TATA CONSULTANCY SERVICES
Experience certainty.

''Event Scry'' – A Data science challenge

Predict the successs, Spot the Outcome, Analyse the Future

10.08.2018

Predict the success

Analyze the future

Spot the outcome

Spot the outcome

Analyze the future

Predict the success

Success or Not, let your model tell us

Contest Starts : 10th August 2018

Prizes worth $ 3,500

Contents

What is Event Scry ?

Scrying (also known by various names such as "seeing" or "peeping") is the practice of looking into a suitable medium in the hope of detecting significant messages or visions.

Opinions about the future are quite easy, but an analytical ability to accurately predict the Future, using data is a challenge that a very few dare to accept

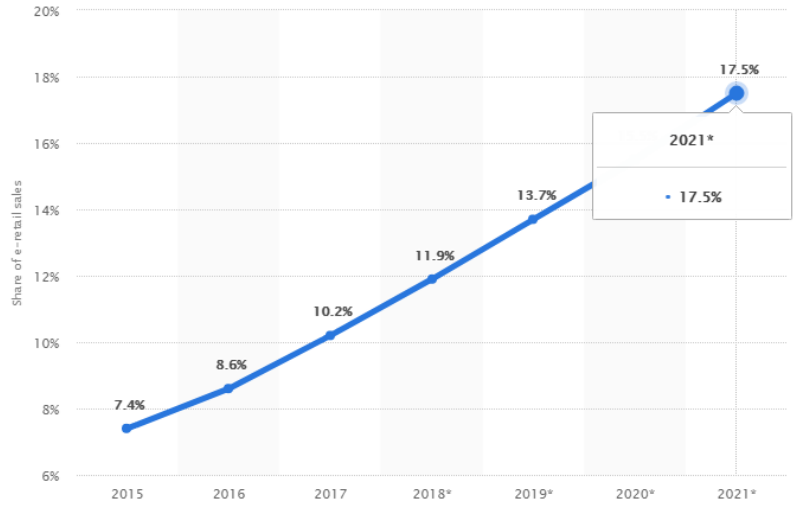
Through the event we are hoping, to find talented Data science enthusiasts in TCS, that are ready to solve customer problems through innovative analytical solutions

The Problem statement !

Growing e-commerce markets – All of us understand the importance of User experience of any Omni-channel application for a service provider, specifically in the field of e-commerce.

The great challenge – As important as it is to have a great e-commerce platform experience, it is also important for the payment solutions to have a seamless experience and Although there are multiple payment market solutions, not all experiences are seamless and user friendly¨

Motto of the hour– Payment solution providers need to design innovative market solutions to provide secure yet simple and frictionless payment options. The focus therefore is on designing an payment solutions with improved customer experience. The emphasis is on 'Instant and frictionless Payments'



Share of e-retail sales to global retail sales

Click here

The Problem - Currently, one of our payment solutions clients, who operates across multiple countries in Europe, is facing a challenge in closing e-commerce sales due to lack of impactful user experience through their payments portal. They are tracking multiple system parameters associated with payment improve their performance.

Thinking brains - Can we TCS'ers help the client by coming up with a model which can accurately predicts whether a given session parameters will lead to a "successful event" or a "non successful event".

What are the Prizes worth ?

Winner – Prizes worth **2000 USD**

First Runner – Prizes worth **1000 USD**

Second Runner – Prizes worth **500 USD**

Key Event rules

- No private sharing outside teams
  - Privately sharing code or data outside of teams is not permitted. It’s okay to share code if made available to all participants on the forum.

- Team limits
  - There is no maximum team size.

- WINNER LICENSE TYPE: Non-Exclusive
  - Competitions are open to employees worldwide from TCS. Other local rules and regulations may apply to you, so please check your local laws to ensure that you are eligible to participate in skills-based competitions.
  - ENTRY IN THIS COMPETITION CONSTITUTES YOUR ACCEPTANCE OF THESE OFFICIAL COMPETITION RULES.

