

SOEN 6611 - Software Measurement

Credit Card Classification

Team 8

Akshay Dhabale - 40163636

Kshitij Yerande - 40194579

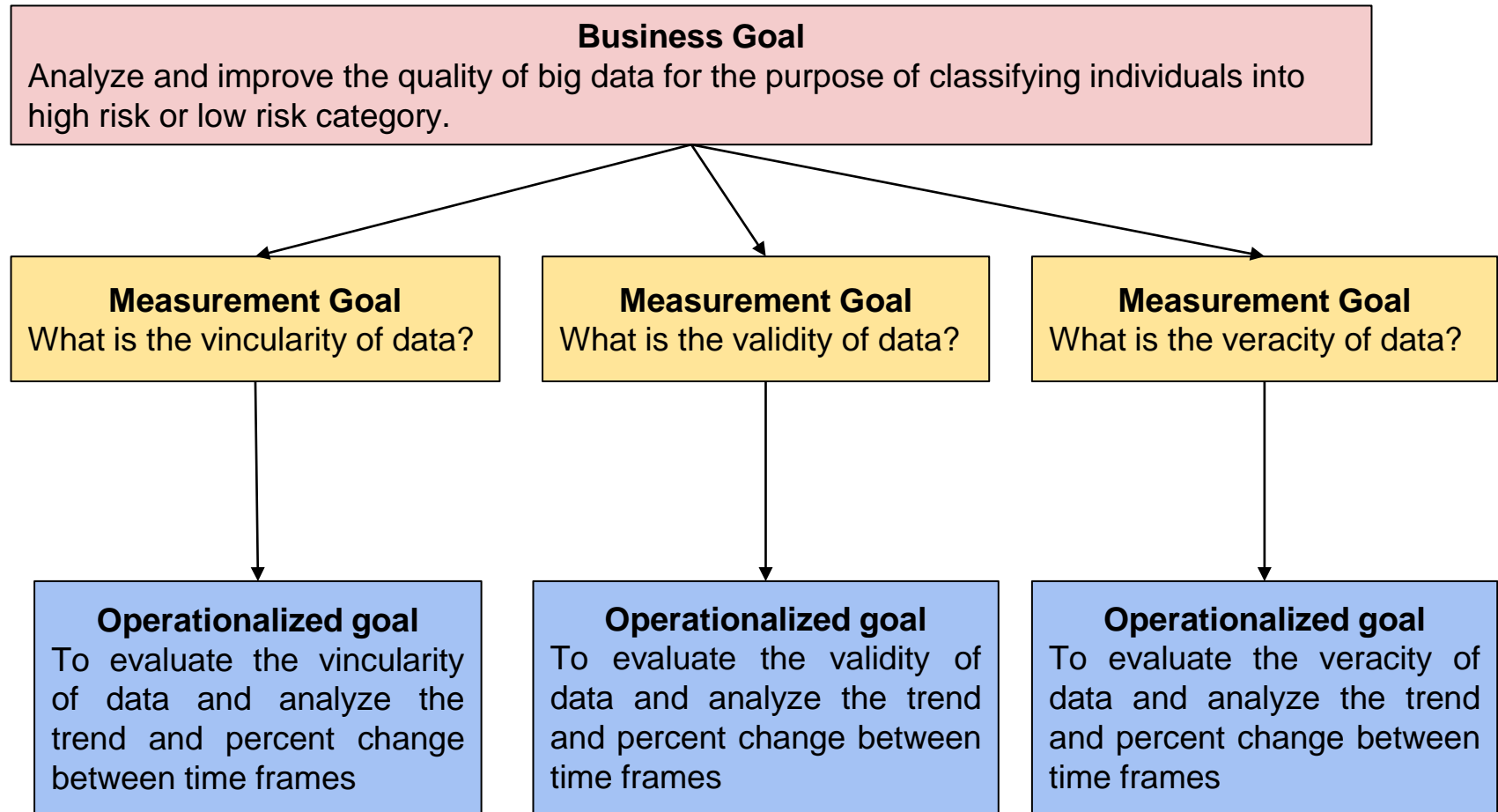
