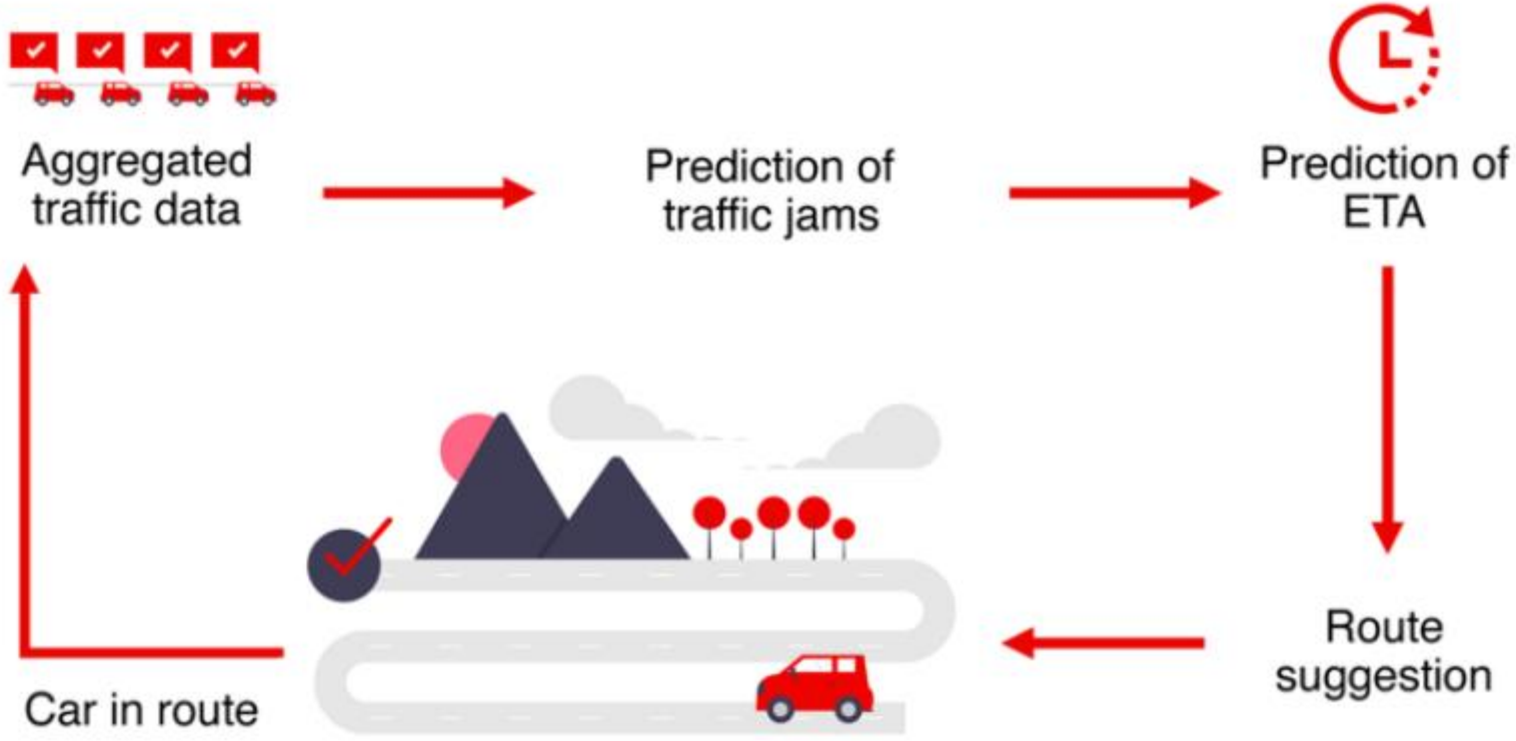
AFTER

Client ID	Client name	Call Type	Call Length	Channel preference	Status
#1229	######	card-lost	2:27	phone	solved
#1203	######	card-lost	12:12	phone	solved
#5661	######	account-balance	8:06	phone	solved
#8791	######	account-balance	1:01	email	solved

Source: Internet

Why good models go bad? (6/6)

- Broken upstream models



Source: Internet

Learnings from Good models going Bad

- Empty carts for Instacart (online grocery shopping service)-
  - Developed a ML model for predicting whether a particular product would be available at a given store with a 93% accuracy rate
  - In March 2020, the accuracy rate of the model suddenly plunged to 61% for many products – changed shopping behavior of customer due to COVID – 19
  - Instacart's quick response – reduced the timescale of the data to AI models from weeks to 10 days

