Group project report of US H1-B visa analysis for profile scanner

By

Mr. Chonchanok Chevaisrakul 6210545459

Sec 450

Submit to

Asst. Prof. Kitsana WAIYAMAI

This report submitted for Database Analytics

Engineers project

Data Analytics

01219367 Second semester, 2023

Table of content

Chapter 1: Introduction	3
Chapter 2: Data exploration and contained data	6
Chapter 3: Data preprocessing	12
Chapter 4: Analytics technique	15
Chapter 5: Result, data presentation and conclusion	18
Reference	25

Chapter 1
Introduction and statement of the problem

1.1 Introduction

The H-1B visa is a type of working visa issued by the Department of State of the United States of america. This type of visa has a unique characteristic for temporary working visas for specialist occupations such as doctor, professor, engineer or other careers that require a minimum education level in bachelor or higher. [1]The approval rate (certificated) based on the department of labor statistics is 92.7%. And the point of the importance of this type of visa is used to be the pathway to the green card or permanent resident permit in the United States.

1.2 Statement of the problems

The purpose of this study is to improve the model of the prediction of the visa application by using Gradient boosting, XG boost and combined with hybrid sampling (upsampling and downsampling) to prove the performance of the model by using cross validation to test the performance of the data.

1.3 The data contains

[2] This data that we use for analytics is provided by Sharan Naribole in Keggle (website for find and research data analytics). Data contains 3 million records of H1-B visa applications from 2011 to 2016. The column of the record has case status(e.g. Certified or denied), employer organization, occupation, job title, time position, the prevailing wage, year of application, town of worksite with latitude and longitude. This data can be categorized as many decision data because the file size is around 0.2GB and 3 million records inside.

1.4 The objective of this analytics

The objective of this analytics application is to improve the data for predicting the profile of the application that passes on the criteria of the requirement based on the previous result of the application in the past. And also with the way to prove the faster speed to execute the application with lower time to execute and higher or equal performance of the prediction.

1.5 The solution for the statement

We designed to use the data in the CSV format to achieve the limitation of the XLSX (MS-Excel format). And use the rapid miner application to make the process of data exploration, data visualization, pre-processing, make a model for analytics, which indicates the performance of the prediction model.

1.6 The limitation of this project

The limitation of the project is that the many decision data with high features cannot be covered all by the personal computer that we use. We have to split the data to make the computer compatible. From the mention, the data may lose some of the important data due the limitation of the power of computing and by the limitation of the application that we use for data analytics in this case is rapid miner version 10.3.

1.7 The use of the other application.

This project after implementation can be used as we expect in the below application. To improve the performance of their organization.

- Primary scan for embassy/consulate or visa agency profile firm. To prescreen in brief for less manpower to screen the visa application. E.g. Schengen visas or US visas often take a long period to make a decision. This project can be used as a baseline to screen profile and detect the suspect profile in primary or as a shortcut to fast track process if the tree gives a result in a positive way.
- Statistical division of the tracking immigrant. This project is also used

as informative data for decisions such as criteria for issuing visas in future or making a warning list from the profile data. Or immigration report which takes a lot of time to approve or deny. They can use this model to make a good profile here and can access some of the priority such as autogate or high possibility to extend their visa.

- Job hunter (e.g. linkedin). Because the H1-B visa is a visa that is issued for high-skill employees from overseas. Some of this category applications can be used to prescreen the data and give a guideline to services inside the application to make it more convenient to companies and people who are finding their job in the US. And make a higher percentage for companies to avoid scam and make employees better prepared before making a contact.

Chapter 2
Data exploration and contained data

2.1 Data profile

This data has originated from the employment and training administration in the department of labor of the United States. Focus on the visa type H1-B visa application from 2011 and 2016. Which all of this data include every status from various people profile applications approximately around 3 million records. Export as the CSV format.

Open in Turbo Prep Auto Model Interactive Analysis Filter (3,000,378 / 3,000,378 examples):

Row No.	ID	CASE_STAT...	EMPLOYER_...	SOC_NAME	JOB_TITLE	FULL_TIME_...	PREVAILING...	YEAR	WORKSITE	lon	lat
1	1	CERTIFIED-...	UNIVERSITY ...	BIOCHEMIST...	POSTDOCTO...	N	36067	2016	ANN ARBOR,...	-83.7430378	42.2808256
2	2	CERTIFIED-...	GOODMAN N...	CHIEF EXEC...	CHIEF OPER...	Y	242674	2016	PLANO, TEXAS	-96.6988856	33.0198431
3	3	CERTIFIED-...	PORTS AME...	CHIEF EXEC...	CHIEF PROC...	Y	193066	2016	JERSEY CITY...	-74.0776417	40.7281575
4	4	CERTIFIED-...	GATES COR...	CHIEF EXEC...	REGIONAL P...	Y	220314	2016	DENVER, CO...	-104.990251	39.7392358
5	5	WITHDRAWN	PEABODY IN...	CHIEF EXEC...	PRESIDENT ...	Y	157518.400	2016	ST. LOUIS, MI...	-90.1994042	38.6270025
6	6	CERTIFIED-...	BURGER KIN...	CHIEF EXEC...	EXECUTIVE ...	Y	225000	2016	MIAMI, FLORI...	-80.1917902	25.7616798
7	7	CERTIFIED-...	BT AND MK E...	CHIEF EXEC...	CHIEF OPER...	Y	91021	2016	HOUSTON, T...	-95.3698028	29.7604267
8	8	CERTIFIED-...	GLOBO MOBI...	CHIEF EXEC...	CHIEF OPER...	Y	150000	2016	SAN JOSE, C...	-121.8863286	37.3382082
9	9	CERTIFIED-...	ESI COMPAN...	CHIEF EXEC...	PRESIDENT	Y	127546	2016	MEMPHIS, TE...	NA	NA
10	10	WITHDRAWN	LESSARD IN...	CHIEF EXEC...	PRESIDENT	Y	154648	2016	VIENNA, VIR...	-77.2652604	38.9012225
11	11	CERTIFIED-...	H.J. HEINZ C...	CHIEF EXEC...	CHIEF INFO...	Y	182978	2016	PITTSBURG...	-79.9958864	40.4406248
12	12	CERTIFIED-...	DOW CORNI...	CHIEF EXEC...	VICE PRESID...	Y	163717	2016	MIDLAND, MI...	-84.2472116	43.6155825
13	13	CERTIFIED-...	ACUSHNET ...	CHIEF EXEC...	TREASURER...	Y	203860.800	2016	FAIRHAVEN, ...	NA	NA
14	14	CERTIFIED-...	BIOCAIR, INC.	CHIEF EXEC...	CHIEF COMM...	Y	252637	2016	MIAMI, FLORI...	-80.1917902	25.7616798
15	15	CERTIFIED-...	NEWMONT M...	CHIEF EXEC...	BOARD MEM...	Y	105914	2016	GREENWOO...	-104.9508141	39.6172101
16	16	CERTIFIED-...	VRICON, INC.	CHIEF EXEC...	CHIEF FINAN...	Y	153046	2016	STERLING, V...	-77.4291298	39.0066993
17	17	CERTIFIED-...	CARDIAC SC...	FINANCIAL M...	VICE PRESID...	Y	90834	2016	WAUKESHA, ...	-88.2314813	43.0116784
18	18	CERTIFIED-...	WESTFIELD ...	CHIEF EXEC...	GENERAL M...	Y	164050	2016	LOS ANGELE...	-118.2436849	34.0522342
19	19	CERTIFIED	QUICKLOGIX...	CHIEF EXEC...	CEO	Y	187200	2016	SANTA CLAR...	-121.9552356	37.3541079
20	20	CERTIFIED	MCCHRYSTA...	CHIEF EXEC...	PRESIDENT, ...	Y	241842	2016	ALEXANDRIA...	-77.0469214	38.8048355
21	21	CERTIFIED-...	CUDDLE BA...	CHIEF EXEC...	CHIEF OPER...	Y	117998	2016	COMMERCE,...	-118.1597929	34.0005691
22	22	CERTIFIED-...	WESTFIELD ...	CHIEF EXEC...	GENERAL M...	Y	164050	2016	LOS ANGELE...	-118.2436849	34.0522342
23	23	CERTIFIED	LOMICS, LLC	CHIEF EXEC...	CEO	Y	99986	2016	SAN DIEGO, ...	-117.1610838	32.715738
24	24	CERTIFIED	UC UNIVERS...	CHIEF EXEC...	CHIEF FINAN...	Y	99986	2016	CHULA VIST...	-117.0841955	32.6400541
25	25	CERTIFIED-...	VMS COMMU...	CHIEF EXEC...	CHIEF OPER...	Y	159370	2016	MIAMI, FLORI...	-80.1917902	25.7616798
26	26	CERTIFIED	QUICKLOGIX...	CHIEF EXEC...	CEO	Y	187200	2016	SANTA CLAR...	-121.9552356	37.3541079

Picture 2.1: Example of data inside the dataset.