Legal guidelines

WINNER LICENSE -

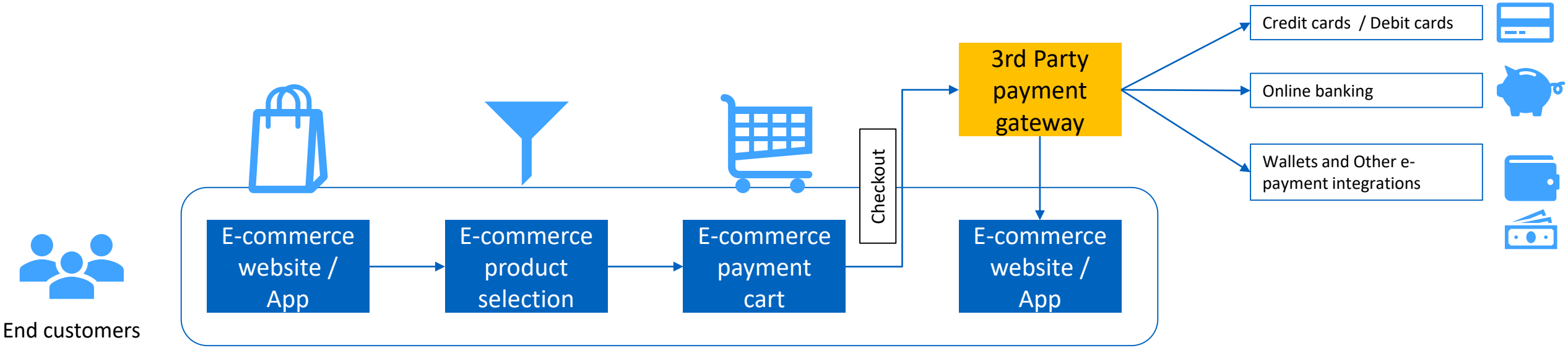
The winners shall grant the program sponsor the following license(s) with respect to your Submission :

- **Non-Exclusive:** You will grant to Competition Sponsor and its designees a worldwide, non-exclusive, sub-licensable, transferable, fully paid-up, royalty-free, perpetual, irrevocable right to use, reproduce, distribute, create derivative works of, publicly perform, publicly display, digitally perform, make, have made, sell, offer for sale and import your winning Submission and the source code used to generate the Submission, in any media now known or hereafter developed, for any purpose whatsoever, commercial or otherwise, without further approval by or payment to Participant.

- **COMPETITION DATA ACCESS USE AND RESTRICTION** After your acceptance of these Competition Rules, you may access and use the Competition Data only for the purposes of the Competition and will need to fully delete the datasets after the competition deadline.
- Data cannot be shared outside TCS, as the content is on ONLY for TCSers

- **EXTERNAL DATA** You may use data, other than the Competition Data, to develop and test your models and Submissions; provided, you have the right and authority to use such external data for the purposes of the Competition, and to share such data with Sponsor.

What is the application all about ? (Illustrative)



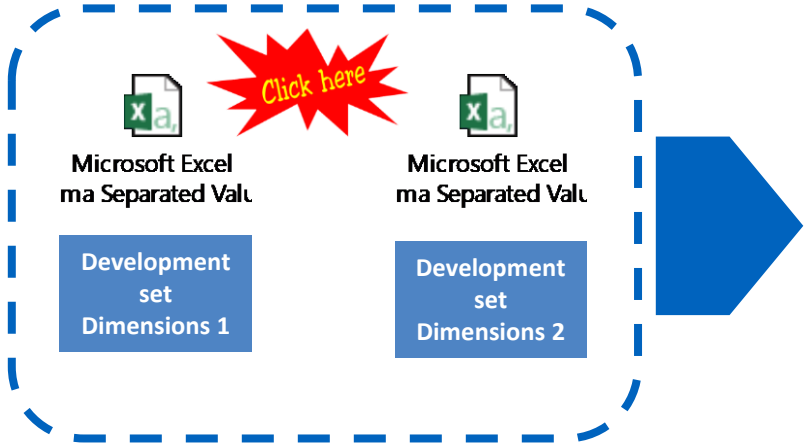
3rd Party payment gateway

- The data that is being shared, is specific to a payment portal, we are tracking the user experience using Google Analytics.

- Payment companies provide payment products to merchants, who sell their products through e-commerce solutions and later provide settlements to the merchants based on the sales of the product

- It is very important for Payment companies to track the user experience of these products, as the User experience of these payment products effect the sales of the merchants and inturn their revenues

Development Data set for the model and Data overview (Development Sample inserted in this slide)



- The Data that is being shared is extracted via Google analytics for the a Payment gateway portal
- We have hidden some of the key parameters of the data (Columns C, D,E,F,G), as they are client specific information, which are primarily Qualitative by nature

Note:

- The data of 2 sets of dimensions that are being extracted is across the same time lines and the same events
- The events across both the dimensions are matching to about 97%, but there is a slight mismatch of the event mapping, if in case you concatenate both the data of Dimension 1 and Dimension 2 data set, using common columns, this is primarily due to challenges with the Google analytics product (You can avoid that mismatch and continue with your model development)
- Your challenge is to create a predictive model which can predict the success of a event with the highest accuracy, leveraging the Dimension data of the several events of the development set.