Mrinal Rai - 40193024

Siddhartha Jha- 40201472

Submitted to: Dr. Olga

Ormandjieva

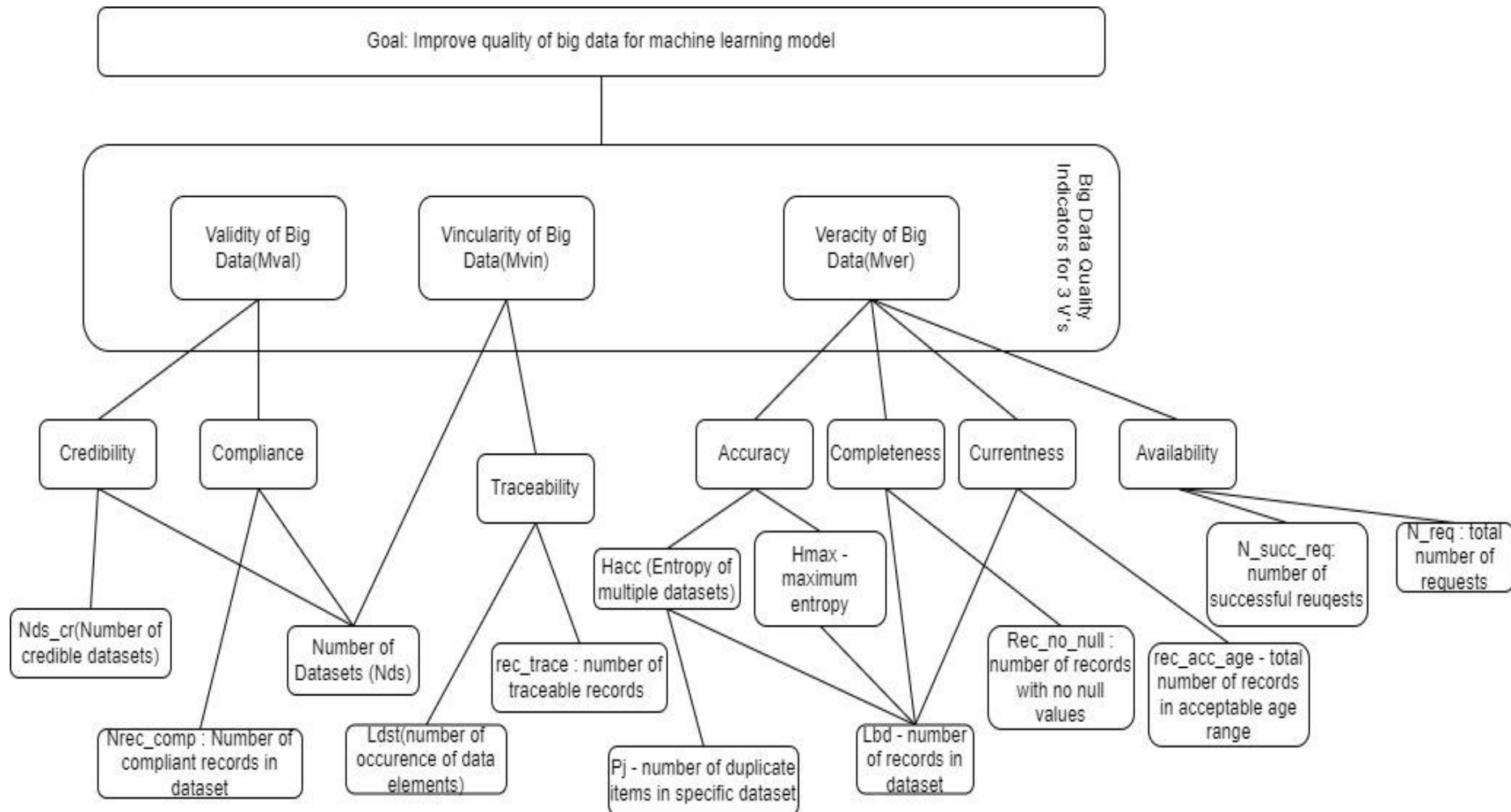
Establishing And Formalizing Measurement Goals