2.2 Data features

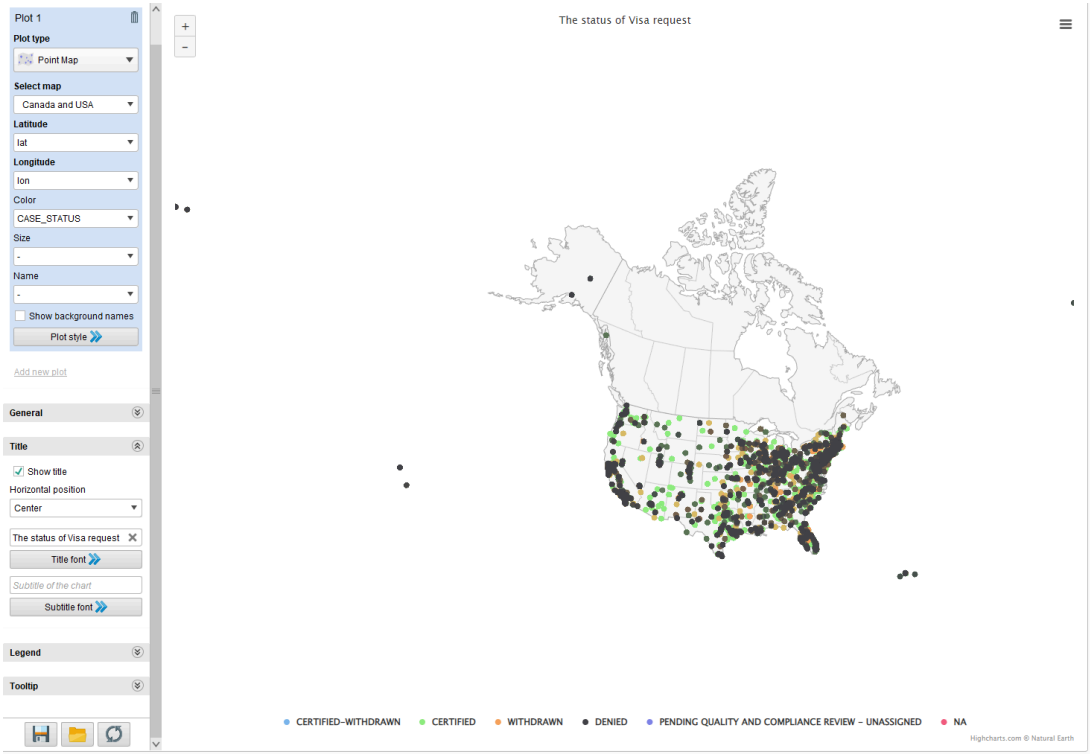
This dataset has 11 features contained inside which are described below.

- Row No. (ID) : This data feature is an ID of the application.
- Case status: The result of status of the visa such as Withdrawn, Certified, Denied etc.
- Employer name: The organization that hires and sponsors the person who is assigned in application. (E.g. company or university)
- SOC name: The occupation name such as Chief Executive, Computer Programmer, Marketing etc.
- Job title: Relate to SOC with the description of the details of Job such as CEO, CFO, President, Quality Assurance.
- Full time position: Contain boolean as Y(Yes) and N(No) to display full time position. In case of N will be a Part time position.
- Prevailing wage: This feature contains the annual salary that the organization offers for their job position.
- Year: The year that application was submitted.
- Worksite: This feature contained the city and state of the employed organization.
- Lat: This feature contain the latitude of the city of worksite
- Lon: This feature contains the longitude of the city of worksite.

2.3. Basic of Data exploration

2.3.1 Population distribution

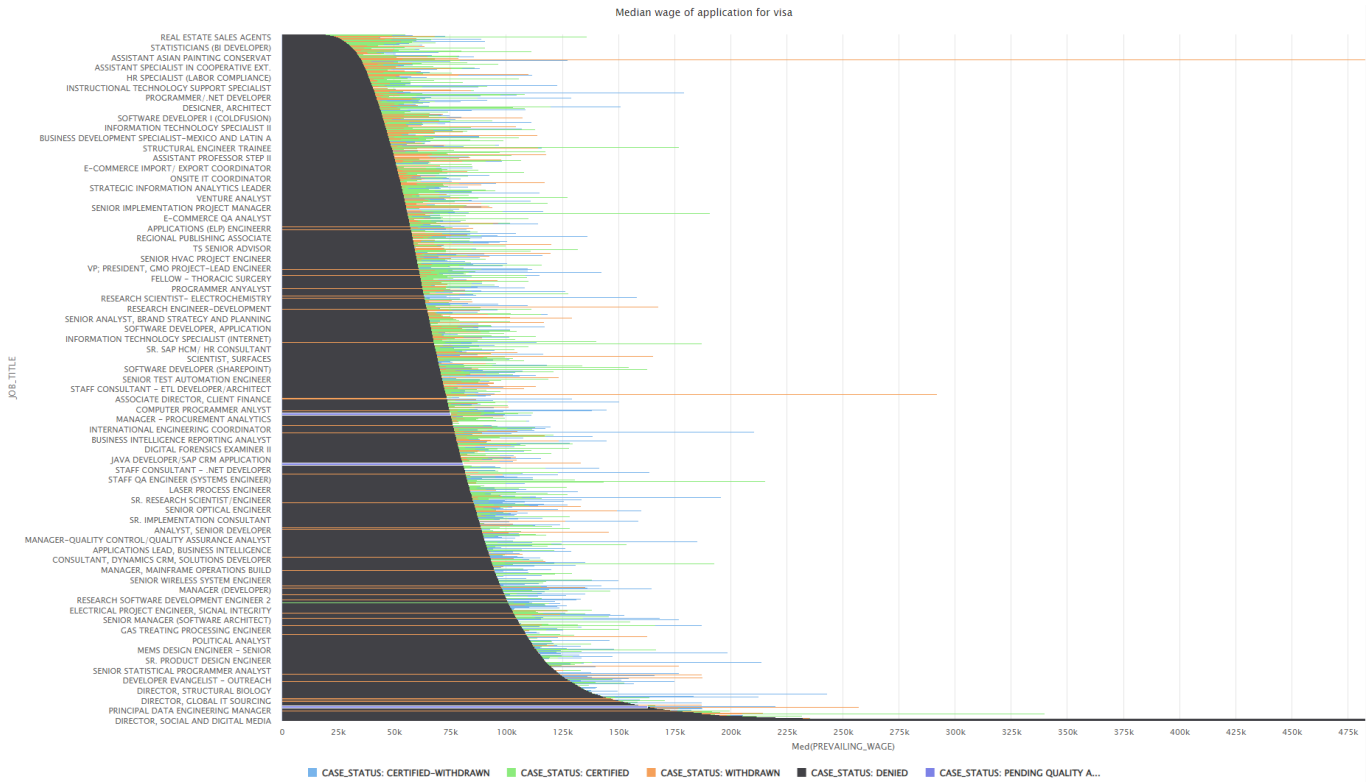
For data exploration, in this case, we decided to use the map for visualizing the latitude and longitude with the visa status of the dataset. This type of plot shows the interesting data of distribution of people in the application.



Picture 2.2: The map plot type with latitude and longitude with visa status

[3]From this data we can see that most of the people have a workplace on the west coast and east coast of the US. And some of the applications are located in the south. But hardly to find in the midwest region in the US, pacific and caribbean region.

2.3.2 Bar chart of the wage offer with the occupation



Picture 2.3: Bar chart of wage offer with the occupation list.

From the above data displayed the wage offer with ascending sorting. Overall we can see that the denied visa application has a lowest salary when compared with the same job title. But no significant difference between certified withdrawn and pending on review. The highest salary profile on this dataset has one of these status which is senior, work in a STEM career such as Engineer, or be a level of manager or higher. The lower part, some of the result has a level of experience as a junior or assistant.