Your challenge

Event starts here

Development set Dimensions 1

Development set Dimensions 2

Microsoft Excel
ma Separated Valu

Microsoft Excel
ma Separated Valu

- The Development Data set, is a data set comprising of 70% of the random sample of the original data set, for which you need to define the model
- You need to develop your model using the development data set leveraging a choice of methodologies, details guidelines are provided below

Stage 1 and Stage 2 here

Test data set Dimensions 1

Test data set Dimensions 2

- The Test Data set, is a data set comprising of 30% of the random sample of the original data set for which you need to test the model and predict the event success (Test data set, shall not have the event labels)
- You need to test your model using the test data set and share us the model performance, and code, detail guidelines are provided below (We shall share the Test data set on 17th August)

Final stage

- In the final stage, we will conduct a F2F or a Webex based presentation and test your model on a totally new data set (from a different time line)to ensure, your model performance is as per your earlier provided relations
- Based on your choice of the model, approach and the final model performance, winners shall be declared and will have an opportunity to present to the customer

Definitions of the columns in each Dimensions

Dimensions 1	Definition
ga:dateHourMinute	Combined values of date, hour and minute formated as YYYYMMDDHHMM
ga:sessionDurationBucket	The length (returned as a string) of a session measured in seconds and reported in second increments
ga:browser	The name of users' browsers, for example, Internet Explorer or Firefox.
ga:operatingSystem	Users' operating system, for example, Windows, Linux, Macintosh, or iOS.
ga:operatingSystemVersion	The version of users' operating system, i.e., XP for Windows, PPC for Macintosh
ga:language	The language, in ISO-639 code format (e.g., en-gb for British English), provided by the HTTP Request for the browser.
ga:mobileDeviceInfo	The marketing name used for the mobile device
ga:sessionsWithEvent	The total number of sessions with events.
Success	Payment success or Failure
Unique code	Unique code refers to the events with a combination of browser, OS,OS version, language and mobile device info

Dimensions 2	Definitions
ga:dateHourMinute	Combined values of date, hour and minute formated as YYYYMMDDHHMM
ga:userType	A boolean, either New Visitor or Returning Visitor, indicating if the users are new or returning.
ga:sessionCount	The session index for a user. Each session from a unique user will get its own incremental index starting from 1 for the first session. Subsequent sessions do not change previous session indices. For example, if a user has 4 sessions to the website, sessionCount for that user will have 4 distinct values of '1' through '4'.
ga:userBucket	Randomly assigned users tag to allow A/B testing and splitting of remarketing lists. Ranges from 1-100.
ga:sessionDurationBucket	The length (returned as a string) of a session measured in seconds and reported in second increments
ga:sessionsWithEvent	The total number of sessions with events.
ga:uniqueDimensionCombinations	Unique Dimension Combinations counts the number of unique dimension-value combinations for each dimension in a report
Sucess	Payment success or Failure

Model results template and guidelines

Details to be shared with the Idea owners

Team details, along with employee Id, Ph.no and Photo (Optional)

Ravindranath S
690993
(ravindranath.s@tcs.com)
Ph: +47-41257885

Model attributes used, along with technique used and why ?

Refer Slide 15

Model performance in comparision to Dev sample and Test sample

Refer Slide 16

The Solution code for testing your model on our dataset

Insert the code into a PPT

Note: Ensure the final presentation should be below 10 slides (Including Overview slide and Thank you slide)

Share us the details of the Attributes that you have used to build the model

- Attribute 1
- Attribute 2
- Attribute 3
- Attribute 4
- Attribute 5
- Attribute 6
- Attribute 7
- Attribute 8
- Attribute 9
- Attribute 10

Predicted Score

Min Score	Max Score	Segments
0.65	0.95	Highly Preferred segment
0.45	0.65	Moderately Preferred segment
0.25	0.45	Risky Segment
0.15	0.25	High Risk Segment

