intuit turbotax quickbooks mint

Exploratory Data Analysis Demo
(Use Case: MOOC dropout prediction)

Feb 09, 2019

Naveen Kumar Kaveti, Data Scientist
Sravya Garapati, Machine Learning Engineer
Viswa Datha Polavarapu, Machine Learning Engineer

Soumya Sulegai, Talent Acquisition Mgr
Priyanka A Giri, CW Talent Acquisition

Agenda

Introduction to Intuit

Prerequisites

Problem Statement

Data Understanding

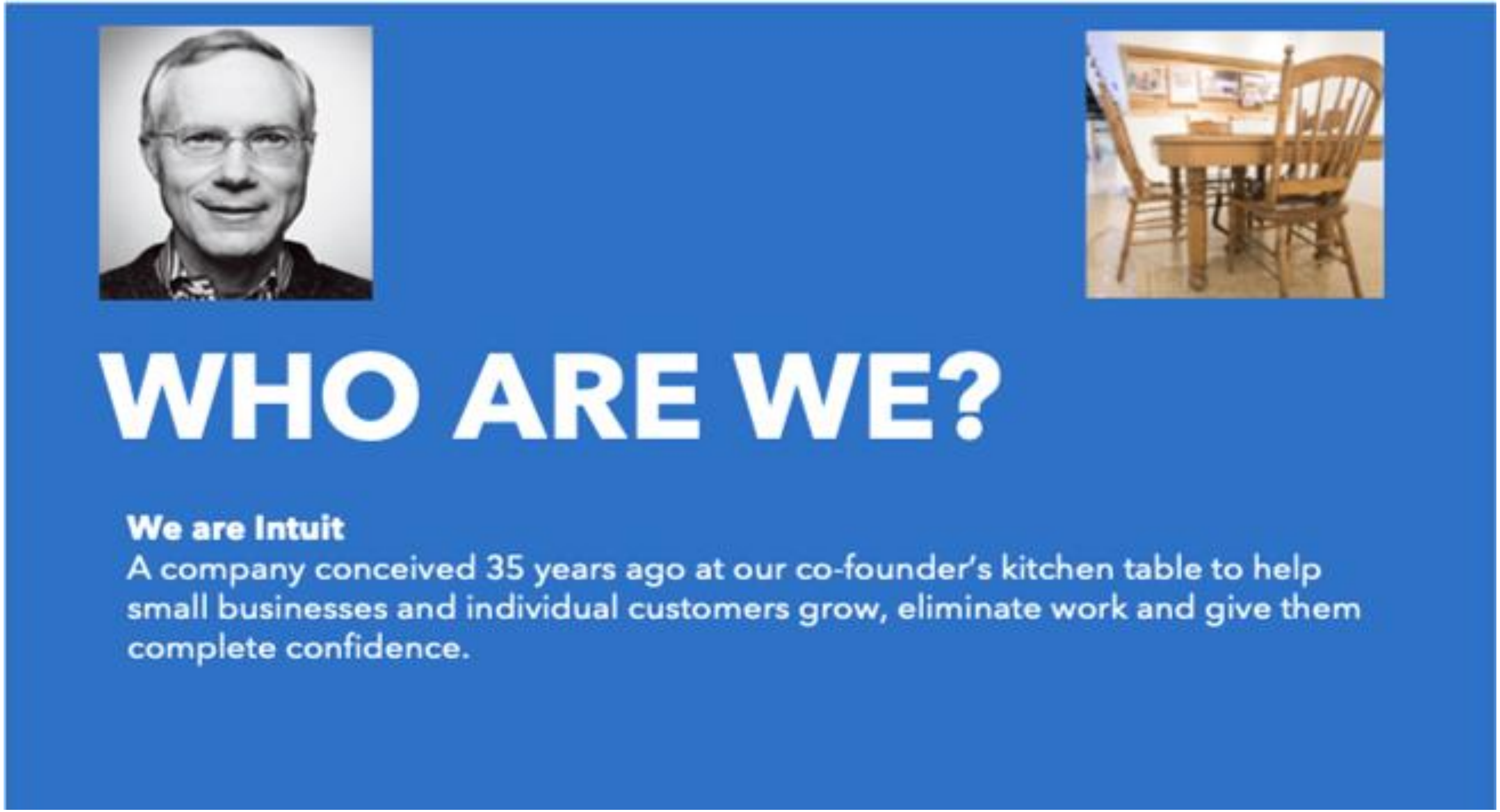
Feature Engineering

EDA (Exploratory Data Analysis)

Model Building

Demo Time

Challenge Time

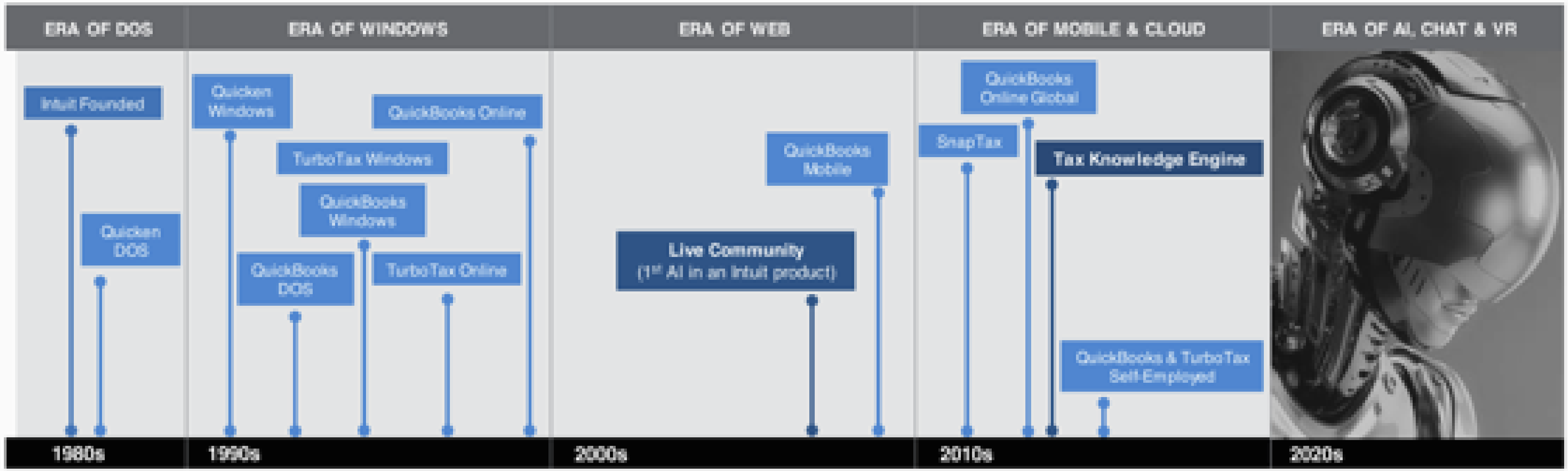


WHO ARE WE?

We are Intuit
A company conceived 35 years ago at our co-founder's kitchen table to help small businesses and individual customers grow, eliminate work and give them complete confidence.

Our Mission

Our journey so far



ERA OF DOS

Intuit Founded

Quicken DOS

1980s

ERA OF WINDOWS

Quicken Windows

TurboTax Windows

QuickBooks Windows

QuickBooks DOS

TurboTax Online

QuickBooks Online

1990s

ERA OF WEB

Live Community (1st AI in an Intuit product)

QuickBooks Mobile

2000s

ERA OF MOBILE & CLOUD

SnapTax

QuickBooks Online Global

Tax Knowledge Engine

QuickBooks & TurboTax Self-Employed

2010s

ERA OF AI, CHAT & VR

2020s

CUSTOMER-OBSESSED • DESIGN-INSPIRED • TECHNOLOGY-POWERED

Products that power prosperity

Our technology has helped us innovate four of our major products that are simplifying work of millions, worth millions.

50M

CUSTOMERS

$6B

COMPANY

intuit quickbooks.

intuit turbotax.

intuit mint.

Agenda

Introduction to Intuit

Prerequisites

Problem Statement

Data Understanding

Feature Engineering

EDA (Exploratory Data Analysis)

Model Building

Demo Time

Challenge Time

Prerequisites

What is distribution?

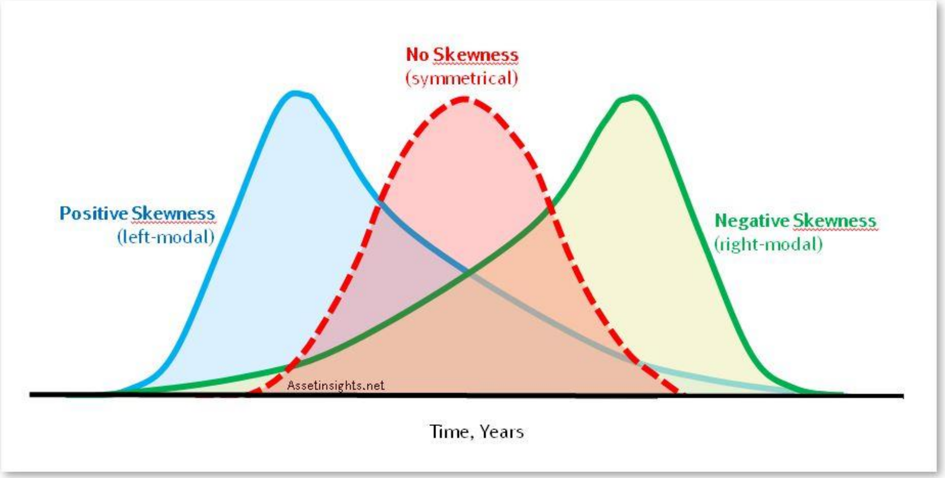
What are the properties of distribution?