Success Criteria And Indicators

Measure	Indicator	Success Criteria
Veracity	<ol style="list-style-type: none"> 1. Mver for three time frames 2. Percentage Change between time frame 	<p>Mver ≥ 0.8</p> <p>Percentage change $< -5\%$ (- represent direction of change)</p>
Vincularity	<ol style="list-style-type: none"> 1. Mvinc for three time frames 2. Percentage Change between time frame 	<p>Mvinc ≥ 0.6</p> <p>Percentage change $< -10\%$ (- represent direction of change)</p>
Validity	<ol style="list-style-type: none"> 1. Mval for three time frames 2. Percentage Change between time frame 	<p>Mval ≥ 0.9</p>

Measurement Hierarchy for Vincularity, Validity and Veracity



Planning the Measurement Process

Data Collection

Measuring Base Measures on raw data

Measuring Derived Measures on raw data

Data Processing

Measuring Base Measures on processed data

Measuring Derived Measures on processed data

Analyzing and interpreting results

Data Collection

Credit Card Classification Dataset

Source: <https://www.kaggle.com/samueltortinhas/credit-card-classification-clean-data>

Description: Credit cards are a popular risk management strategy in the financial sector. To estimate the likelihood of future defaults and credit card borrowing, it uses the personal information and data provided by credit card applicants.

Machine Learning Problem Statement:

To predict the whether an individual is at high credit risk or low credit risk.

General Information:

- **Numbers of rows: 9921**
- **Number of columns : 20**

Column	Description	Value
ID	Unique identifier of the client	Unique number assigned to person
Gender	Gender of the person	1 = Male 0=Female
Own_Car	Does the person own car	1=Yes 0 =No
Own_Property	Does the person own property	1=Yes 0 =No
Work_phone	Does the person have work phone	1=Yes 0 =No
Phone	Does the person have personal phone	1=Yes 0 =No
Email	Does the person have email ID	1=Yes 0 =No
Unemployed	Is the person unemployed	1=Yes 0 =No

Refer to report for details on additional column

Data Processing

Data processing was performed to clean data. Following operation were performed to clean data:

- Remove records with null values
- Remove duplicate records
- Remove records where values does not adhere to column standards. E.g (alphabets in numeric columns, column value out of range)
- Age column value was rounded to integer as age is supposed to be a integer value.

Issue	Value	
Number of null records in data set	Gender 16 Own_car 16 Own_property 16 Work_phone 44 Phone 20 Email 16 Unemployed 16 Num_children 16 Num_family 44 Account_length 16 Total_income 16 Age 16 Years_employed 16 Income_type 16 Education_type 16 Family_status 16 Housing_type 16 Occupation_type 16 Target 16 Date 16	
Number of duplicate records	11	
Columns which does not have acceptable values	Gender: 8 Phone: 124 Target: 144	Work Phone: 13 Unemployed: 135

Calculating Base Measures

Measure	Description	Value - Before Processing	Value After Processing
Nds	Number of Dataset	1	1
Nds_cr	Number of credible dataset	1	1
Nrec_comp	Number of compliant records	9756	9745
Rec_trace	number of traceable records	9756	9745
Ldst	total number of records	9920	9745
Lbd	total number of records	9920	9745
Rec_acc_age	number of records within acceptable time range	4957	4880
N_succ_req	number of successful requests	1	1
N_req	number of requests	1	1
rec_no_null	number of records with no null values	9852	9745
P(j)	number of duplicate records in dataset	11	0