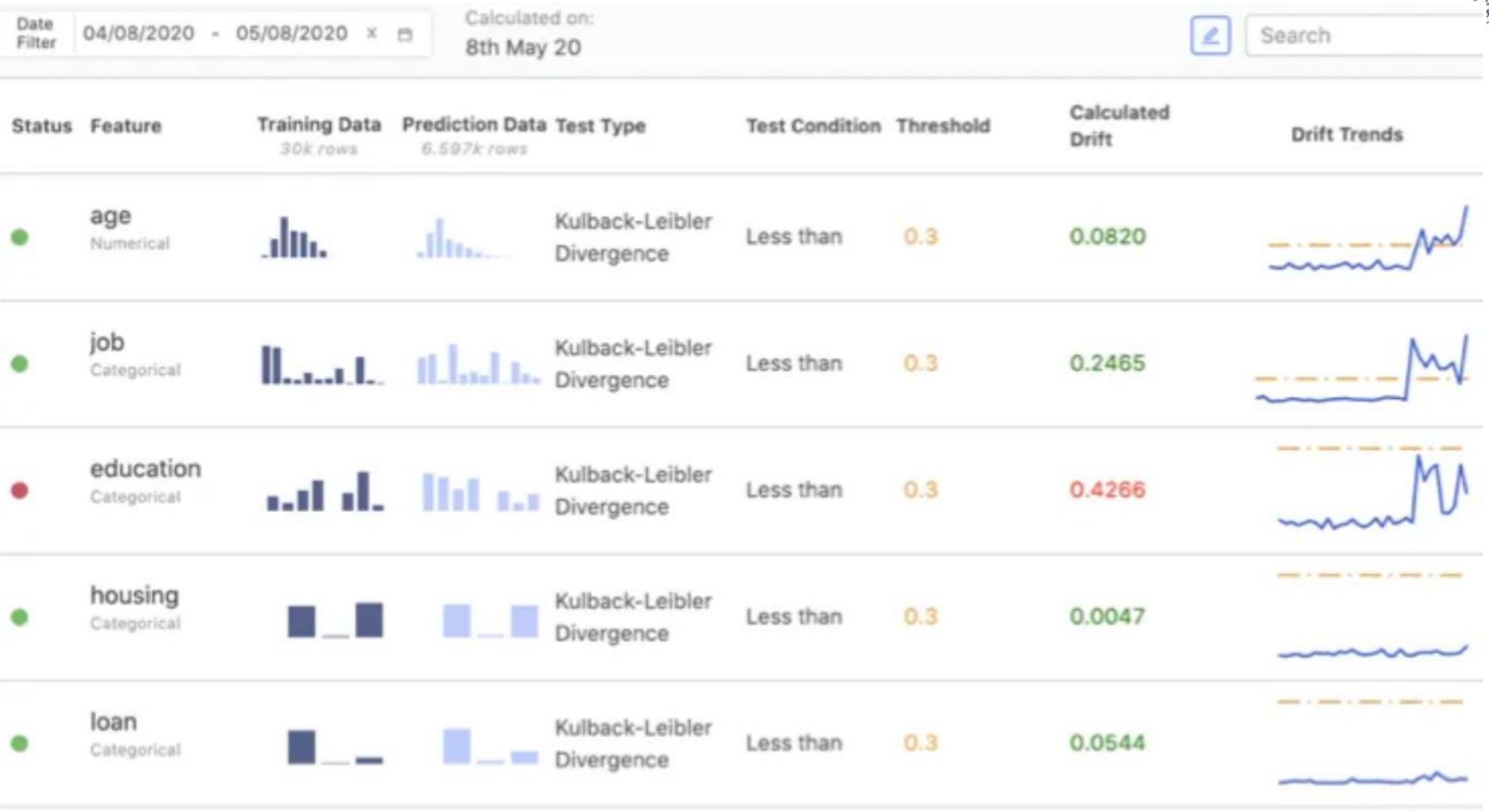
How Organizations can deal with Model degradation?

- Do nothing and wait for it to fail
- Do Ad-hoc drift tests
- Re-train models periodically
- Fix the data pipeline
- Continuous and Standardized monitoring

How to perform Model monitoring?

- Measure drift of independent features
  - **Monitor the statistical features** like distinct values of categorical features, range, histogram, missing values etc.
  - **Monitor data distribution of each feature** using Chi-square test, Kullback Leibler divergence test etc.
- Measure drift of target variable
  - Distribution of target variable
  - Compare the predicted target with actual target
- Continuous monitoring of data pipeline and creating automated alert system for any error

Model Monitoring Interface



Date Filter 04/08/2020 - 05/08/2020

Calculated on: 8th May 20

Search

Status	Feature	Training Data 30k rows	Prediction Data 6.597k rows	Test Type	Test Condition	Threshold	Calculated Drift	Drift Trends
	age Numerical			Kulback-Leibler Divergence	Less than	0.3	0.0820	
	job Categorical			Kulback-Leibler Divergence	Less than	0.3	0.2465	
	education Categorical			Kulback-Leibler Divergence	Less than	0.3	0.4266	
	housing Categorical			Kulback-Leibler Divergence	Less than	0.3	0.0047	
	loan Categorical			Kulback-Leibler Divergence	Less than	0.3	0.0544	

Thank you!

You can reach me on :

Email : shipra@iitism.ac.in

LinkedIn : https://www.linkedin.com/in/shipra-maurya1205/?originalSubdomain=in