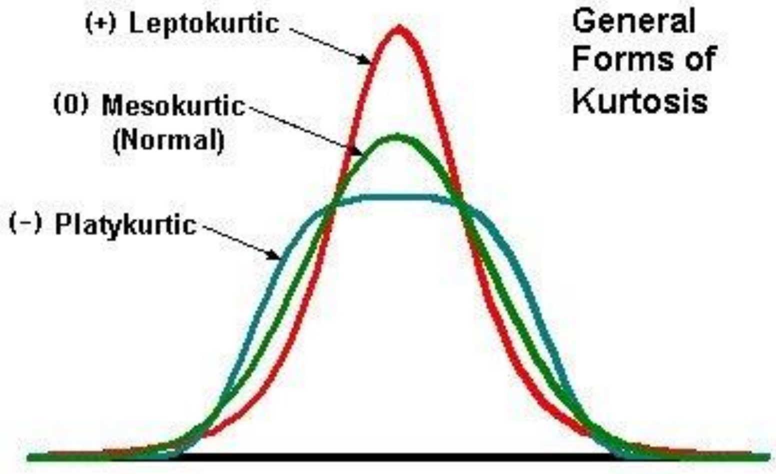
Mean

Variance

Skewness

Kurtosis





Prerequisites

Correlations:

Pearson’s Correlation Coefficient - Measure of the linear correlation between two variables X and Y

$$\rho_{X,Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}$$

Spearman’s Rank Correlation Coefficient - Measures the monotonic relationship between two variables

$$r_s = 1 - \frac{6 \sum_i d_i^2}{n(n^2 - 1)}$$

Mutual Information - Measures the amount of information flow between two variables

$$\mathrm{I}(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \frac{p(x,y)}{p(x)\,p(y)} \qquad \frac{\mathrm{I}(X;Y)}{\mathrm{H}(X) + \mathrm{H}(Y)}$$

Agenda

Introduction to Intuit

Prerequisites

Problem Statement

Data Understanding

Feature Engineering

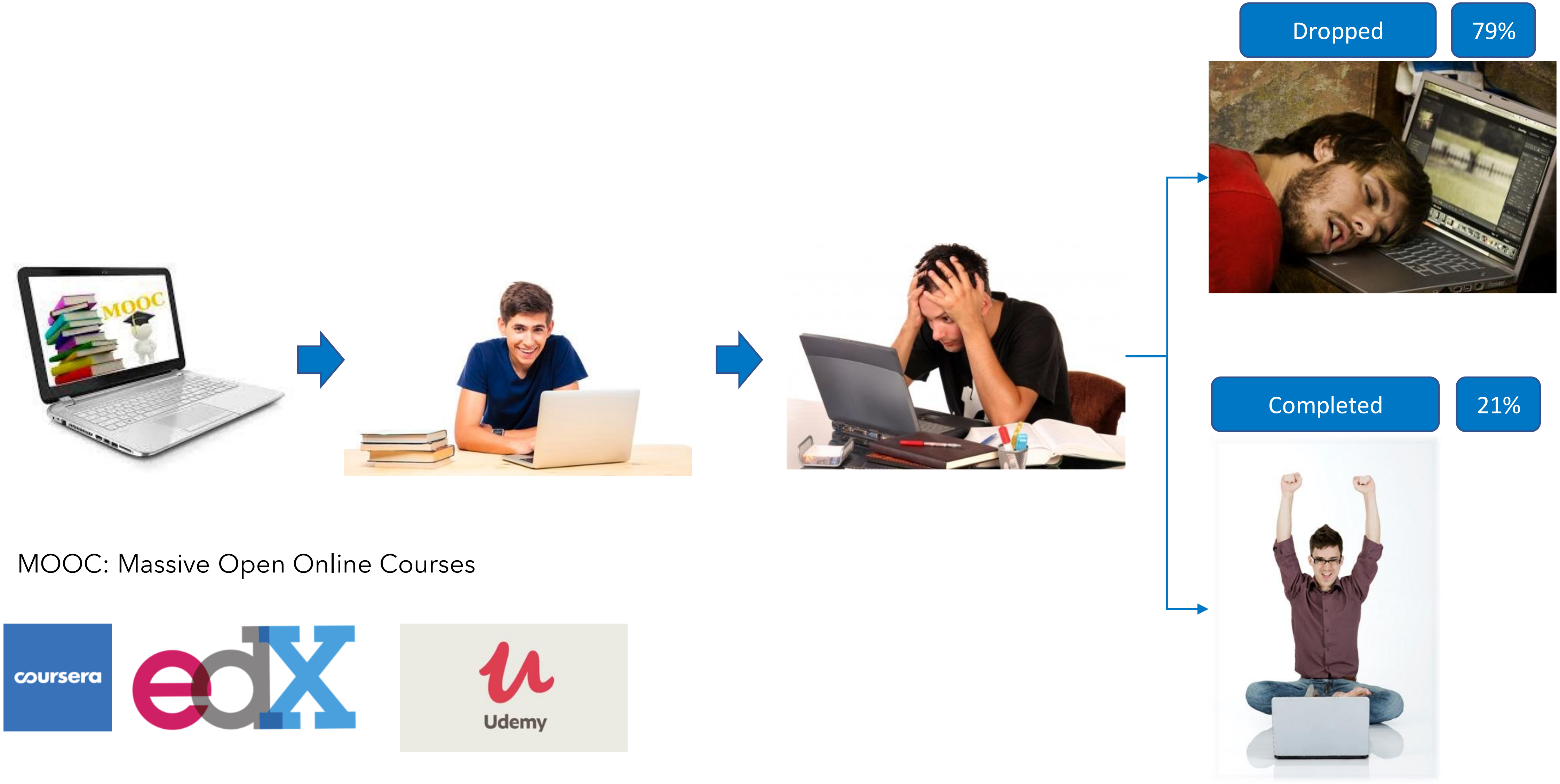
EDA (Exploratory Data Analysis)

Model Building

Demo Time

Challenge Time

Problem Statement



MOOC: Massive Open Online Courses

Problem Statement

The Challenge:

The competition participants need **to predict whether a user will drop a course within next 10 days based on his or her prior activities**. If a user C leaves no records for course C in the log during the next 10 days, we define it as dropout from course C.

But Why?

Students' high dropout rate on MOOC platforms has been heavily criticized, and predicting their likelihood of dropout would be useful **for maintaining and encouraging students' learning activities**.

学堂在线
xuetangx.com

Reference: http://moocdata.cn/challenges/kdd-cup-2015

Agenda

Introduction to Intuit

Prerequisites

Problem Statement

Data Understanding

Feature Engineering

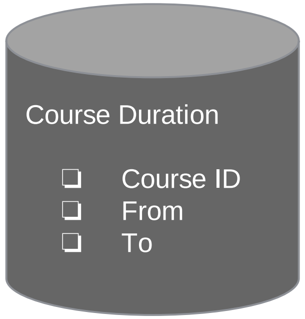
EDA (Exploratory Data Analysis)

Model Building

Demo Time

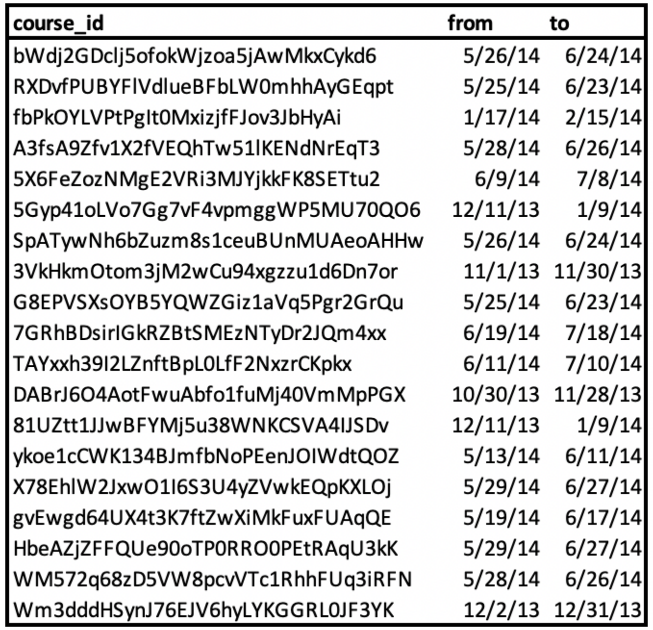
Challenge Time

Data Understanding - Course Level Information



Course Duration

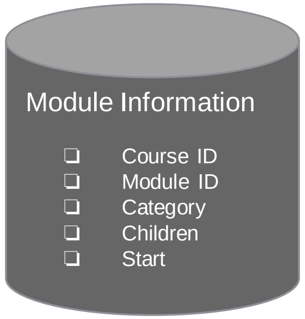
- ❑ Course ID
- ❑ From
- ❑ To



course_id	from	to
bWdj2GDclj5ofokWjzoa5jAwMkxCykd6	5/26/14	6/24/14
RXDvfPUBYFlVdlueBFbLW0mhhAyGEqpt	5/25/14	6/23/14
fbPkOYLVPtPgIt0MxizjfFJov3JbHyAi	1/17/14	2/15/14
A3fsA9Zfv1X2fVEQhTw51lKENdNrEqT3	5/28/14	6/26/14
5X6FeZozNMgE2VRi3MJYjkkFK8SETtu2	6/9/14	7/8/14
5Gyp41oLVo7Gg7vF4vpmggWP5MU70QO6	12/11/13	1/9/14
SpATywNh6bZuzm8s1ceuBUnMUAeoAHHw	5/26/14	6/24/14
3VkHkmOtom3jM2wCu94xgzzu1d6Dn7or	11/1/13	11/30/13
G8EPVSXsOYB5YQWZGiz1aVq5Pgr2GrQu	5/25/14	6/23/14
7GRhBDsirlGkRZBtSMEzNTyDr2JQm4xx	6/19/14	7/18/14
TAYxxh39I2LZnftBpL0LfF2NxzrCKpkx	6/11/14	7/10/14
DABrJ6O4AotFwuAbfo1fuMj40VmMpPGX	10/30/13	11/28/13
81UZtt1JJwBFYMj5u38WNKCSVA4IJSDv	12/11/13	1/9/14
ykoe1cCWK134BJmfbNoPEenJOIWdtQOZ	5/13/14	6/11/14
X78EhlW2JxwO1I6S3U4yZVwkEQpKXLOj	5/29/14	6/27/14
gvEwgd64UX4t3K7ftZwXiMkFuxFUAqQE	5/19/14	6/17/14
HbeAZjZFFQUe90oTP0RRO0PEtRAqU3kK	5/29/14	6/27/14
WM572q68zD5VW8pcvVTc1RhhFUq3iRFN	5/28/14	6/26/14
Wm3dddHSynJ76EJV6hyLYKGGRL0JF3YK	12/2/13	12/31/13