Calculating Derived Measures

Measure	Value before processing	Value after processing
Accuracy	0.99	1
Completeness	0.99	1
Currentness	0.49	0.50
Availability	0.83	0.89
Traceability	0.98	1
Compliance	1	1
Credibility	0.98	1

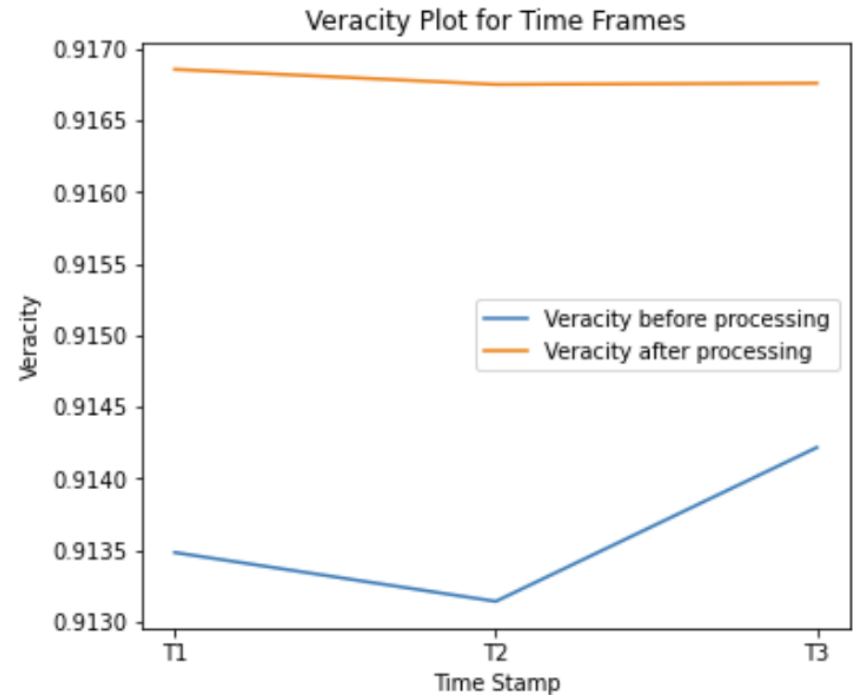
Note: Values above are the average over timeframe.

Results- Veracity

Veracity before processing and after processing is above the acceptable threshold ≥ 0.8

Average Veracity after processing has increased slightly by 0.35% signifying improvement.

Percentage change between time frame t2-t3 is positive meaning more accurate, complete records are added to dataset.



Veracity							
T1(before)	T2(before)	T3(before)	Average	T1(after)	T2(after)	T3(after)	Average
0.9134	0.9132	0.9142	0.9136	0.9168	0.9167	0.9167	0.9167

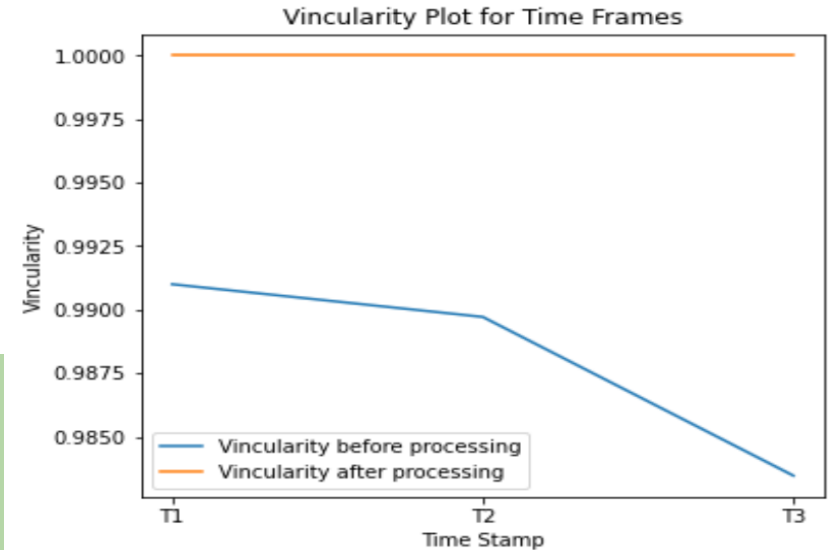
% Change					
T1 - T2(Before)	T1-T3(Before)	T2-T3(Before)	T1-T2(After)	T1-T3(After)	T2-T3(After)
-0.04	0.08	0.12	-0.01	-0.01	0