2.4 Technique explanation

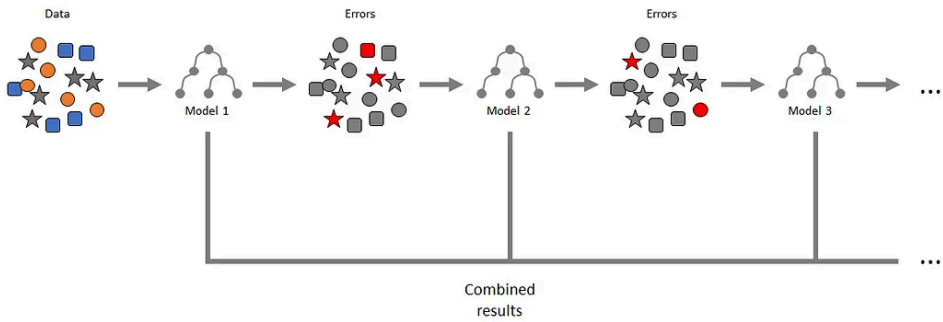
Difference Between Boosting Algorithms

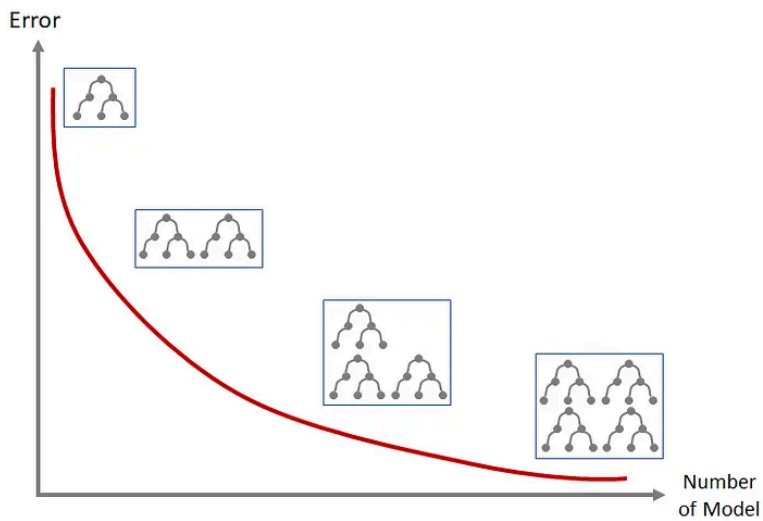
Algorithms	Gradient Boosting	AdaBoost	XGBoost	CatBoost	LightGBM
Year	–	1995	2014	2017	2017
Handling Categorical Variables	May require preprocessing like one-hot encoding	No	NO	Automatically handles categorical variables	No
Speed/Scalability	Moderate	Fast	Fast	Moderate	Fast
Memory Usage	Moderate	Low	Moderate	High	Low
Regularization	NO	No	Yes	Yes	Yes
Parallel Processing	No	No	Yes	Yes	Yes
GPU Support	No	No	Yes	Yes	Yes
Feature Importance	Available	Available	Available	Available	Available

Picture 2.4: Comparison in Boosting algorithm family
Source: https://www.geeksforgeeks.org/gradientboosting-vs-adaboost-vs-xgboost-vs-catboost-vs-lightgbm/

2.4.1 Gradient boosting and XG boost

[4] The gradient boosting decision tree (GBDT) is a decision tree similar to a random forest. Each tree will make a relation with the previous model by the latest. By focusing on the wrong prediction to make more accuracy to make a better decision model. But with the high-efficient model still has disadvantage in speed of compute and overfitting.





Picture 2.5 and 2.6: Concept of Gradient boosting decision tree
Source: https://medium.com/kbtg-life/tree-based-algorithms-%E0%B9%81%E0%B8%9A%E0%B8%9A-high-level-4058e909e0c5

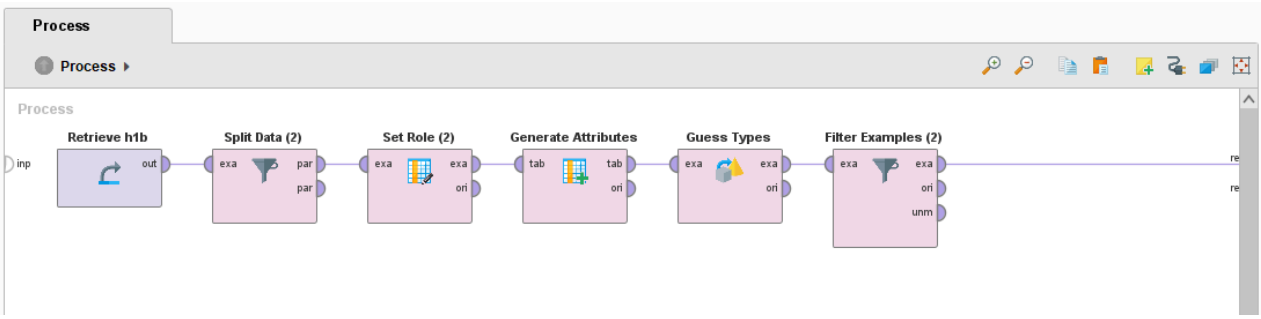
[5][6] From above explanation make we tried to introduce the solution of the problem by using **XGBoost** which improving speed and scalability, supportability and regularization to reduce generalization and overfitting or underfitting

Chapter 3
Data preprocessing

3.1. Data import configuration

The data after download from the source must be imported by setting the configuration of file encoding to UTF-8 (in case MS-Windows will default here as windows-1252) to make sure that it will be compatible with other devices.

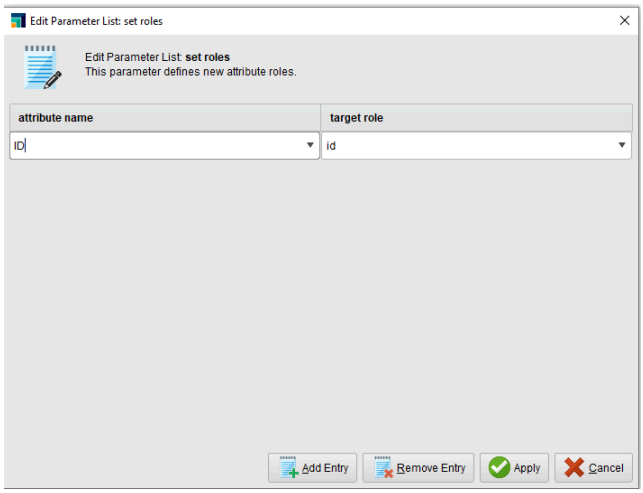
3.2 Pre-processing for finding visa status.



Picture3.1: Preprocessing process on Rapidminer 10.3 for visa status

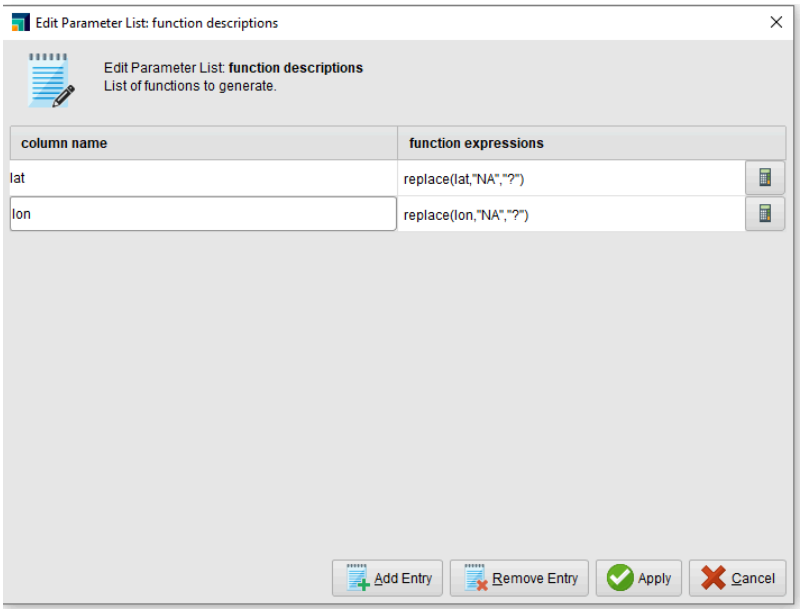
For preprocessing, start from retrieving the dataset that we need to process. Then split the data to make the load of processing less for not over the runtime of the device. In the case of using 3 million records the device cannot hold the process and timeout because the size is too overload. In this case may use approximately 300,000 - 750,000 that device still functional for executing the process.

Then set roles to make sure that the data will not be used in the processing. In this case is to set the id column as an ID like in the case below.