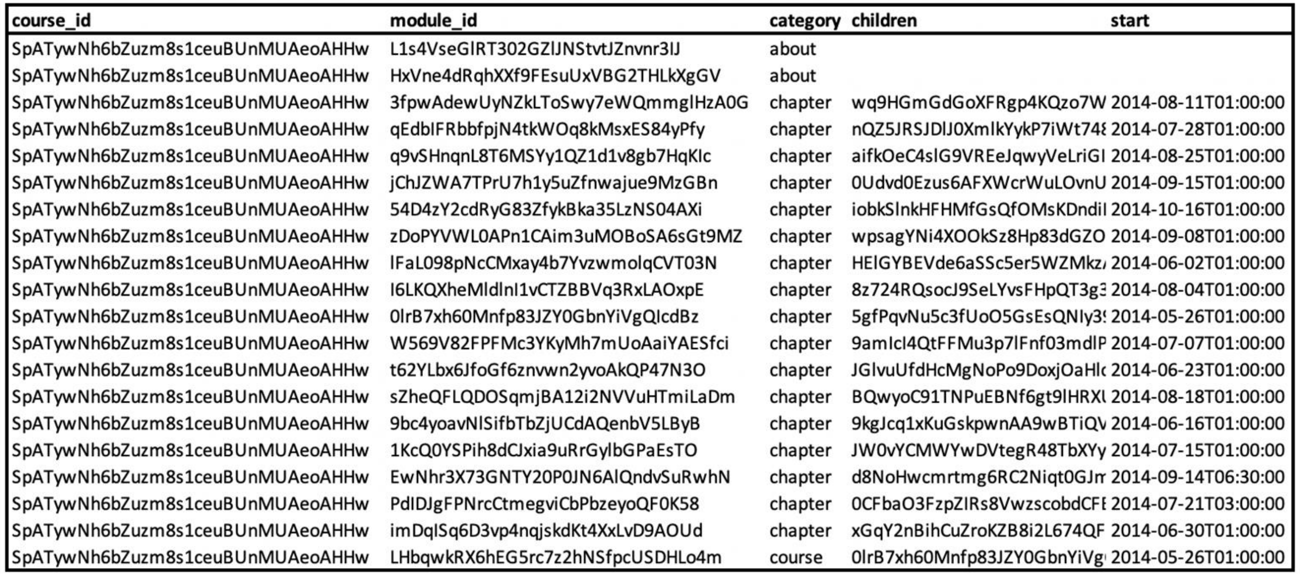
Description:

Each line contains the timespan of each course (both train and test data).



Module Information

- ❑ Course ID
- ❑ Module ID
- ❑ Category
- ❑ Children
- ❑ Start

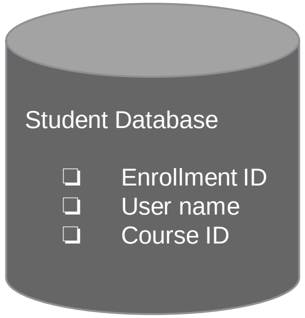


course_id	module_id	category	children	start
SpATywNh6bZuzm8s1ceuBUnMUAeoAHHw	L1s4VseGlRT302GZlJNStvtJZnvnr3lJ	about		
SpATywNh6bZuzm8s1ceuBUnMUAeoAHHw	HxVne4dRqhXXf9FEsuUxVBG2THLkXgGV	about		
SpATywNh6bZuzm8s1ceuBUnMUAeoAHHw	3fpwAdewUyNZkLToSwy7eWQmmglHzA0G	chapter	wq9HGmGdGoXFRgp4KQzo7W	2014-08-11T01:00:00
SpATywNh6bZuzm8s1ceuBUnMUAeoAHHw	qEdblFRbbfpjN4tkWOq8kMsxES84yPfy	chapter	nQZ5JRSJDlJ0XmlkYykP7iWt74	2014-07-28T01:00:00
SpATywNh6bZuzm8s1ceuBUnMUAeoAHHw	q9vSHnqnL8T6MSYy1QZ1d1v8gb7HqKlc	chapter	aifkOeC4slG9VREeJqwyVeLriGI	2014-08-25T01:00:00
SpATywNh6bZuzm8s1ceuBUnMUAeoAHHw	jChJZWA7TPrU7h1y5uZfnwajue9MzGBn	chapter	0Udvd0Ezus6AFXWcrWuLOvnU	2014-09-15T01:00:00
SpATywNh6bZuzm8s1ceuBUnMUAeoAHHw	54D4zY2cdRyG83ZfykBka35LzNS04AXi	chapter	iobkSlnkHFHMfGsQfOMsKDndil	2014-10-16T01:00:00
SpATywNh6bZuzm8s1ceuBUnMUAeoAHHw	zDoPYVWL0APn1CAim3uMOBoSA6sGt9MZ	chapter	wpsagYNi4XOOkSz8Hp83dGZO	2014-09-08T01:00:00
SpATywNh6bZuzm8s1ceuBUnMUAeoAHHw	lFaL098pNcCMxay4b7YvzwmolqCVT03N	chapter	HElGYBEVde6aSSc5er5WZMkz	2014-06-02T01:00:00
SpATywNh6bZuzm8s1ceuBUnMUAeoAHHw	I6LKQXheMldlnI1vCTZBBVq3RxLAOxpE	chapter	8z724RQsocJ9SeLYvsFHpQT3g	2014-08-04T01:00:00
SpATywNh6bZuzm8s1ceuBUnMUAeoAHHw	0lrB7xh60Mnfp83JZY0GbnYiVgQlcdBz	chapter	5gfPqvNu5c3fUoO5GsEsQNIy3	2014-05-26T01:00:00
SpATywNh6bZuzm8s1ceuBUnMUAeoAHHw	W569V82FPFMc3YKyMh7mUoAaiYAESfci	chapter	9amIcl4QtFFMu3p7lFnf03mdlP	2014-07-07T01:00:00
SpATywNh6bZuzm8s1ceuBUnMUAeoAHHw	t62YLbx6JfoGf6znvwn2yvoAkQP47N3O	chapter	JGlvuUfdHcMgNoPo9DoxjOaHl	2014-06-23T01:00:00
SpATywNh6bZuzm8s1ceuBUnMUAeoAHHw	sZheQFLQDOSqmjBA12i2NVVuHTmiLaDm	chapter	BQwyoC91TNPuEBNf6gt9lHRX	2014-08-18T01:00:00
SpATywNh6bZuzm8s1ceuBUnMUAeoAHHw	9bc4yoavNlSifbTbZjUCdAQenbV5LByB	chapter	9kgJcq1xKuGskpwnAA9wBTiQ	2014-06-16T01:00:00
SpATywNh6bZuzm8s1ceuBUnMUAeoAHHw	1KcQ0YSPih8dCJxia9uRrGylbGPaEsTO	chapter	JW0vYCMWYwDVtegR48TbXYy	2014-07-15T01:00:00
SpATywNh6bZuzm8s1ceuBUnMUAeoAHHw	EwNhr3X73GNTY20P0JN6AlQndvSuRwhN	chapter	d8NoHwcmrtmg6RC2Niqt0GJ	2014-09-14T06:30:00
SpATywNh6bZuzm8s1ceuBUnMUAeoAHHw	PdIDJgFPNrcCtmegviCbPbzeyoQF0K58	chapter	0CFbaO3FzpZIRs8VwzscobdCF	2014-07-21T03:00:00
SpATywNh6bZuzm8s1ceuBUnMUAeoAHHw	imDqlSq6D3vp4nqjskdKt4XxLvD9AOUd	chapter	xGqY2nBihCuZroKZB8i2L674QF	2014-06-30T01:00:00
SpATywNh6bZuzm8s1ceuBUnMUAeoAHHw	LHbqwkRX6hEG5rc7z2hNSfpcUSDHLo4m	course	0lrB7xh60Mnfp83JZY0GbnYiVg	2014-05-26T01:00:00

Description:

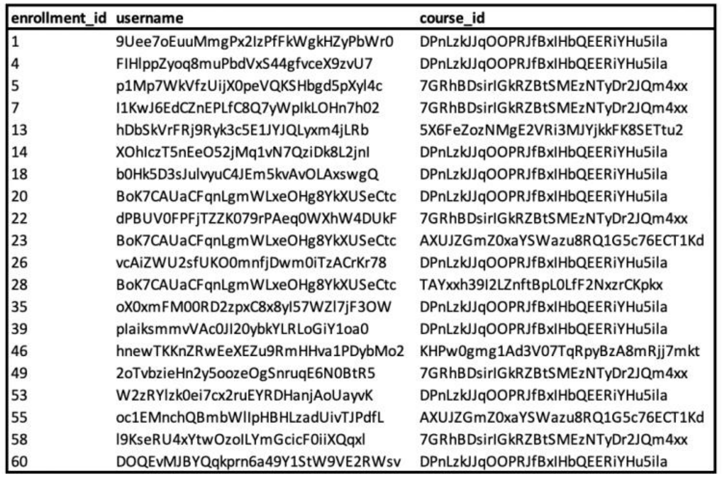
Each line in this file describes a module in a course with its category, children objects and release time.