Input → Model → Output

Techniques (Illustrative)

Dummy Variable Method	**Each attribute is categorized into smaller sub-groups by methods known as fine classing and coarse classing. All but one of these sub-groups so derived are then evaluated as potential predictors.**
Weight of evidence Method	**Each attribute is categorized into smaller sub-groups and for every sub-group weight of evidence defined as log(subgroup odds/sample odds) is computed. The attribute is then replaced by it's grouped version as a predictor.**
Ensemble Techniques	**Predictions made by multiple models are combined using ensemble techniques like democracy and autocracy . These result in improved accuracy and robustness of predictions over time.**

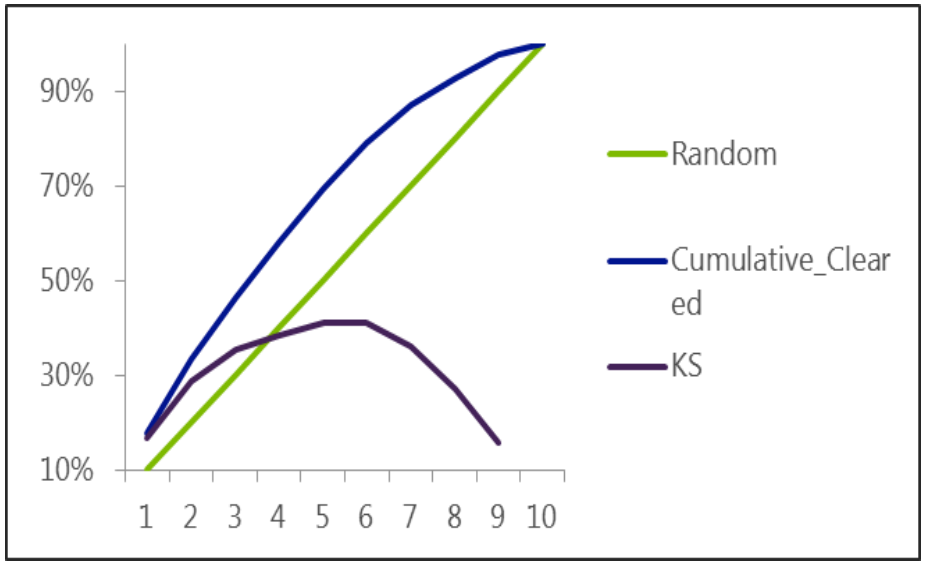
Share us the values of how your model is performing with the development set and test data set

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	Significance Level
(Intercept)					
Attribute 1					
Attribute 2					
Attribute 3					
Attribute 4					
Attribute 5					
Attribute 6					
Attribute 7					
Attribute 8					
Attribute 9					
Attribute 10					
Attribute 11					
Attribute 12					
Attribute 13					
Attribute 14					
Attribute 15					