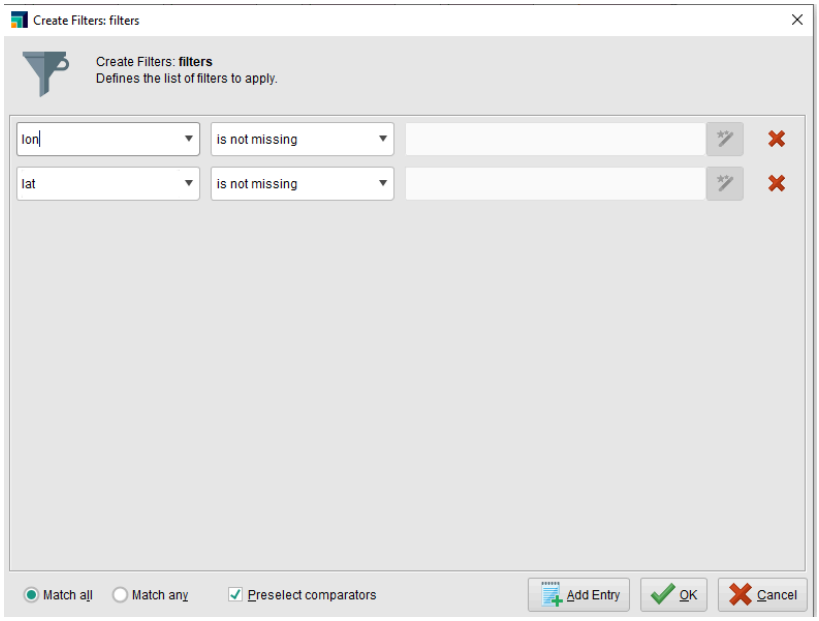


Picture3.2: Set role for ID

After setting the role for the dataset, we realize that the data in this case for latitude and longitude are still nominal and the missing value in this data is still assigned as NA which makes the data not numerical yet. We have to generate the attribute and change the type then filter out the missing value.



Picture3.3: Function expression for replacing NA to missing value



Picture3.4: Filter missing value out from the data by custom option.

In our process, after we change NA to “?” as a missing value already. We have to use an operation named Guess Type. This function has more convenience than we look and change by ourselves. The application will find the best option to convert and process to the best match of type. And also with the advantage we can see abnormal things that may still include in the column of the dataset. In the case that the function expression does not conclude all of the data in the column the sign will be found on the visualization and statistics after we run the process.

Filter (11 / 11 attributes):

Name	Type	Missing	Statistics		
ID (Id)	Integer	0	Min 6	Max 3002458	Average 1495005.892
CASE_STATUS	Polynominal	0	Least INVALIDATED (0)	Most CERTIFIED (631853)	Values CERTIFIED (631853), CERTIFIED-WITHDRAWN (49102), ...[6 more]
EMPLOYER_NAME	Polynominal	0	Least ĖNIMAI, INC. (0)	Most INFOSYS LIMITED (326...	Values INFOSYS LIMITED (32618), TATA CON [...] S LIMITED (16075), ...[235732 more]
SOC_NAME	Polynominal	0	Least Woodwork [...] Other (0)	Most Computer [...] s (71971)	Values Computer [...] Analysts (71971), Computer Programmers (55590), ...[2116 more]
JOB_TITLE	Polynominal	0	Least TEST ANALYST - US (0)	Most PROGRAMM [...] T (613...	Values PROGRAMMER ANALYST (61367), SOFTWARE ENGINEER (29938), ...[287222 more]
FULL_TIME_POSITION	Polynominal	0	Least NA (6)	Most Y (621796)	Values Y (621796), N (101755), ...[1 more]
PREVAILING_WAGE	Real	16	Min 0	Max 413472579	Average 141723.468
YEAR	Integer	6	Min 2011	Max 2016	Average 2013.866
WORKSITE	Polynominal	0	Least FORT WA [...] VANIA (0)	Most NEW YORK [...] K (47989)	Values NEW YORK, NEW YORK (47989), HOUSTON, TEXAS (20768), ...[18495 more]
lon	Real	0	Min -157.858	Max 145.730	Average -92.109
lat	Real	0	Min 13.437	Max 64.838	Average 38.162

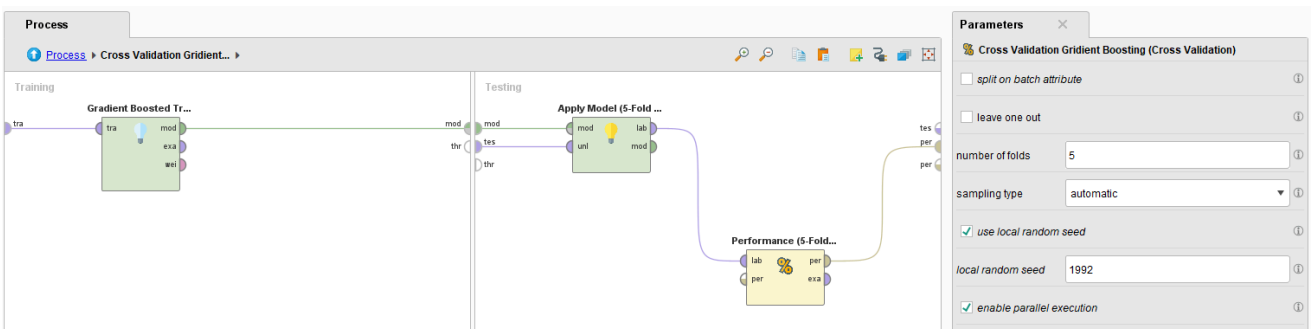
Picture 3.5 : Data after preprocessing in statistics.

After I cross check with the data statistic, I realized that year and wage do not remove all of the missing data. Because of my new experience swapping from Python to Rapidminer and too much focusing to optimize the model to run on the machine and focus too much on the preprocessing latitude and longitude. So I can improve here by using a filter and set in both from above which is not missing for 2 more add entries.

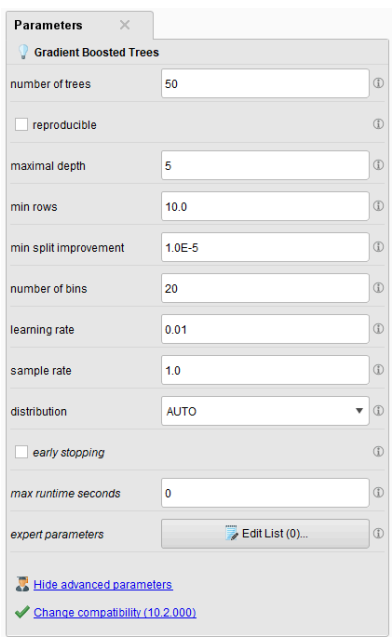
Chapter 4
Analytics technique

4.1 Using a gradient boost tree for finding the model of the tree for the statement.

After we have done our data processing. We have to set roles again but in this case we will set a label for finding. In this case we will find the status of the visa for making a tree. And connect to the operator of the gradient boost tree with a test by cross validation to find the result and performance of the data. In this case we will set the number of trees that we need to process at default(50) and have a test by 5-fold-cross validation.

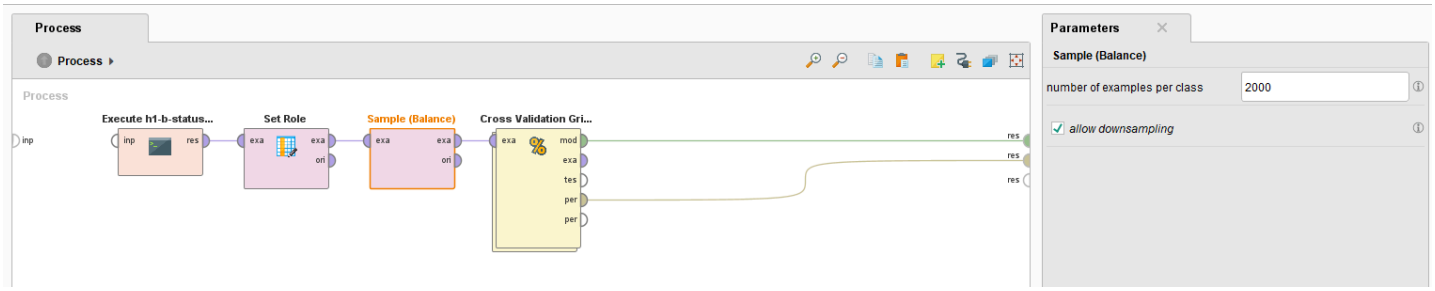


Picture4.1: Process inside cross validation for gradient boost



Picture4.2: Setting of default gradient boost tree.