Data Understanding - Enrollment Level Information



Student Database

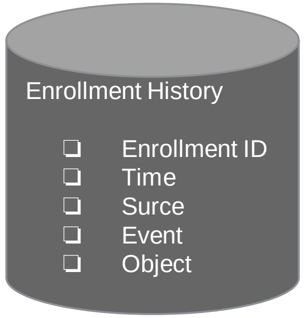
- ❑ Enrollment ID
- ❑ User name
- ❑ Course ID



enrollment_id	username	course_id
1	9Uee7oEuuMmgPx2IzPfFkWgkHZyPbWr0	DPnLzkJJqOOPRJfBxIHbQEERiYHu5ila
4	FIHlppZyoq8muPbdVxS44gfvceX9zvU7	DPnLzkJJqOOPRJfBxIHbQEERiYHu5ila
5	p1Mp7WkVfzUijX0peVQKSHbgd5pXyl4c	7GRhBDsirIGkRZBtSMEzNTyDr2JQm4xx
7	I1KwJ6EdCZnEPLfC8Q7yWpIkLOHn7h02	7GRhBDsirIGkRZBtSMEzNTyDr2JQm4xx
13	hDbSkVrFRj9Ryk3c5E1JYJQLyxm4jLRb	5X6FeZozNMgE2VRi3MJYjkkFK8SETtu2
14	XOhIczT5nEeO52jMq1vN7QziDk8L2jnI	DPnLzkJJqOOPRJfBxIHbQEERiYHu5ila
18	b0Hk5D3sJulvyuC4JEm5kvAvOLAxswgQ	DPnLzkJJqOOPRJfBxIHbQEERiYHu5ila
20	BoK7CAUaCFqnLgmWLxeOHg8YkXUSeCtc	DPnLzkJJqOOPRJfBxIHbQEERiYHu5ila
22	dPBUV0FPFjTZZK079rPAeq0WXhW4DUkF	7GRhBDsirIGkRZBtSMEzNTyDr2JQm4xx
23	BoK7CAUaCFqnLgmWLxeOHg8YkXUSeCtc	AXUJZGmZ0xaYSWazu8RQ1G5c76ECT1Kd
26	vcAiZWU2sfUKO0mnfjDwm0iTzACrKr78	DPnLzkJJqOOPRJfBxIHbQEERiYHu5ila
28	BoK7CAUaCFqnLgmWLxeOHg8YkXUSeCtc	TAYxxh39I2LZnftBpL0LfF2NxzrCKpkx
35	oX0xmFM00RD2zpxC8x8yI57WZl7jF3OW	DPnLzkJJqOOPRJfBxIHbQEERiYHu5ila
39	plaiksmmvVAc0JI20ybkYLRLoGiY1oa0	DPnLzkJJqOOPRJfBxIHbQEERiYHu5ila
46	hnewTKKnZRwEeXEZu9RmHHva1PDybMo2	KHPw0gmg1Ad3V07TqRpyBzA8mRjj7mkt
49	2oTvbzieHn2y5oozeOgSnruqE6N0BtR5	7GRhBDsirIGkRZBtSMEzNTyDr2JQm4xx
53	W2zRYlzk0ei7cx2ruEYRDHanjAoUayvK	DPnLzkJJqOOPRJfBxIHbQEERiYHu5ila
55	oc1EMnchQBmbWlIpHBHLzadUivTJPdfL	AXUJZGmZ0xaYSWazu8RQ1G5c76ECT1Kd
58	I9KseRU4xYtwOzolLYmGcicF0iiXQqxl	7GRhBDsirIGkRZBtSMEzNTyDr2JQm4xx
60	DOQEvMJBYQqkprn6a49Y1StW9VE2RWsv	DPnLzkJJqOOPRJfBxIHbQEERiYHu5ila

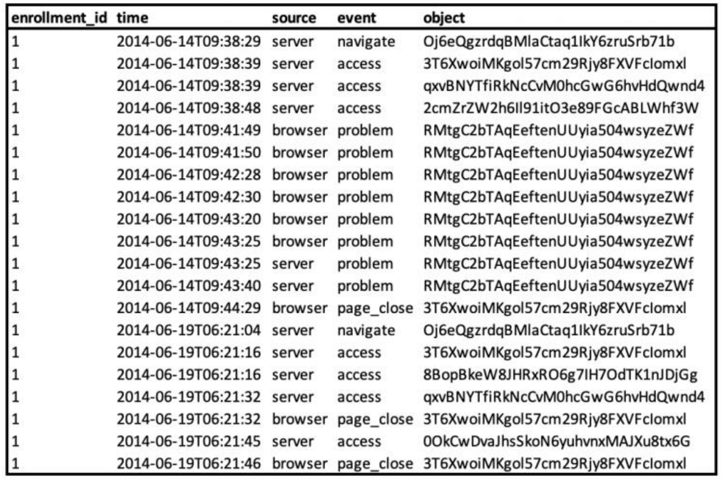
Description:

Each line is a course enrollment record with an enrollment id, a username U and a course id C, indicating that U enrolled in course C.



Enrollment History

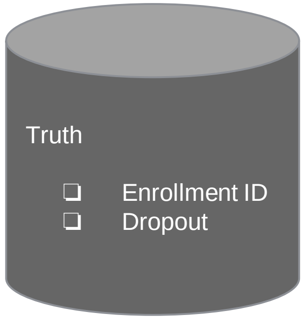
- ❑ Enrollment ID
- ❑ Time
- ❑ Surce
- ❑ Event
- ❑ Object



enrollment_id	time	source	event	object
1	2014-06-14T09:38:29	server	navigate	Oj6eQgzrdqBMlaCtaq1IkY6zruSrb71b
1	2014-06-14T09:38:39	server	access	3T6XwoiMKgol57cm29Rjy8FXVFclomxl
1	2014-06-14T09:38:39	server	access	qxvBNYTfiRkNcCvM0hcGwG6hvHdQwnd4
1	2014-06-14T09:38:48	server	access	2cmZrZW2h6Il91itO3e89FGcABLWhf3W
1	2014-06-14T09:41:49	browser	problem	RMtgC2bTAqEeftenUUyia504wsyzeZWf
1	2014-06-14T09:41:50	browser	problem	RMtgC2bTAqEeftenUUyia504wsyzeZWf
1	2014-06-14T09:42:28	browser	problem	RMtgC2bTAqEeftenUUyia504wsyzeZWf
1	2014-06-14T09:42:30	browser	problem	RMtgC2bTAqEeftenUUyia504wsyzeZWf
1	2014-06-14T09:43:20	browser	problem	RMtgC2bTAqEeftenUUyia504wsyzeZWf
1	2014-06-14T09:43:25	browser	problem	RMtgC2bTAqEeftenUUyia504wsyzeZWf
1	2014-06-14T09:43:25	server	problem	RMtgC2bTAqEeftenUUyia504wsyzeZWf
1	2014-06-14T09:43:40	server	problem	RMtgC2bTAqEeftenUUyia504wsyzeZWf
1	2014-06-14T09:44:29	browser	page_close	3T6XwoiMKgol57cm29Rjy8FXVFclomxl
1	2014-06-19T06:21:04	server	navigate	Oj6eQgzrdqBMlaCtaq1IkY6zruSrb71b
1	2014-06-19T06:21:16	server	access	3T6XwoiMKgol57cm29Rjy8FXVFclomxl
1	2014-06-19T06:21:16	server	access	8BopBkeW8JHRxRO6g7IH7OdTK1nJDjGg
1	2014-06-19T06:21:32	server	access	qxvBNYTfiRkNcCvM0hcGwG6hvHdQwnd4
1	2014-06-19T06:21:32	browser	page_close	3T6XwoiMKgol57cm29Rjy8FXVFclomxl
1	2014-06-19T06:21:45	server	access	0OkCwDvaJhsSkoN6yuhvnxMAJXu8tx6G
1	2014-06-19T06:21:46	browser	page_close	3T6XwoiMKgol57cm29Rjy8FXVFclomxl

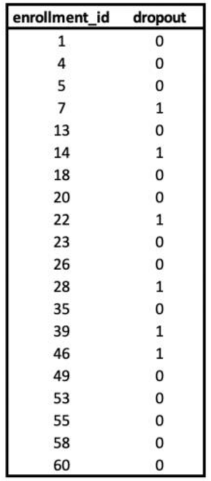
Description:

Each line is an action taken by a user within an enrollment.



Truth

- ❑ Enrollment ID
- ❑ Dropout

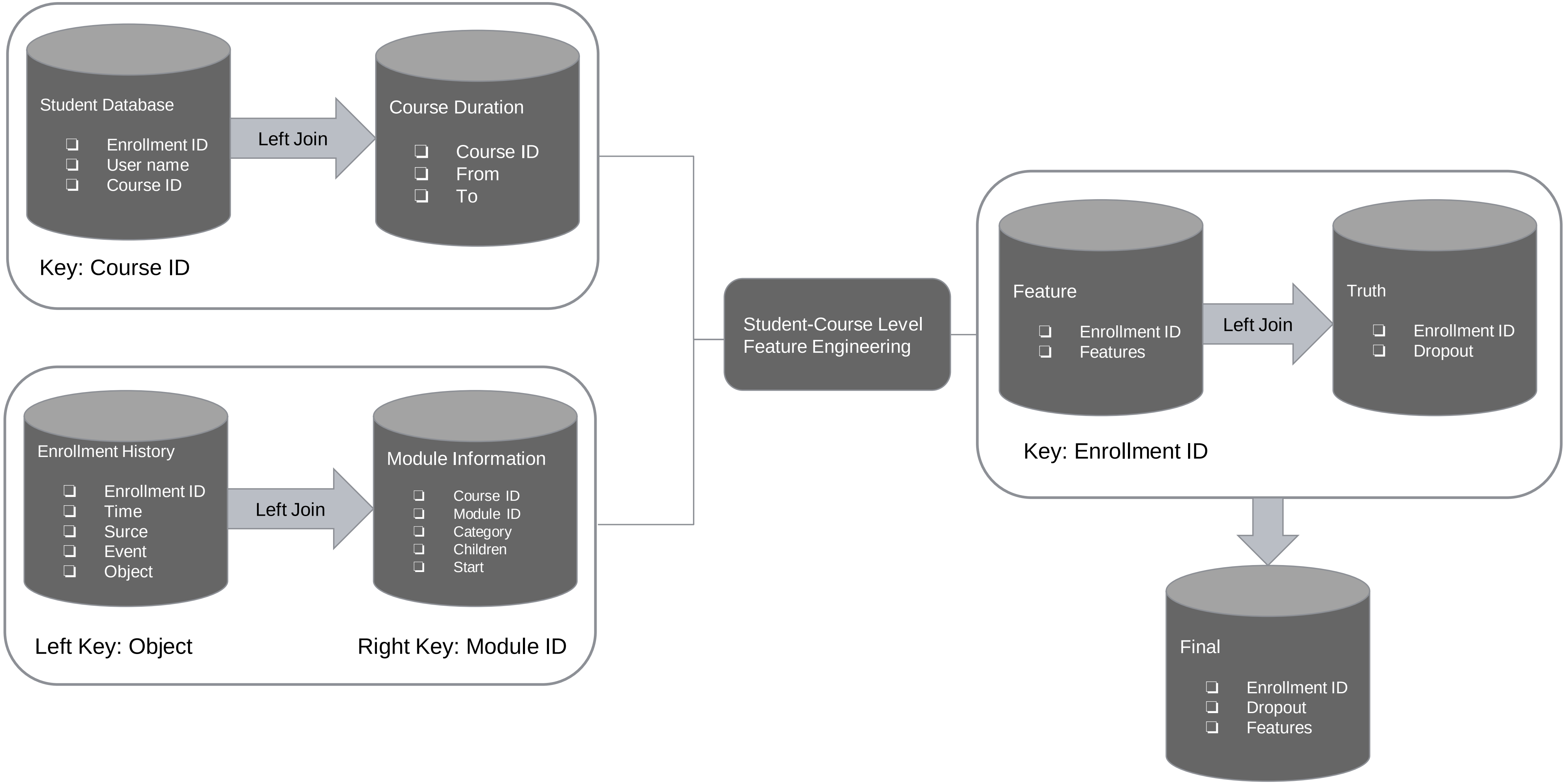


enrollment_id	dropout
1	0
4	0
5	0
7	1
13	0
14	1
18	0
20	0
22	1
23	0
26	0
28	1
35	0
39	1
46	1
49	0
53	0
55	0
58	0
60	0

Description:

Each line contains information about the ground truth of enrollments in the training set.