Association of Predicted Probabilities and Observed Responses

Percent Concordant	Somers' D
Percent Discordant	Gamma
Percent Tied	Tau-a
Pairs	c



Confusion Matrix

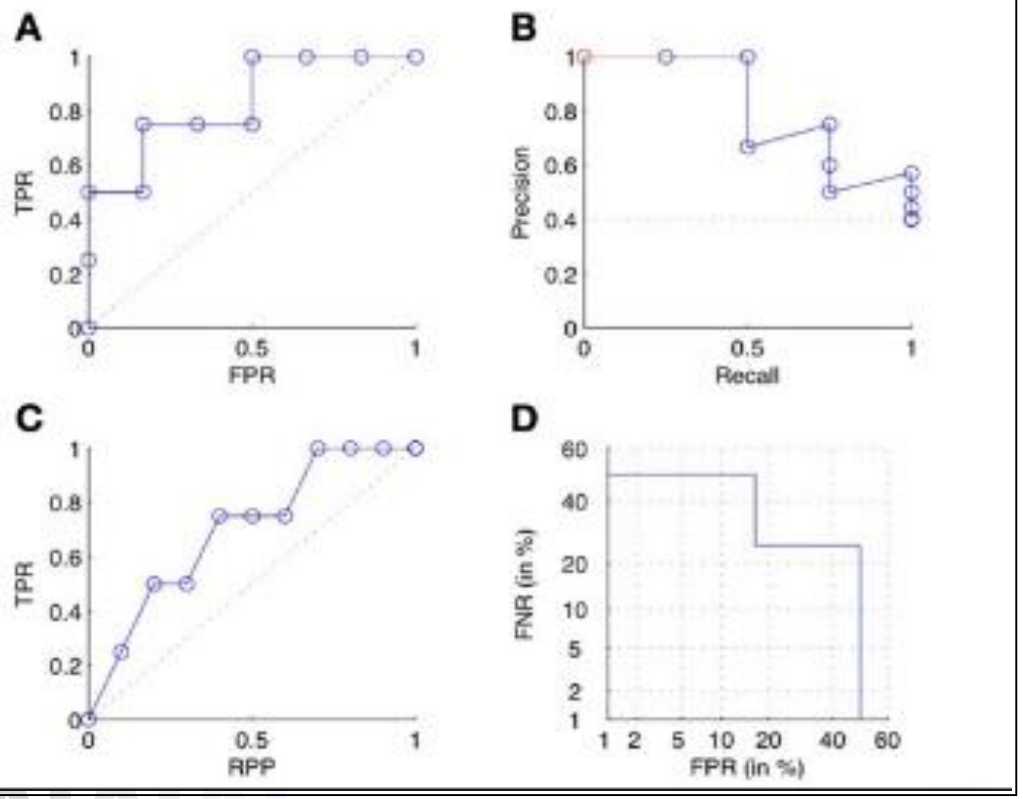
	FALSE	TRUE
0		
1		

Some models that you can use

Logistic Regression Output

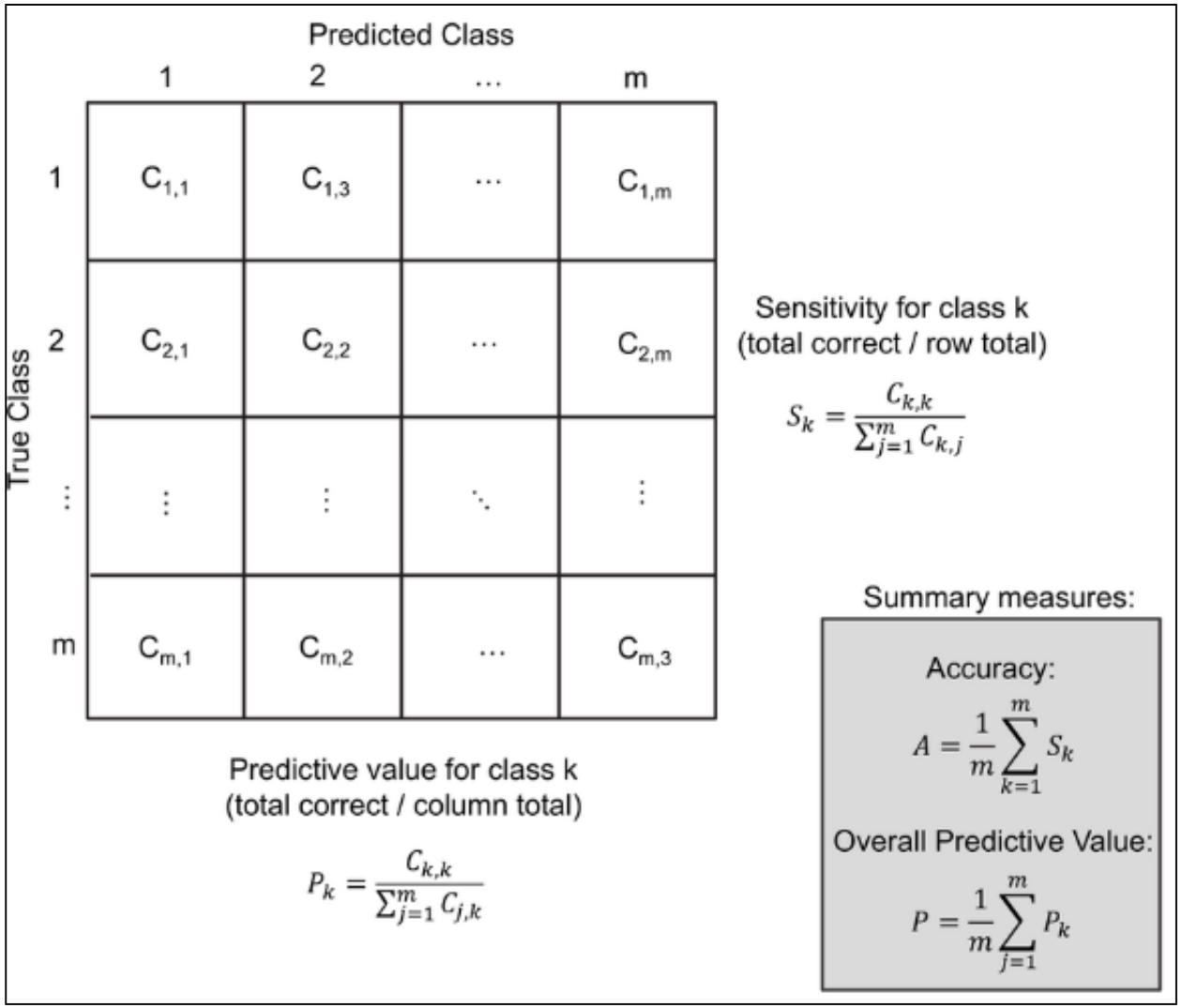
Sorted predictions	0.91	0.86	0.85	0.57	0.54	0.26	0.18	0.16	0.14	0.13
Rank	1	2	3	4	5	6	7	8	9	10
Actual values	**1**	**1**	0	**1**	0	0	**1**	0	0	0
FPR	0	0	1/6	1/6	1/3	1/2	1/2	2/3	5/6	1
TPR/Recall	1/4	1/2	1/2	3/4	3/4	3/4	1	1	1	1
Precision	1	1	2/3	3/4	3/5	1/2	4/7	1/2	4/9	2/5