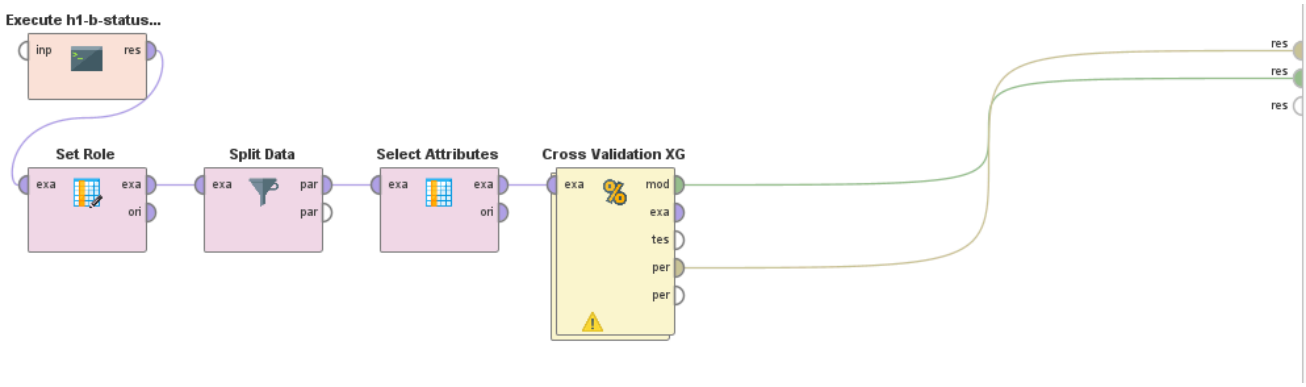
4.2 The proven of the performance by using hybrid sampling (upsampling and downsampling)



Picture4.3: Setting of hybrid sampling.

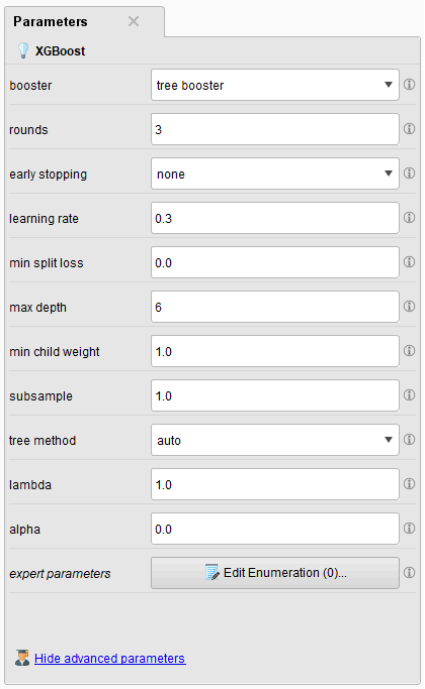
After we finished the process 4.1 and data exploration we see that percent of certified have a very high percentage (around 87.1 percent of dataset). So we decide to find the improvement by setting different samples per class to find the performance that is better or worse accuracy. This extension is required to pre-download from the marketplace.

4.3 The proven of the performance using XG boost

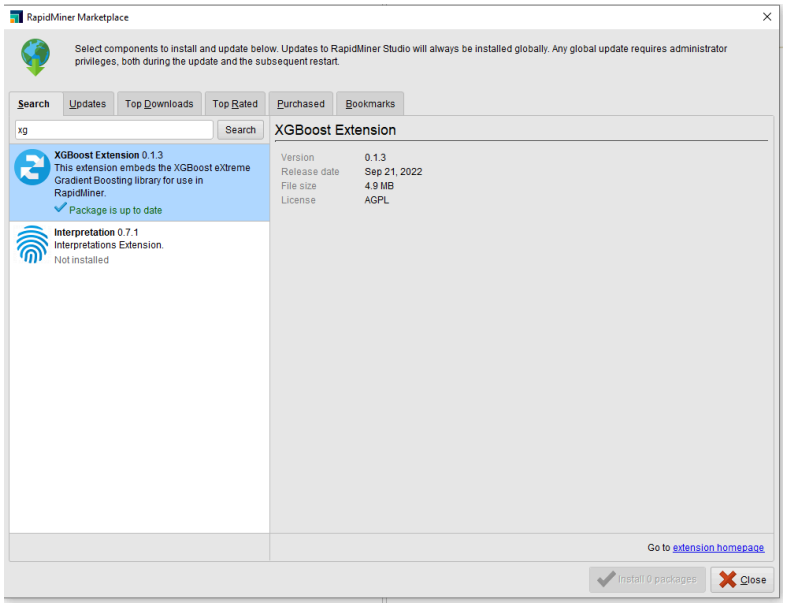


Picture4.4: Setting of XG boost.

After we use a gradient boosting decision tree to improve the performance we decide to use XG boost to improve the speed to compute. This is an extension of the Rapidminer application which is required to download before use.



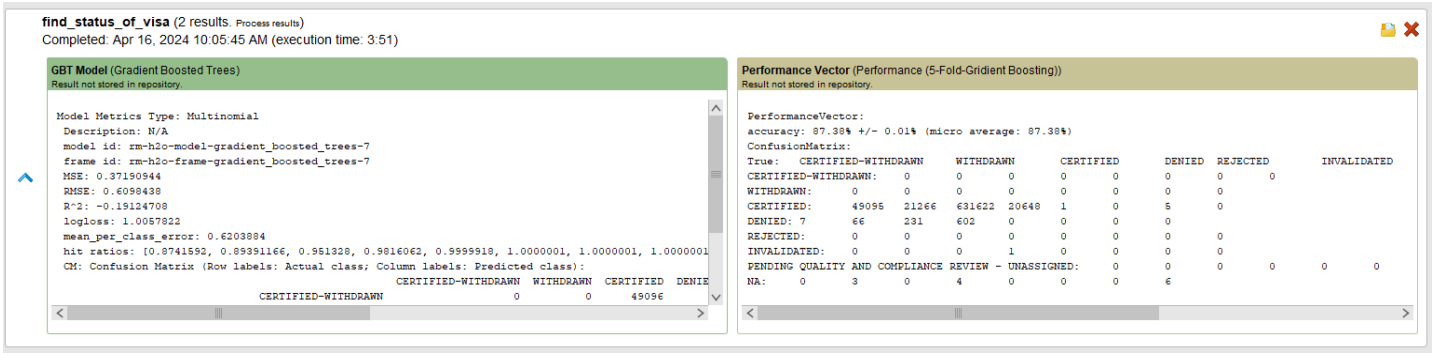
Picture4.5: Configuration of XG boost.