Data Understanding



Student Database

- ❑ Enrollment ID
- ❑ User name
- ❑ Course ID

Left Join

Course Duration

- ❑ Course ID
- ❑ From
- ❑ To

Key: Course ID

Enrollment History

- ❑ Enrollment ID
- ❑ Time
- ❑ Surce
- ❑ Event
- ❑ Object

Left Join

Module Information

- ❑ Course ID
- ❑ Module ID
- ❑ Category
- ❑ Children
- ❑ Start

Left Key: Object

Right Key: Module ID

Student-Course Level Feature Engineering

Feature

- ❑ Enrollment ID
- ❑ Features

Left Join

Truth

- ❑ Enrollment ID
- ❑ Dropout

Key: Enrollment ID

Final

- ❑ Enrollment ID
- ❑ Dropout
- ❑ Features

Agenda

Introduction to Intuit

Prerequisites

Problem Statement

Data Understanding

Feature Engineering

EDA (Exploratory Data Analysis)

Model Building

Demo Time

Challenge Time

Feature Engineering

User Level Features

- ❑ Number of courses enrolled
- ❑ Lifetime of the user

Course Level Features

- ❑ Number of users enrolled
- ❑ Dropout percentage
- ❑ Average delay between chapter start times

Enrollment Level Features

- ❑ Average delay between chapter complete times
- ❑ Event (Problem, Video and Discussion) counts
- ❑ Event (Problem, Video and Discussion) duration

Agenda

Introduction to Intuit

Prerequisites

Problem Statement

Data Understanding

Feature Engineering

EDA (Exploratory Data Analysis)

Model Building

Demo Time

Challenge Time

EDA (Exploratory Data Analysis)

Make a Hypothesis

Test a Hypothesis

Testing of Hypothesis (Two Sample t-test)

Step1:

Null Hypothesis (Make an hypothesis about population): Mean of two samples are equal ($\mu_1 = \mu_2$)

Alternative Hypothesis (Negate Null Hypothesis): Mean of two samples are not equal ($\mu_1 \neq \mu_2$)

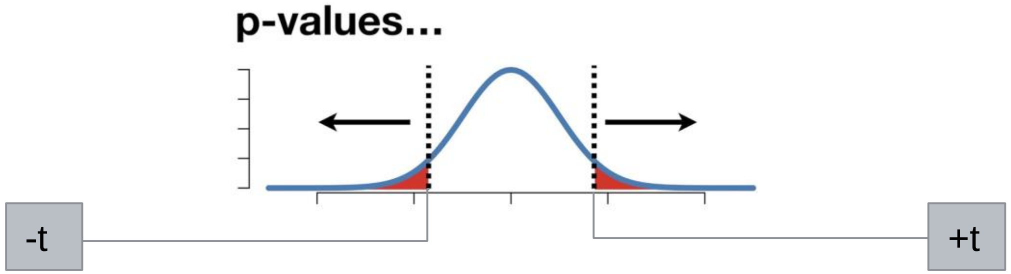
Step 2:

Test the hypothesis about population using available data

$$t = \frac{|\overline{X_1} - \overline{X_2}|}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

Step 3:

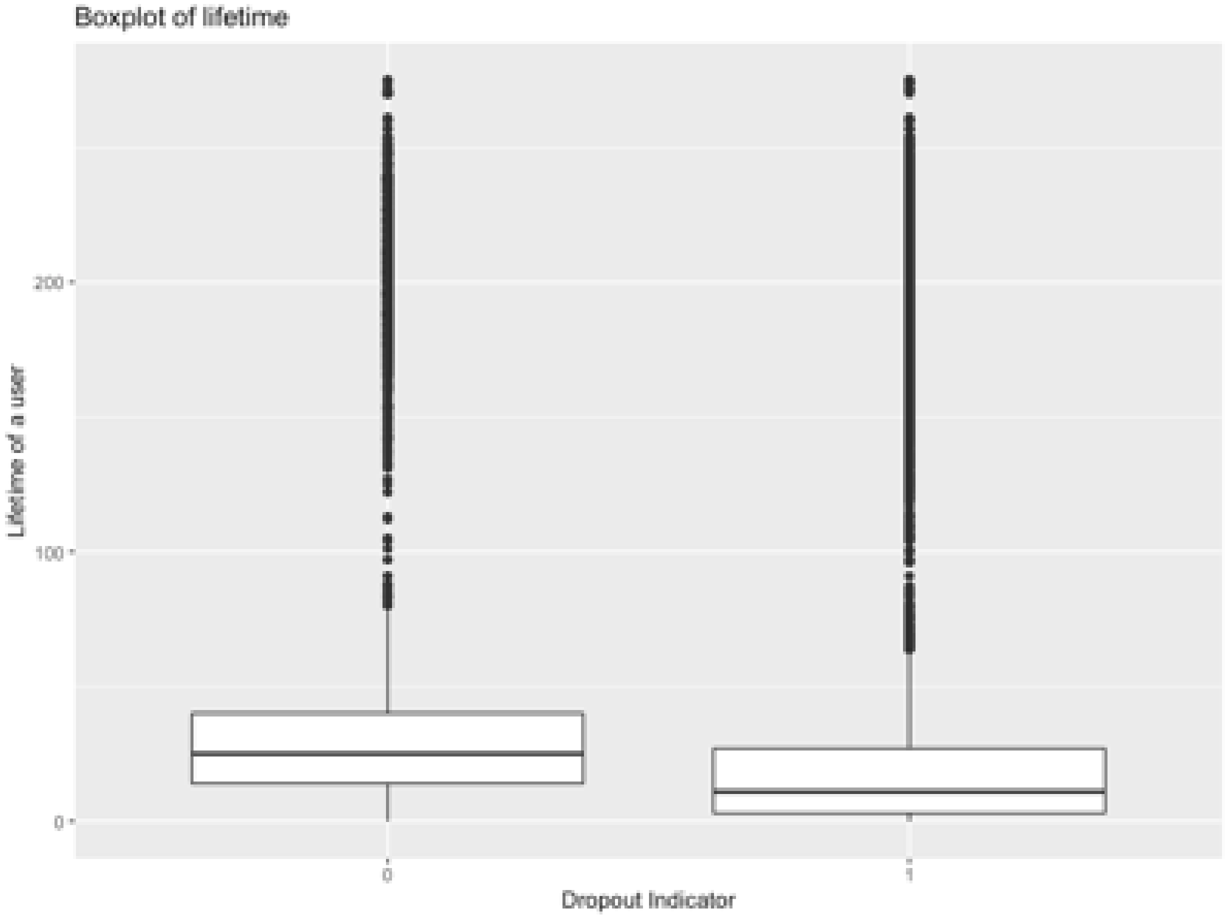
Compute p-value based on t-statistic

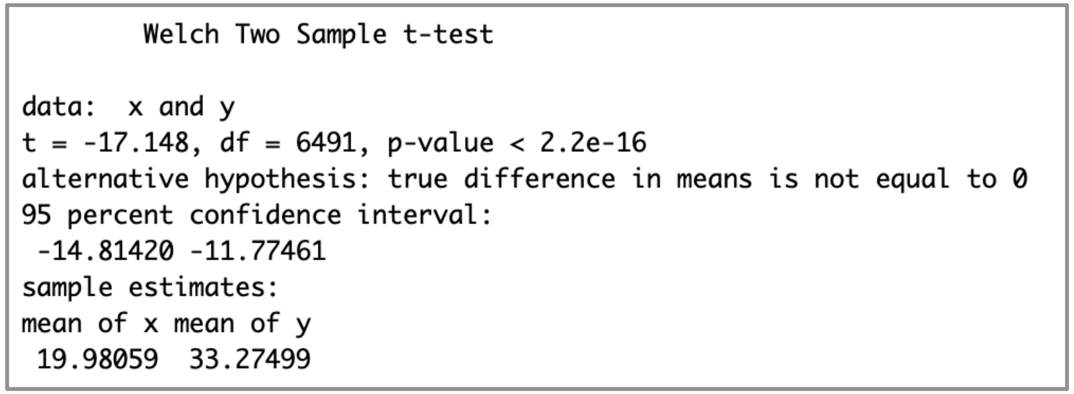


Step 4: Compare p-value with the assumed level of significance (say, 0.05) and reject the null hypothesis if p-value is less than 0.05 and fail to reject the null hypothesis if p-value is greater than 0.05

EDA (Exploratory Data Analysis)

Hypothesis: Does lifetime of user impacts the user’s willingness to complete the course?