Constants: $P = 4, N = 6$



Source: https://www.researchgate.net/figure/ROC-curve-A-precision-recall-curve-B-lift-chart-C-and-DET-curve-D-for-the_fig2_259354565

Random Forest Output



Source: https://www.researchgate.net/figure/Example-confusion-matrix-for-an-m-class-classification-problem-Ci-j-denotes-the-number_fig3_250920710?_sg=9QJz8cWUFanJEsnpUO3prrWjmEQXQXJbiBWZ-khoxbl0NEA7KGw1xsA8pAjztZ0f1xjfUH8wIO2H78j_aFbyWw

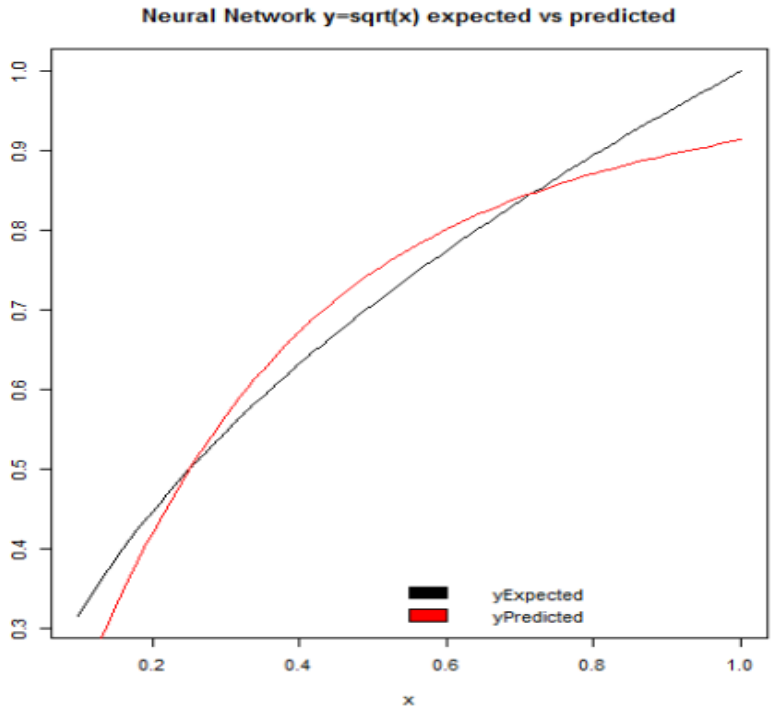
Some models that you can use

SVM Output (Representative)