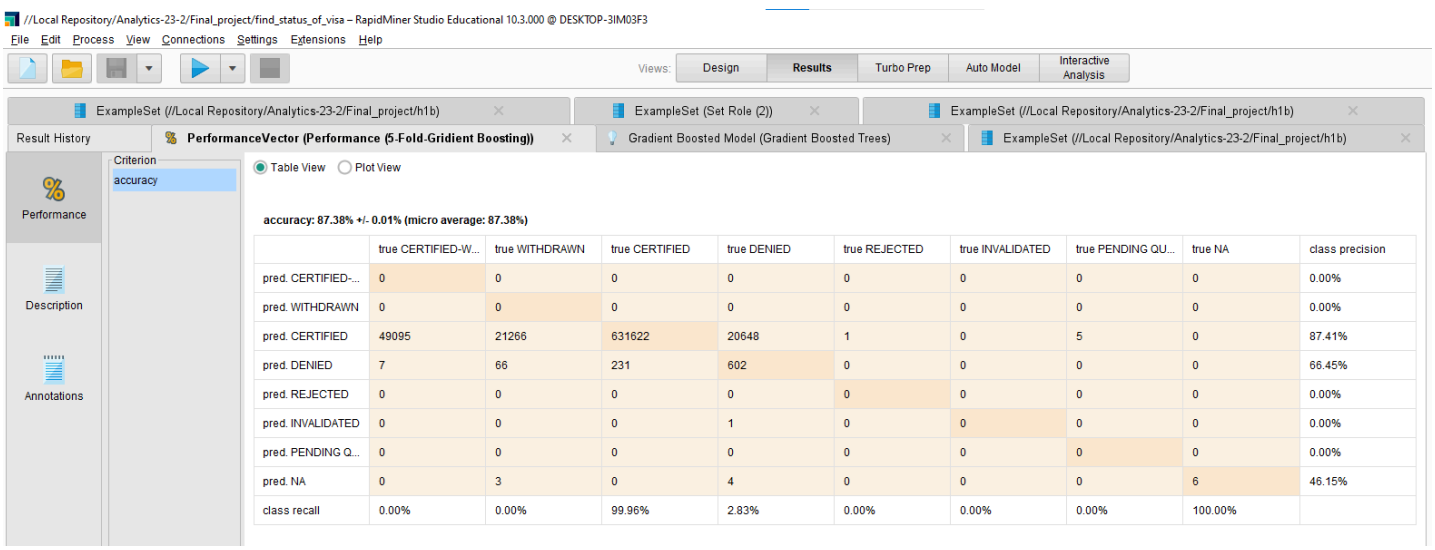


Picture4.6: Marketplace in Rapidminer Studio 10.3

Chapter 5
Result, data presentation and conclusion

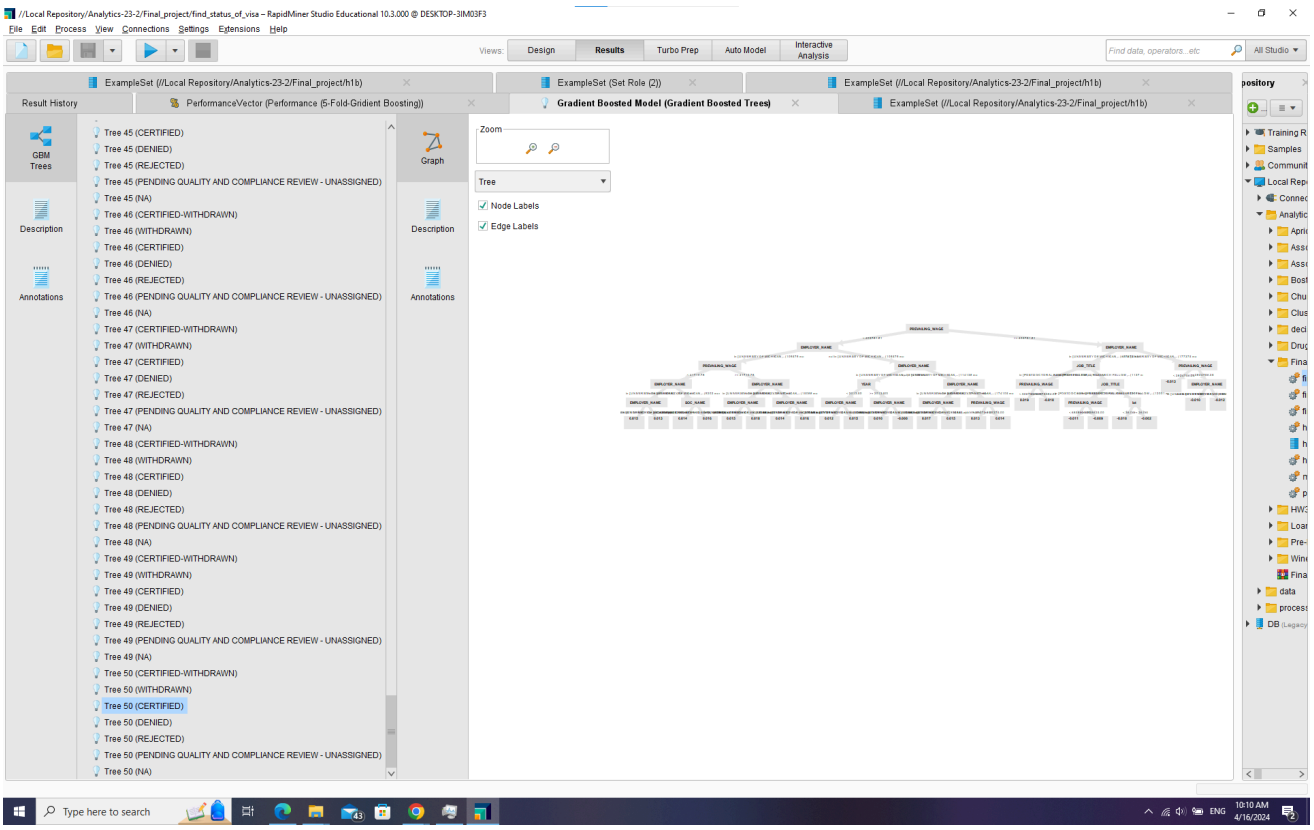
5.1 Result base on 4.1 process





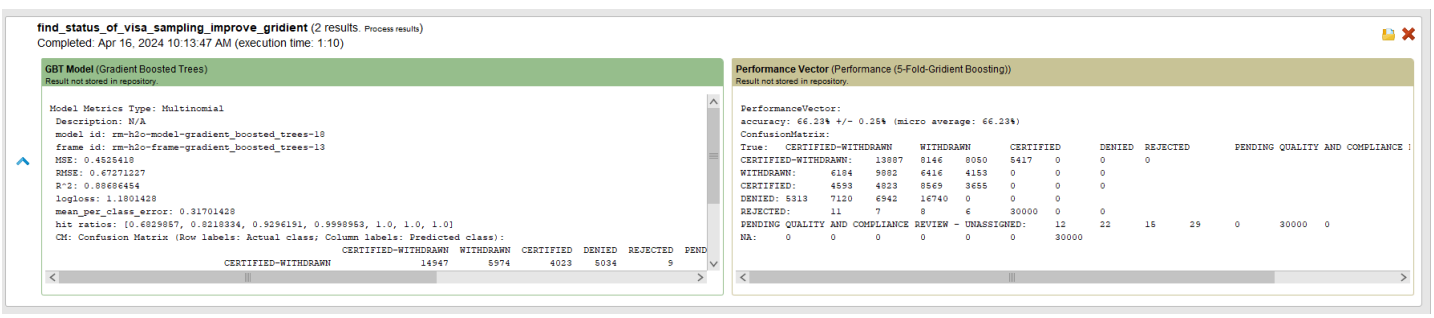
	true CERTIFIED-W...	true WITHDRAWN	true CERTIFIED	true DENIED	true REJECTED	true INVALIDATED	true PENDING QU...	true NA	class precision
pred. CERTIFIED-...	0	0	0	0	0	0	0	0	0.00%
pred. WITHDRAWN	0	0	0	0	0	0	0	0	0.00%
pred. CERTIFIED	49095	21266	631622	20648	1	0	5	0	87.41%
pred. DENIED	7	66	231	602	0	0	0	0	66.45%
pred. REJECTED	0	0	0	0	0	0	0	0	0.00%
pred. INVALIDATED	0	0	0	1	0	0	0	0	0.00%
pred. PENDING Q...	0	0	0	0	0	0	0	0	0.00%
pred. NA	0	3	0	4	0	0	0	6	46.15%
class recall	0.00%	0.00%	99.96%	2.83%	0.00%	0.00%	0.00%	100.00%	

From above, the result by design filters 723,557 samples to make 50 gradient boosted trees and tests by 5-fold cross validation the result of time executed is 3 minutes and 51 seconds. With the accuracy of prediction at 84.38%. However the result in true denied for recall is very low. Our prediction is that the reason for this may come from the low number of examples to predict the denied or too close to predict much about how it differs between status. But this case has a condition that certified-withdraw has the same status as certified that the accuracy on our design must be higher than this. And the tree result is below as an example. Also with the high accuracy may come from that a lot of data here at majority is certified.