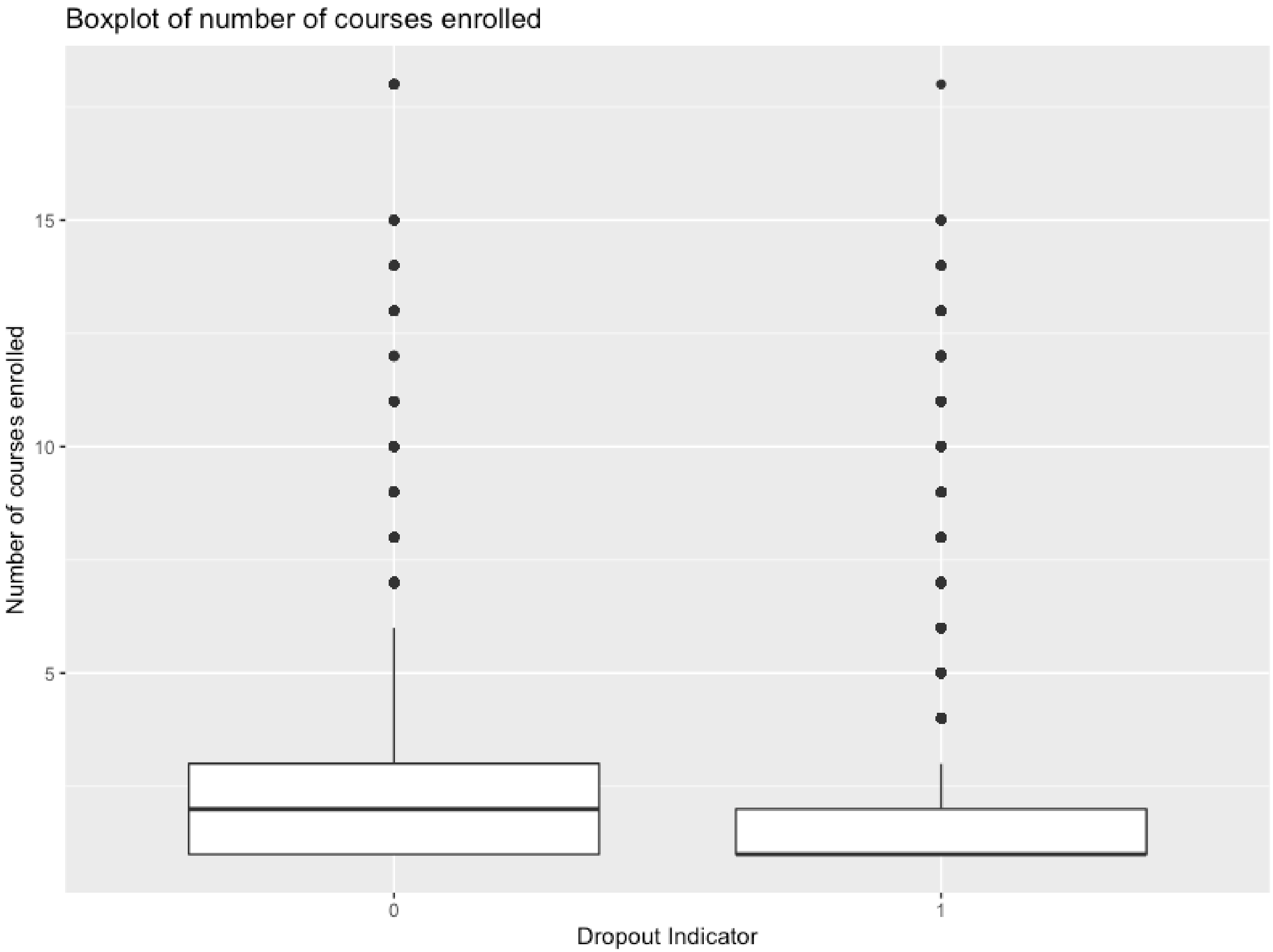


```
        Welch Two Sample t-test

data:  x and y
t = -17.148, df = 6491, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -14.81420 -11.77461
sample estimates:
mean of x mean of y
 19.98059  33.27499
```

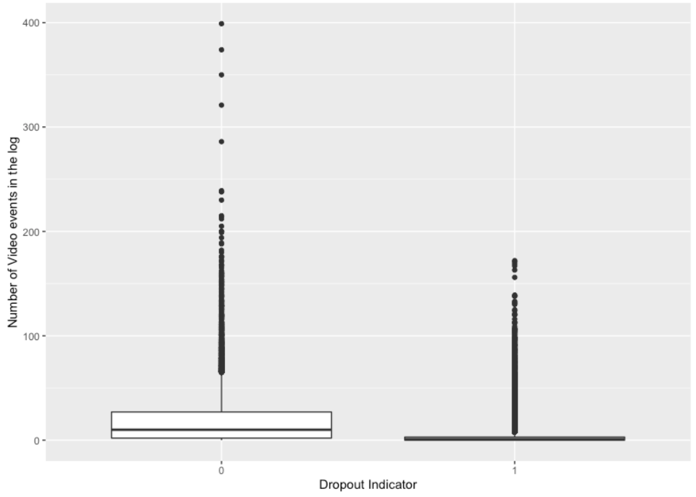
EDA (Exploratory Data Analysis)

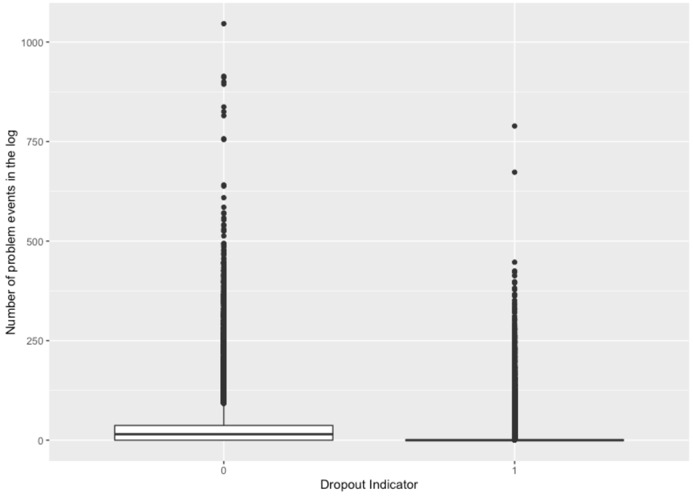
Hypothesis: Does number of courses enrolled by the user impact the user's willingness to complete the course?

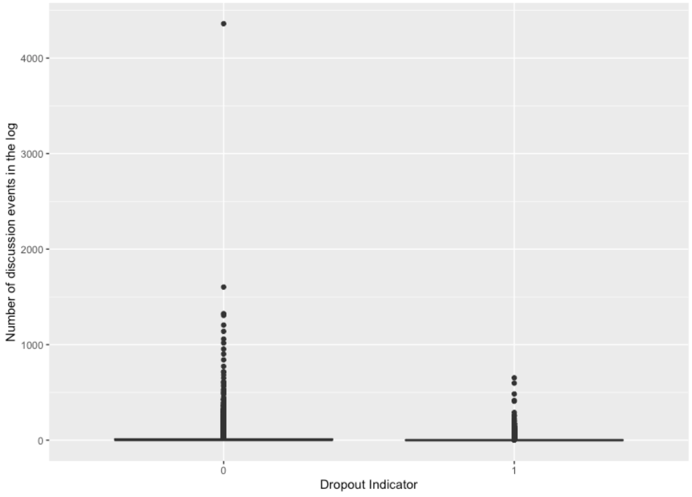


EDA (Exploratory Data Analysis)

Hypothesis: Does event (problem/video/discussion) counts impact the user’s willingness to complete the course?


Boxplot of number of video events


Boxplot of number of problem events


Boxplot of number of discussion events

t = -43.033; p-value = < 2.2e-16

Mean of x = 3.46; Mean of y = 18.78

Conclusion: The difference in means is not equals to 0

t = -31.896; p-value = < 2.2e-16

Mean of x = 4.93; Mean of y = 33

Conclusion: The difference in means is not equals to 0

t = -14.87; p-value = < 2.2e-16

Mean of x = 2.07; Mean of y = 18.14

Conclusion: The difference in means is not equals to 0

Agenda

Introduction to Intuit

Prerequisites

Problem Statement

Data Understanding

Feature Engineering

EDA (Exploratory Data Analysis)

Model Building

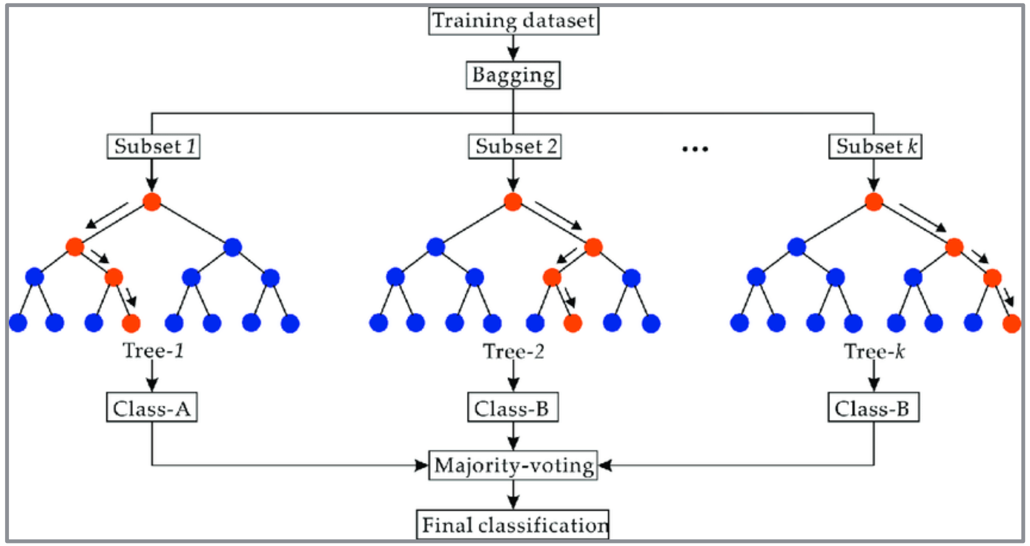
Demo Time

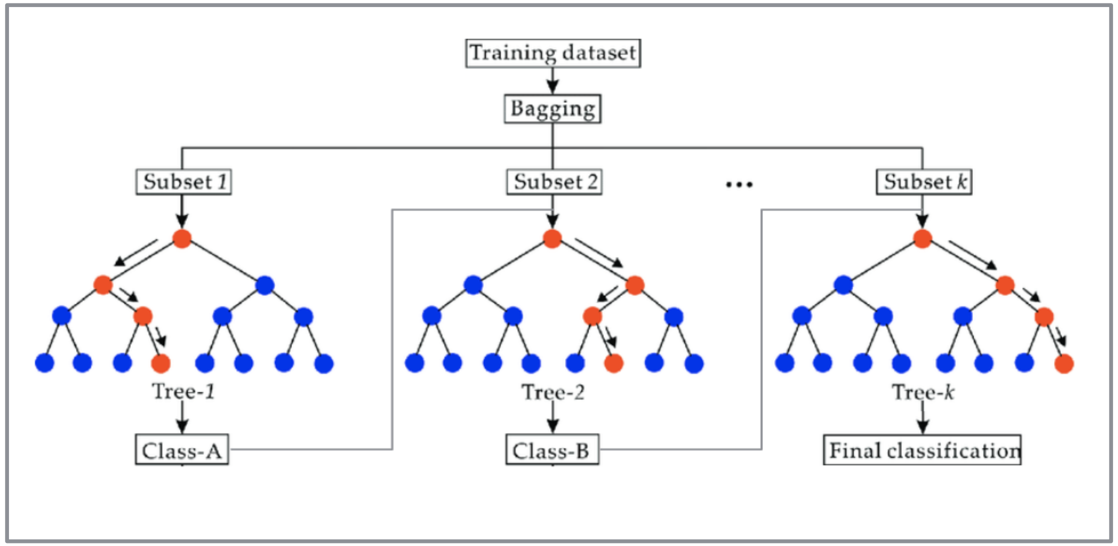
Challenge Time

Bagging Vs Boosting

Bagging (Parallel)