Results- Vincularity

Vincularity before processing and after processing is above the acceptable threshold ≥ 0.6

Average Vincularity after processing has increased slightly by 1.21% signifying some improvement.

Vincularity before processing is on a downward slope as new data is added however is well above the acceptable threshold and drop threshold.



Vincularity							
T1(before)	T2(before)	T3(before)	Average	T1(after)	T2(after)	T3(after)	Average
0.991	0.9897	0.9834	0.9880	1	1	1	1

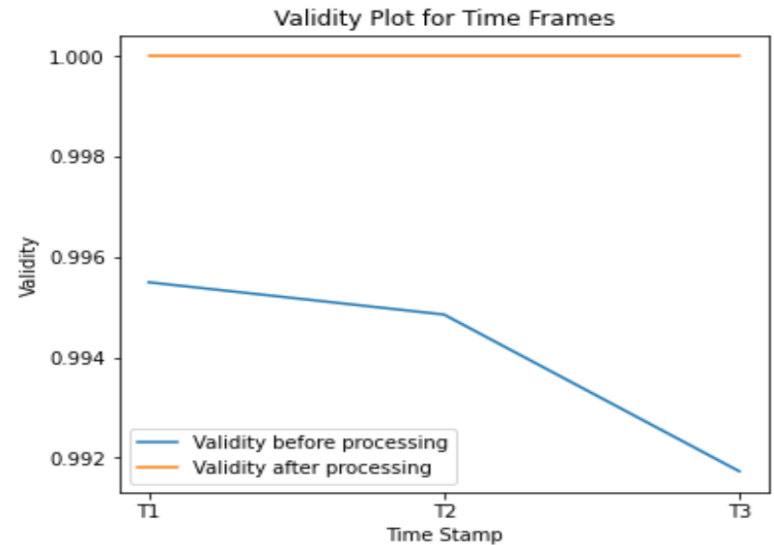
% Change					
T1 - T2(Before)	T1-T3(Before)	T2-T3(Before)	T1-T2(After)	T1-T3(After)	T2-T3(After)
-0.13	-0.76	-0.63	0	0	0

Results- Validity

Validity before processing and after processing is above the acceptable threshold ≥ 0.9

Average Validity after processing has increased slightly by 0.6% signifying some improvement.

Validity before processing dropped slightly in t2 - t3 time frame due increase in very few non compliant records.



Validity							
T1(before)	T2(before)	T3(before)	Average	T1(after)	T2(after)	T3(after)	Average
0.9955	0.9948	0.9917	0.9940	1	1	1	1

% Change					
T1 - T2(Before)	T1-T3(Before)	T2-T3(Before)	T1-T2(After)	T1-T3(After)	T2-T3(After)
-0,06	-0.38	-0.31	0	0	0

Lessons Learned

Good data quality helps strengthening the trust on the analysis results

Structured approach of measurement process makes things clear and stay on track

Learned to apply measurement process on a given problem.

Learned about new V's (Vincularity, Validity and Veracity) and how to measure them.

What could be changed to deliver better results?

Better planning of tasks and analysis of scope would have avoided last minute changes

Lack of work experience creates difficulty in understanding the issues faced when working with actual real-time data

What still puzzles us?

How Quality measurement procedures work in parallel with SDLC

**Thank
You**