5.2 Result base on 4.2 process

5.2.1 At sampling 30000 samples per class.



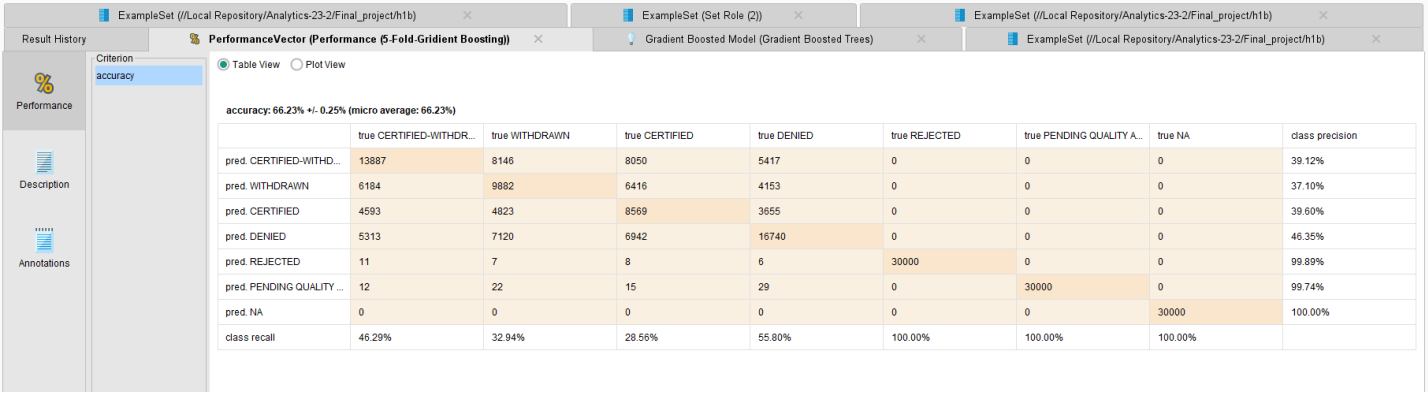
find_status_of_visa_sampling_improve_gridient (2 results. Process results)
Completed: Apr 16, 2024 10:13:47 AM (execution time: 1:10)

GBT Model (Gradient Boosted Trees)
Result not stored in repository.

```
Model Metrics Type: Multinomial
 Description: N/A
 model id: rm-h2o-model-gradient_boosted_trees-18
 frame id: rm-h2o-frame-gradient_boosted_trees-13
 MSE: 0.4525418
 RMSE: 0.67271227
 R^2: 0.88686454
 logloss: 1.1801428
 mean_per_class_error: 0.31701428
 hit ratios: [0.6829857, 0.8218334, 0.9296191, 0.9998953, 1.0, 1.0, 1.0]
 CM: Confusion Matrix (Row labels: Actual class; Column labels: Predicted class):
                                                CERTIFIED-WITHDRAWN  WITHDRAWN  CERTIFIED  DENIED  REJECTED  PEND
                          CERTIFIED-WITHDRAWN                14947       5974       4023    5034         9
```

Performance Vector (Performance (5-Fold-Gridient Boosting))
Result not stored in repository.

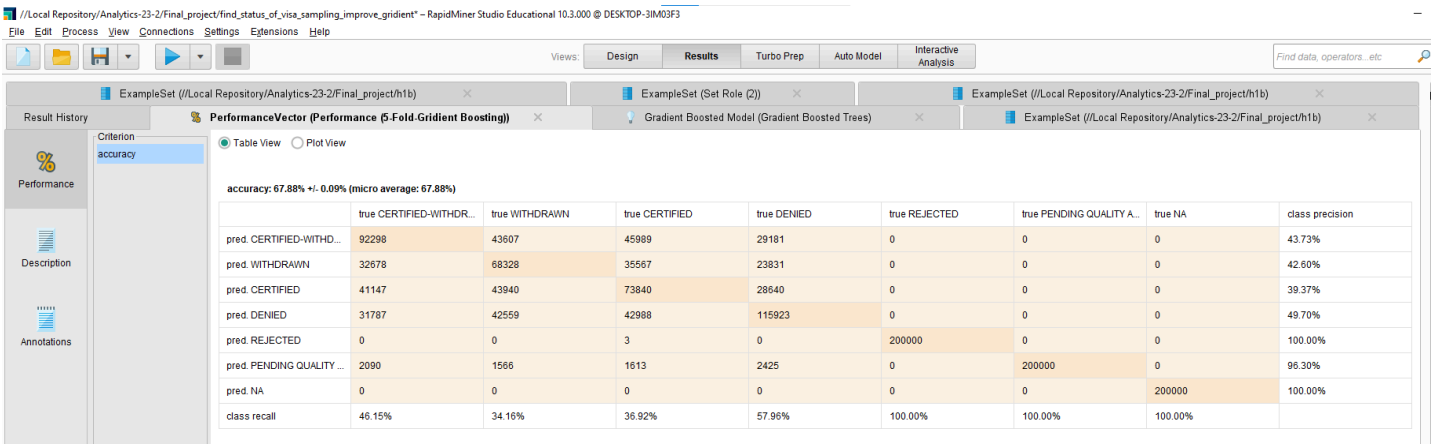
```
PerformanceVector:
accuracy: 66.23% +/- 0.25% (micro average: 66.23%)
ConfusionMatrix:
True:   CERTIFIED-WITHDRAWN    WITHDRAWN      CERTIFIED      DENIED  REJECTED       PENDING QUALITY AND COMPLIANCE :
CERTIFIED-WITHDRAWN:    13807   8146    8050    5417    0       0       0
WITHDRAWN:      6184    9882    6416    4153    0       0       0
CERTIFIED:      4553    4823    8569    3655    0       0       0
DENIED: 5313    7120    6942    16740   0       0       0
REJECTED:       11      7       8       6       30000   0       0
PENDING QUALITY AND COMPLIANCE REVIEW - UNASSIGNED:     12      22      15      29      0       30000   0
NA:     0       0       0       0       0       0       30000
```



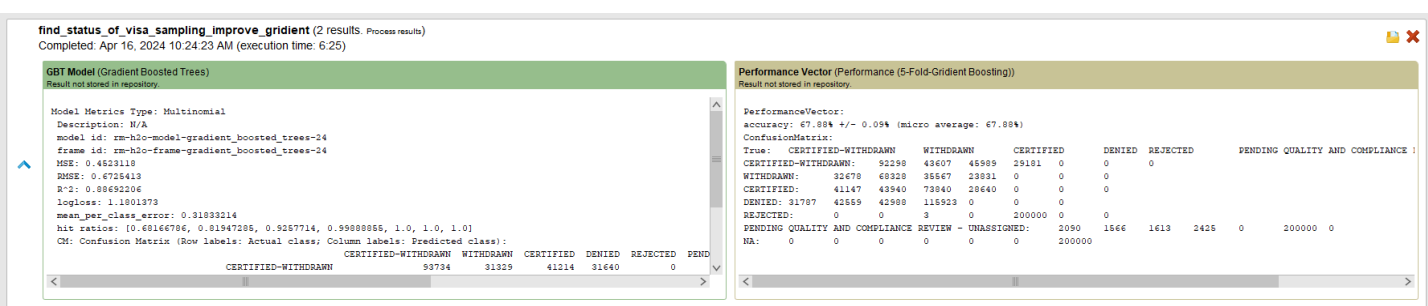
	true CERTIFIED-WITHDR...	true WITHDRAWN	true CERTIFIED	true DENIED	true REJECTED	true PENDING QUALITY A...	true NA	class precision
pred. CERTIFIED-WITHD...	13887	8146	8050	5417	0	0	0	39.12%
pred. WITHDRAWN	6184	9882	6416	4153	0	0	0	37.10%
pred. CERTIFIED	4593	4823	8569	3655	0	0	0	39.60%
pred. DENIED	5313	7120	6942	16740	0	0	0	46.35%
pred. REJECTED	11	7	8	6	30000	0	0	99.89%
pred. PENDING QUALITY ...	12	22	15	29	0	30000	0	99.74%
pred. NA	0	0	0	0	0	0	30000	100.00%
class recall	46.29%	32.94%	28.56%	55.80%	100.00%	100.00%	100.00%	

After we use hybrid sampling to extend the minor value of the status to have more examples,the runtime has a result at high satisfied speed. Also with the result that the prediction may seem like it has lower accuracy. But by the model we can see that the smaller of visa status has a better result of prediction but trade with the overall accuracy.