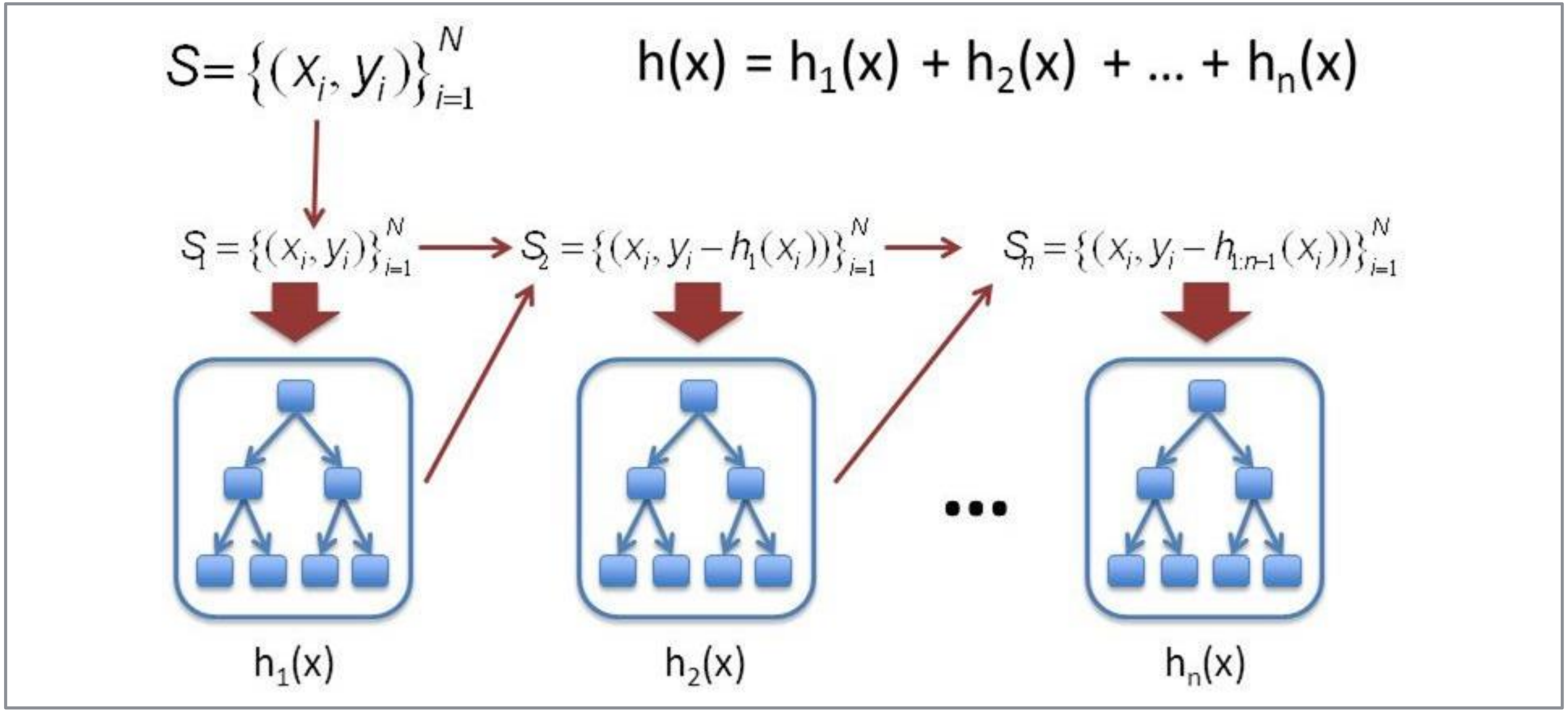
Boosting (Sequential)





Reference: GIS-based mineral prospectivity mapping using machine learning methods: A case study from Tongling ore district, eastern China

Gradient Boost Machine



$$S = \{(x_i, y_i)\}_{i=1}^{N}$$

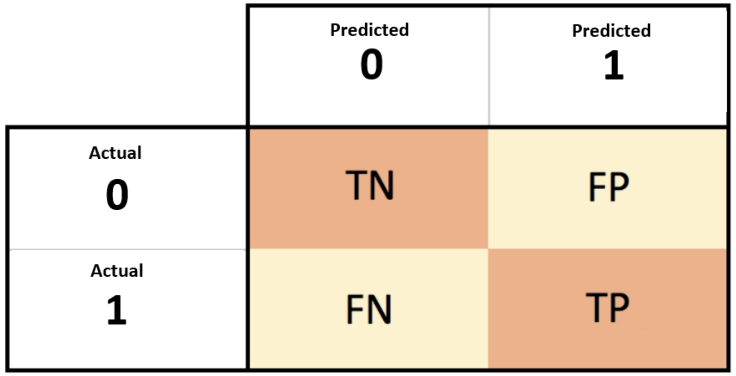
$$h(x) = h_1(x) + h_2(x) + \dots + h_n(x)$$

$$S_1 = \{(x_i, y_i)\}_{i=1}^{N} \rightarrow S_2 = \{(x_i, y_i - h_1(x_i))\}_{i=1}^{N} \rightarrow S_n = \{(x_i, y_i - h_{1:n-1}(x_i))\}_{i=1}^{N}$$

Reference: https://dimensionless.in/gradient-boosting/

Metrics to Validate Classification Model

Confusion Matrix:



	Predicted 0	Predicted 1
Actual 0	TN	FP
Actual 1	FN	TP

Reference:
Packtpub.com

Accuracy:

$$\frac{TN + TP}{TN + TP + FP + FN}$$

Precision:

$$\frac{TP}{TP + FP}$$

Recall:

$$\frac{TP}{TP + FN}$$

F1 Score:

$$\frac{2*P*R}{P + R}$$

Accuracy: Proportion of correct classifications

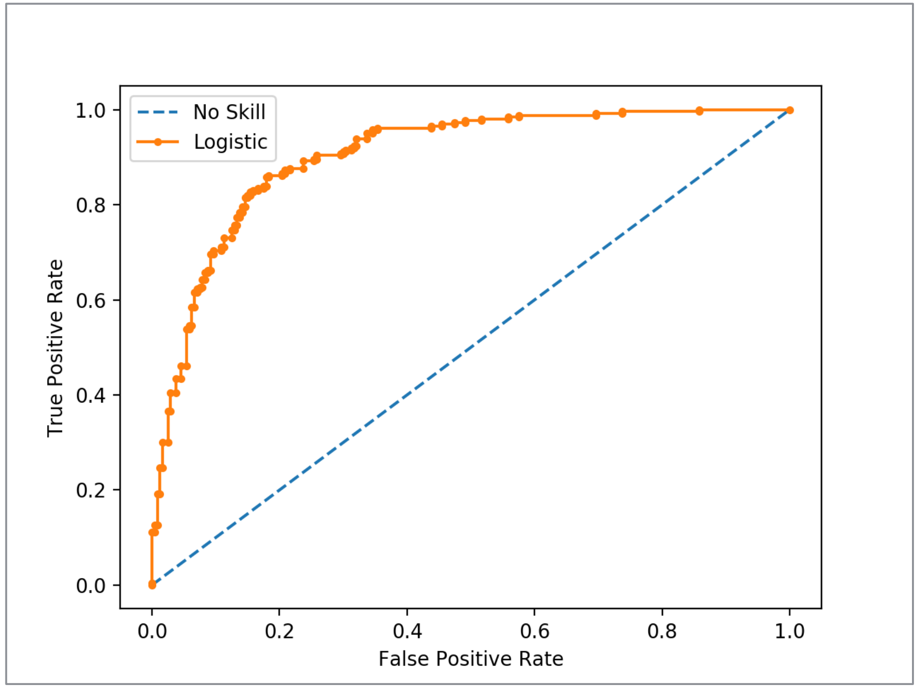
Precision: Quantifies the number of correct positive predictions made. It’s a good metric to validate if the cost of false positives is very high.

Recall: Quantifies the number of correct positive predictions made out of all positive predictions that could have been made. It’s a good metric to validate if the cost of false negatives is very high.