Statistics

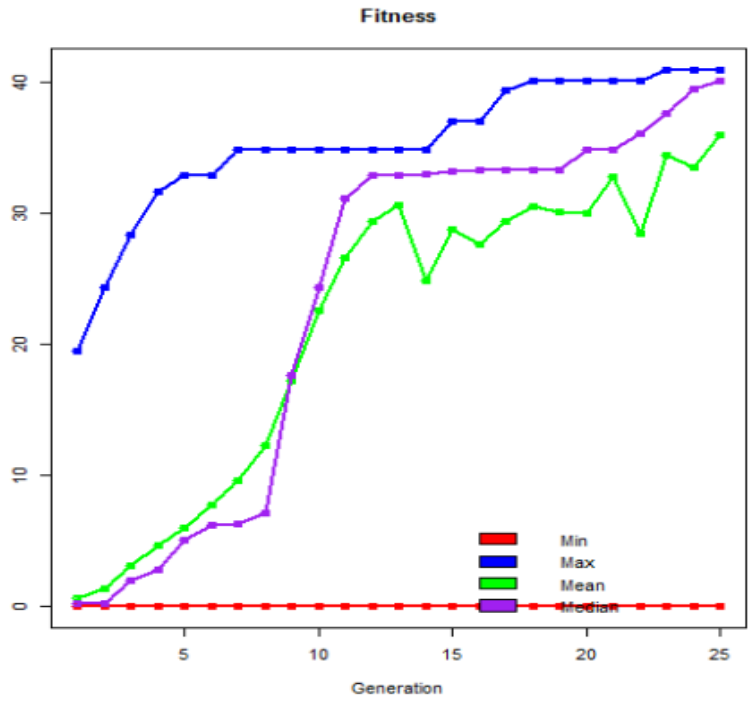
Accuracy	0.9246
95% CI	(0.9123, 0.9654)
No Information Rate	0.5849
p -Value [Acc > NIR]	<2e-16
Kappa	0.8452
Mcnemar's Test P-Value	0.02498
Sensitivity	0.9854
Specificity	0.8754
Pos Pred Value	0.9125
Neg Pred Value	0.9236
Prevalence	0.6421
Detection Rate	0.5896
Balanced Accuracy	0.9019
Positive' Class	0

Neural Network Output (Representative)



Performance Chart

The performance of the network can be seen in the bottom left chart of the image above, there is considerable differences between the expected output and the actual output.



Iteration Comparison

The maximum, mean and median fitness are generally increasing with each generation.

- Additionally, in case SVM or Neural network method have been used, team needs to provide Rationale behind method, libraries and architecture along with justification.
- If the competing teams consider learning rules then appropriate reasoning should be provided

Evaluation criteria

The models shall be evaluated based on the following criteria –

Stability – Score distributional shift between development and test samples needs to be measured and shifts beyond 10% needs to be eliminated

Strength – Model ROC Curve / Gain charts and Lift charts needs to be demonstrated and models can be ranked on the basis of capture / lift metrics

Rank ordering and Goodness of fit – The models needs to demonstrate the monotonicity of the scores and relative comparision between actual and predicted curves should be the lowest

Attribute significance and directional impacts – Attributes should be significant at 95% confidence level and the directional signs should not be counter intuitive and should be in sync between development and validation samples

Key dates of the events

Organizers to share
- Development data set
- Guidelines for participation
- Rules and regulations regarding the event and the data
- Templates to provide results

Event start date

10th Aug

Participants to share
- Interest to participate in the event (Event registration)
- Team details

Organizers to share
- Test data set without labels
- Guidelines to generate the labels of the test data set
- Expected output template

Stage 1

17th Aug

Organizers to share
- Final ranks of the participants based on their model performance and result of the test data set

Stage 2

24th Aug

Participants to share
- Code of the final model
- Output of the test data set as per expected output

Organizers to share
- Final ranks of the participants
- Top 5 rankers shall be invited to the finale

Final stage

27-31st Aug

FAQ Blog for Event Scry

https://knome.ultimatix.net/blogposts/443617-event-scry-a-data-science-challenge-discussion-forum

Please use this blog as a discussion forum for all of us to discuss !

Comment your queries here and we shall respond back at the earliest.

Contact details

Primary contacts

Ravindranath S
(ravindranath.s@tcs.com)
Ph: +47-41257885

Chandersekher Joshi
(cs.joshi@tcs.com)
Ph:+47-96715439

Jury Team

Lipika Dey
(Chief Data Scientist, CTO)

Dibyendu Mukherjee
(Data scientist, A&I)

Manoj Apte
(Data scientist, CTO)

TATA CONSULTANCY SERVICES
Experience certainty.

TATA

Thank You

IT Services
Business Solutions
Consulting