5.2.2 At sampling 200000 samples per class.



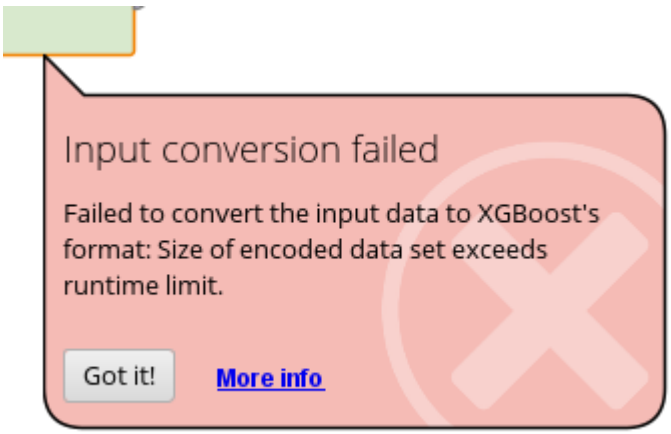
	true CERTIFIED-WITHDR...	true WITHDRAWN	true CERTIFIED	true DENIED	true REJECTED	true PENDING QUALITY A...	true NA	class precision
pred. CERTIFIED-WITHD...	92298	43607	45989	29181	0	0	0	43.73%
pred. WITHDRAWN	32678	68328	35567	23831	0	0	0	42.60%
pred. CERTIFIED	41147	43940	73840	28640	0	0	0	39.37%
pred. DENIED	31787	42559	42988	115923	0	0	0	49.70%
pred. REJECTED	0	0	3	0	200000	0	0	100.00%
pred. PENDING QUALITY ...	2090	1566	1613	2425	0	200000	0	96.30%
pred. NA	0	0	0	0	0	0	200000	100.00%
class recall	46.15%	34.16%	36.92%	57.96%	100.00%	100.00%	100.00%	

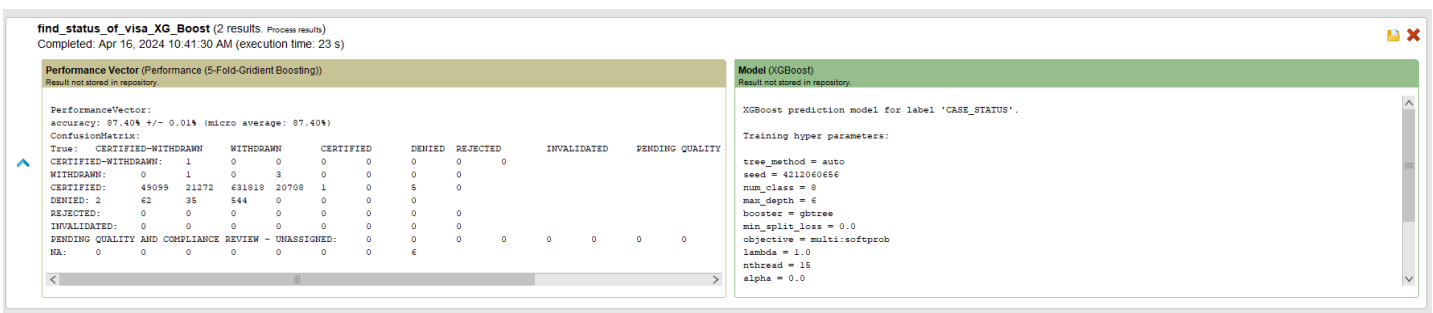


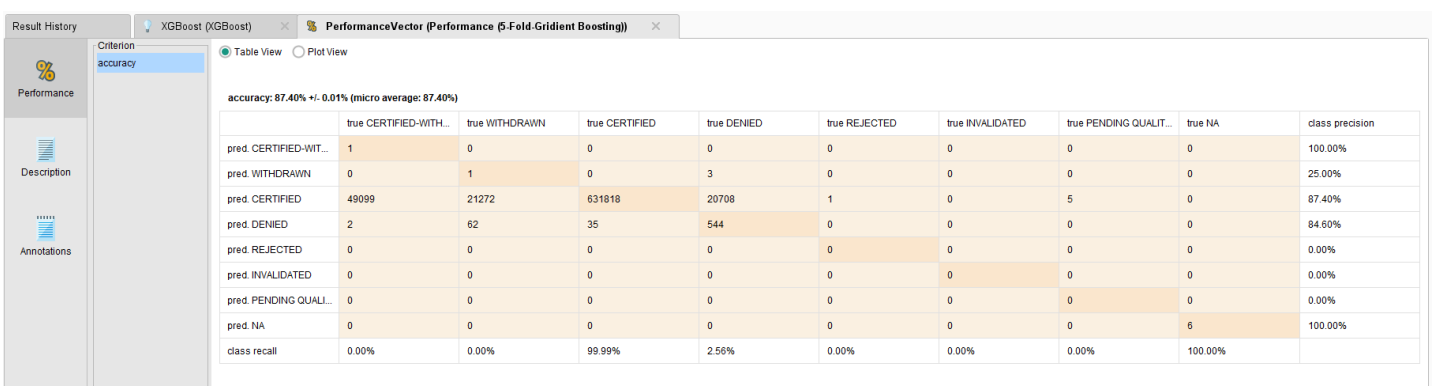
After we successfully went on the process of 5.2.1. We decided to extend our

sampling by using the same technique as process 5.2.1 but change the visa status example from 30,000 to 200,000 per class to prove how our model can make the prediction better. The result has a better view on the execution. However with the small percentage of improvement, we have to spend 6 times of execution time to run the process out as a result. But the small class of the prediction has a better performance in both precision, recall and accuracy

5.3 Result base on 4.3 process







	true CERTIFIED-WITH...	true WITHDRAWN	true CERTIFIED	true DENIED	true REJECTED	true INVALIDATED	true PENDING QUALIT...	true NA	class precision
pred. CERTIFIED-WIT...	1	0	0	0	0	0	0	0	100.00%
pred. WITHDRAWN	0	1	0	3	0	0	0	0	25.00%
pred. CERTIFIED	49099	21272	631818	20708	1	0	5	0	87.40%
pred. DENIED	2	62	35	544	0	0	0	0	84.60%
pred. REJECTED	0	0	0	0	0	0	0	0	0.00%
pred. INVALIDATED	0	0	0	0	0	0	0	0	0.00%
pred. PENDING QUALI...	0	0	0	0	0	0	0	0	0.00%
pred. NA	0	0	0	0	0	0	0	6	100.00%
class recall	0.00%	0.00%	99.99%	2.56%	0.00%	0.00%	0.00%	100.00%	

From above the execution time XG Boost tree can make a prediction rapidly by spending only 23 seconds when compared with ordinary gradient boost tree at 3

minutes and 51 seconds. In the view of the result, they have the same significant result as the result in section 5.1. But the disadvantage is that the data from this model does not show the design of the tree same as the factory install module of gradient boost tree that originated in rapidminer 10.3. The module has many differences in the interface result when compared to the pre-install of gradient boost tree.

Also with the critical disadvantage of this module is when the process decides that the runtime limit is not enough on our device the program will stop the process and have a failure on the execution. We have to lightweight our feature to make the process run over the runtime that may bring to missing some of the factor in some cases.

5.4 Conclusion

From all of the above results we can conclude that the application that we make from the dataset has not been much satisfied because of the limitation of the machine and the many decision data of the dataset. So we decided to make various ways to display the possibility of making a model of prediction to show equality prediction in every class. And also with the improvement of the performance together with our data. With on our research and development on the design of model, the comparison between all of result are display at below

Characteristic	Only gradient boost tree	Gradient boost tree with 30000 hybrid sampling	Gradient boost tree with 200,000 hybrid sampling	XG boost tree
Runtime speed	3:51	1:10	6:25	0:23
Accuracy (%)	87.38	66.23	67.88	87.40
Focusing of recall and precision on small class	Poor	Good	Better	Poor
Visualization	Provide tree model	Provide tree model	Provide tree model	None

5.5 Suggestion

From all of what we have done on this project. We suggest how to improve the data. After group discussion we can make some suggestions like below.

1. Use discretization by user specification to make the latitude and longitude to the region of the US such as east coast, west coast, midwest, north and south. To visualize the region which can be seen in the picture of grouping better by the mathematical.
2. Clustering is still a good option to predict when compared with all of the process tree.
3. Hybrid sampling should be made since data preprocessing to make the

result better. However the machine cannot accept too much high workload that we decide to make it as a separate process.

4. The data, if the time and all of the performance can do. We suggest using the original dataset from the department of labor instead to make it more complete in the model. Some of the data that is provided on Keggle may cut some very important data to make the model more efficient.
5. The cross check is a mandatory requirement to make a project lowest error. This process should be in every loop of development and require more than one person to avoid human error in every stage of the project.

References

[1]https://www.dol.gov/sites/dolgov/files/ETA/oflc/pdfs/H-1B_Selected_Statistics_FY2019_Q4.pdf

[2]https://www.kaggle.com/datasets/nsharan/h-1b-visa

[3]https://en.wikipedia.org/wiki/List_of_regions_of_the_United_States

[4]https://www.geeksforgeeks.org/gradientboosting-vs-adaboost-vs-xgboost-vs-catboost-vs-lightgbm/

[5]https://medium.com/kbtg-life/tree-based-algorithms-%E0%B9%81%E0%B8%9A%E0%B8%9A-high-level-4058e909e0c5

[6]https://blog.pjjop.org/modern-regularization-with-data-augmentation-batch-normalization-and-dropout/

[7]https://grad.dpu.ac.th/upload/content/files/year9-3/9-30.pdf