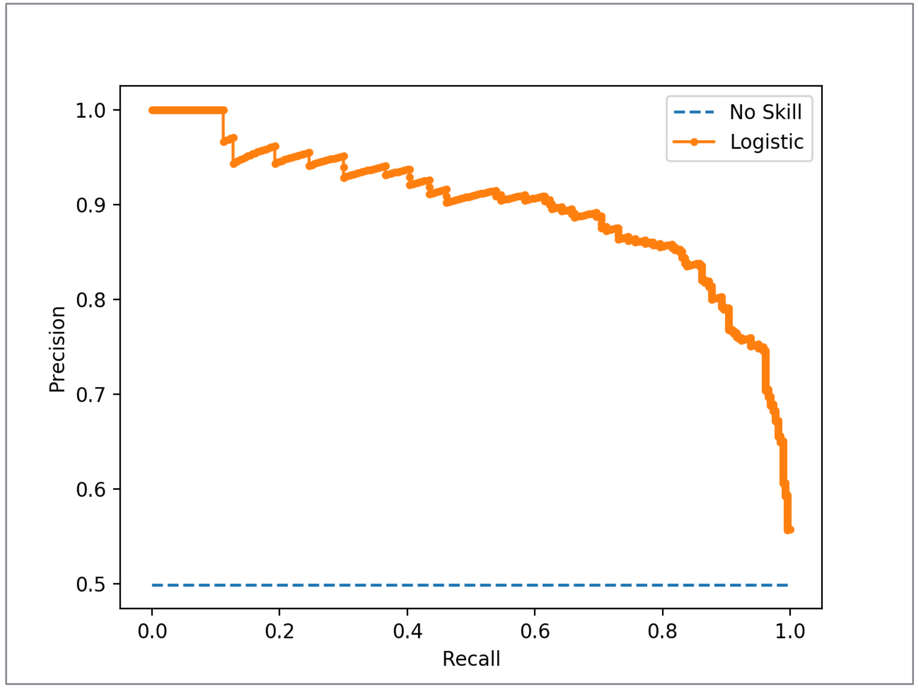
F1 Score: Balances between precision and recall

AUC-ROC and AUC-PR

AUC-ROC



AUC-PR



Recall/TPR:

$$\frac{TP}{TP + FN}$$

FPR:

$$\frac{FP}{FP + TN}$$

Reference: https://machinelearningmastery.com/roc-curves-and-precision-recall-curves-for-imbalanced-classification/

Model Building

Train Metrics

Trained Model: Gradient Boost Machine (GBM)

Number of enrollments in train: 72,395

Confusion Matrix for F1-optimal threshold

7,968	7,061
1,923	55,443

87.6%

AUC-ROC: 0.87

AUC-PR: 0.95

Max F1: 0.92

Threshold: 0.47

Test Metrics

Number of enrollments in test: 24,013

Confusion Matrix for F1-optimal threshold

2,411	692
2,491	18,419

86.7%

AUC-ROC: 0.85

AUC-PR: 0.94

References

1. KDD Cup 2015 Challenge
2. Code

Try this out: Will Bill Solve it?

Agenda

Introduction to Intuit

Prerequisites

Problem Statement

Data Understanding

Feature Engineering

EDA (Exploratory Data Analysis)

Model Building

Demo Time

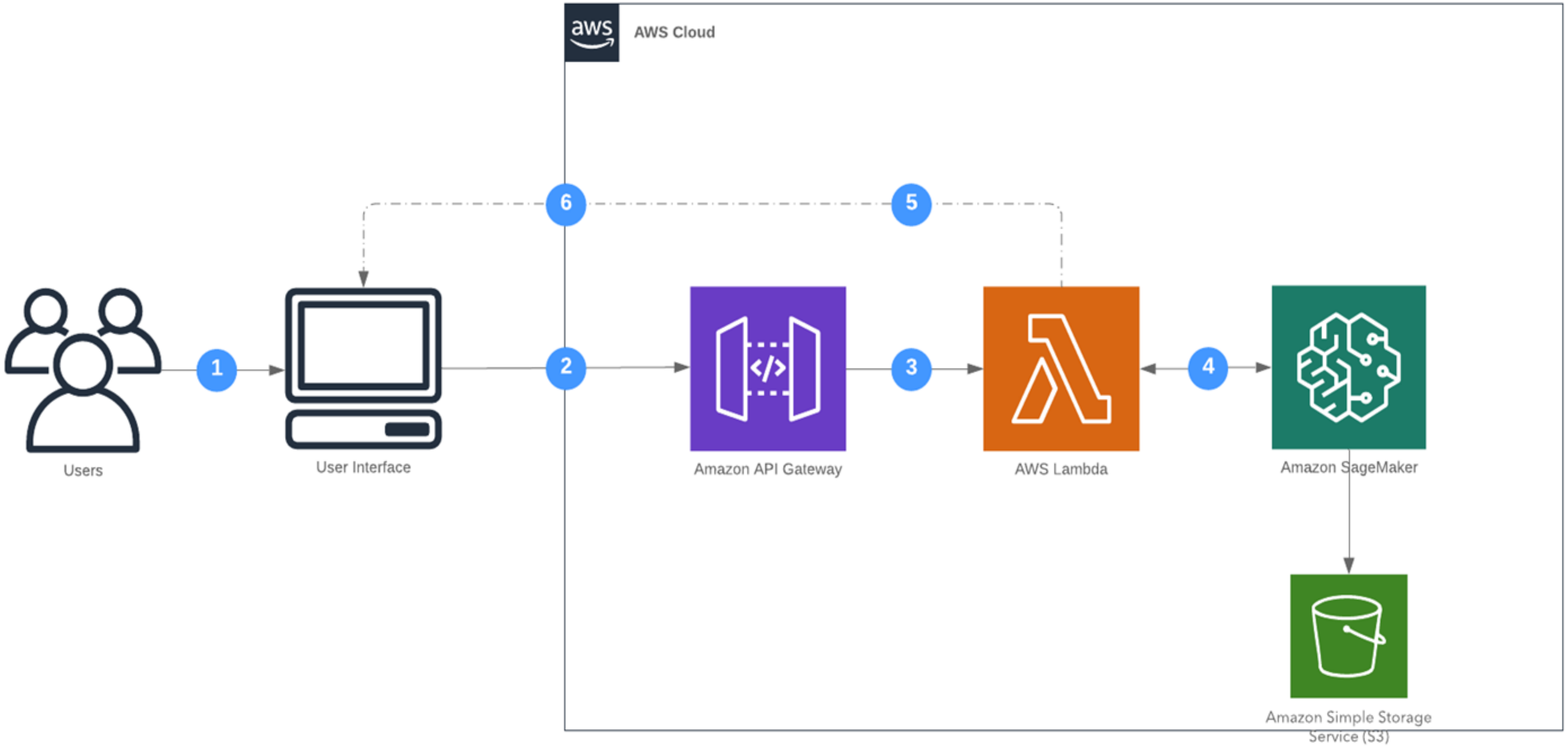
Challenge Time

Automated EDA

Monotonous work by data scientists trying to explore data.

- Code-free Data Analysis on large datasets
- Basic Statistical Metrics
- Variable Importance and Information Gain

Architecture



Financial and Technological Behavior of People in Rural India

The dataset used for this exercise contains demographic and behavioral information from a representative sample of survey respondents from India and their usage of traditional financial and mobile financial services. The dataset is a product of InterMedia's research to help the world's poorest people take advantage of widely available mobile phones and other digital technology to access financial tools and participate more fully in their local economies. Women in these communities, in particular, are often largely excluded from the formal financial system. By predicting gender, the datathon teams will explore the key differences in behavior patterns of men and women, and how that may impact their use of new financial services. Ideally, these findings will influence plans to reach women in developing economies and encourage them to adopt new financial tools that will help to lift them and their families out of poverty.

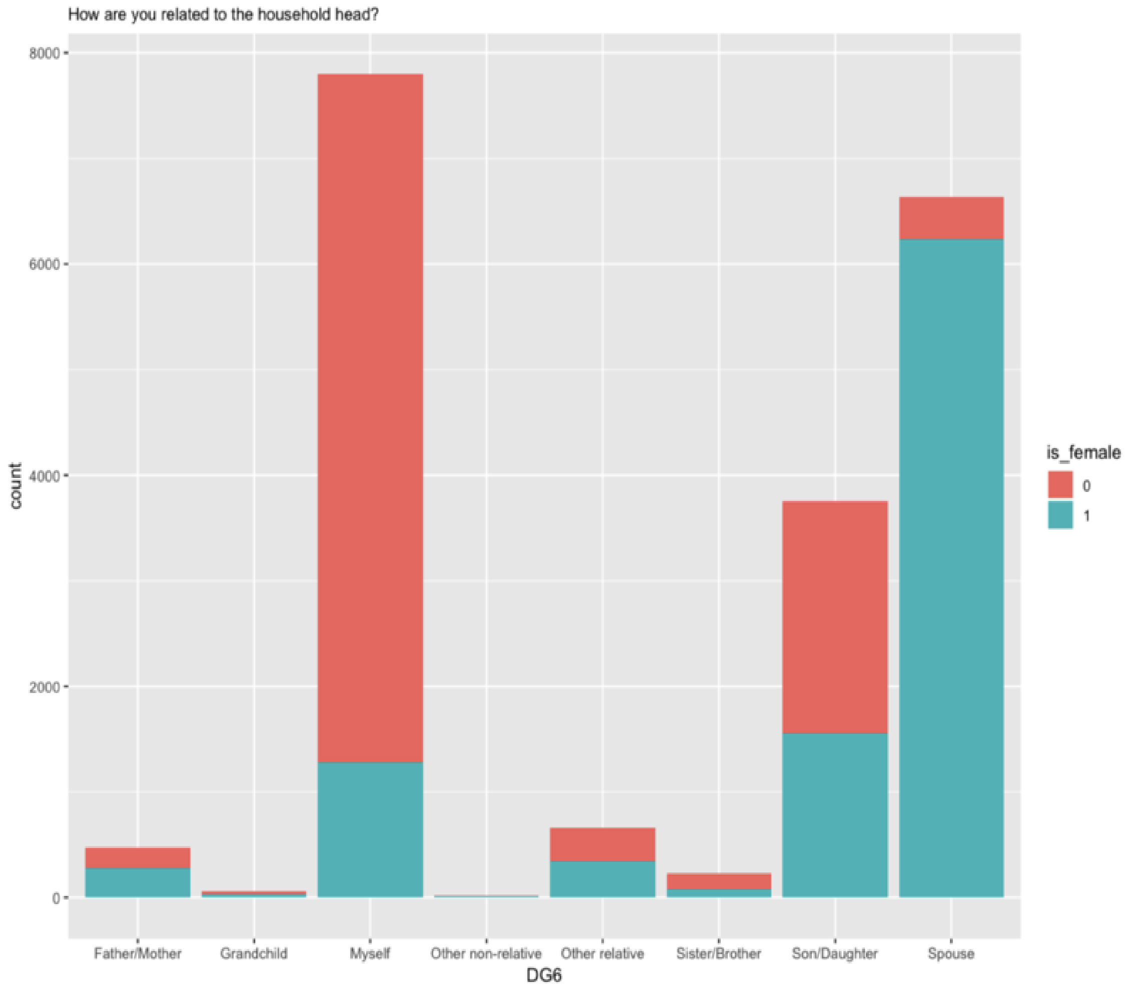
Demo

What are we looking for?

There are multiple choice/numerical questions in the dataset!!

Which of the features do You Think are Important?

Build a model to predict which variables most strongly predict individually (and together) who is a female and who is not.



Challenge Time

Q&A

Your opportunity to ask and learn