Business Problem Proposal

Amazon Employee Access Authorization

Amazon.com- Employee Access Challenge

Background:

Often employees who join need access to a variety of resources. At time of joining, they need to be provided with an access to resources they need, to be able to complete their work. This access may be in terms of reading/ manipulating resources and employees working in a particular department are often said to require the same type of resources. Often the access needs of employees are figured out as and when they carry out their daily work.

Employees may get stuck at places due to access restrictions and need to contact their senior officials to grant them the required access to resume their work. As the situation is expanded to a larger sphere of the company, a lot of time and money is wasted to overcome access restrictions. Employees are not able to log on the system or submit useful changes to the system which can lead to a failure to meet deadline and much worse results.

I studied the trends in the resource access in various technology giants like Google, Amazon, Walmart etc. Having observed the trends and traditions, it seems that the amount of time is wasted for both the employee and the manager in assigning resources is very large. This time if utilized somewhere else can be more productive.

Thus it is important to formulate a system to help complete access requirements. There is a large amount of data being collected everyday based on the employee rank, department, and role and access requirements. Given this data and provisioned access, it can be comparatively easy to figure out and determine an employee's access needs when he/she is a newcomer or an employee is leaving the company. Such a system can help reduce human involvement used to grant access to employees or revoke it and thus save a lot of time and money.

Acknowledgement:

I would like to thank Dr Arbi Ghazarian for his assistance in the project and selection of data set. The data set used is taken from the Kaggle website. A larger segment of the data is also available at the UCI machine learning data set repository. I would also like to thank the data creator and donator Ken Montanez from information security department, Amazon Corp for providing the data set.

Kanika Mathur

Objective (The business problem):

The basic objective of our business problem is to build a model, which learns through historical data(training set) determines employee access needs such that manual access grants and revokes are minimized as employee status changes with time.

How will this help the business?

It will help save a lot of time of the human time and money. Resource access can be done with least human intervention which will help boost the employee working and output. Thus it will help improve efficiency.

Thus we plan to build a system to take the place of resource administrator and save time and money.

I/P: Employee role information and resource code

O/P: Returns whether access is granted or not

The model can be extended easily to make even more predictions about the data.

How will a normal user use the system?

A normal user (the employee) will ask for access to a particular resource, the request will be processed by the above model, which has all the information of the employee, his role, his manager, his role description etc and based on these attributes, predictions will be made whether or not to grant access to that particular employee or not. Maximum cases will be sufficed by the system, for cases which need a special consideration, they can be redirected to the manager for evaluation. This will thus help in reducing human in the resource allocation system and help save time and money.

Measure of success for the system?

Success rate for the system will be determined by the number of cases the system accurately predicts the resource allocation. The number of human intervention in the system is inversely proportional to the success of the system.

Less the number of human intervention needed in the system, more is the success rate of the system.

What Business entity does an instance/example correspond to?

The examples in the data have attributes which describe the various features of the employee eg his role, his ID, his manager, his description. After all the data is stored for an employee, we have an attribute ACTION which predicts the decision of the system

We even describe the attributes as having numerical values.

The result also predicts the value in a numeric system, 0= resource denied and 1= resource granted.

The real nature of data is for a person. It has various features of an employee. Decisions made from this system further correspond to what the employee will be able to access and what not.

Description of Data(Are the attributes defined precisely):

The data is provided by Amazon Inc in 2010-11 published by Kaggle platform. There are three .csv files

1) **The training set**:
   - ➔ It consists of 32769 rows (samples). Each of the training set sample has one label attribute namely "ACTION" which takes up values 1 or 0.
     1 is the value when the resource request is granted
     0 is the value when the resource request is disapproved.
   - ➔ Additionally each training set sample has attributes as described in the table below: The attributes have been defined very precisely.

Table 1: Attribute Description

Attribute_name	Description
ACTION	It is the final outcome of the model. An output of 1= request approved and 0= request denied.
RESOURCE	It refers to the ID of the resource.

MGR_ID	It is the employee ID of the manager of the particular employee under consideration. ** An employee can have only one manager at a time.
ROLE_ROLLUP_1	It refers to the company role grouping category ID 1 (eg US Engineering).
ROLE_ROLLUP_2	It refers to the Company Role grouping category ID 2 (eg US Retail).
ROLE_DEPTNAME	It refers to the description of the role of the department (eg Retail).
ROLE_TITLE	It refers to the company role business title description (eg Senior Engineering Retail Manager).
ROLE_FAMILY_DESC	It refers to the Company role family extended description (eg Retail manager, Software Engineering).
ROLE_FAMILY	It refers to the Company role family description (eg Retail Manager).
ROLE_CODE	It refers to the Company role code (eg manager). ** ROLE_CODE is unique to each role.

2) **<u>The testing set:</u>**

It contains the data set on which testing needs to be done. We build the model which learns from historical data and then the model is used to test the outcome on the testing set data.

It has 58921 test set samples for which the model needs to be tested. Also included are the various attributes of the data which are as listed in the table below:

Table 2: Attribute Description for test set

Attribute name	Description
ID	It has the ID of the employee for which we need to predict whether a resource has to be allocated or not.
RESOURCE	It refers to the ID of the resource.

MGR_ID	It is the employee ID of the manager of the particular employee under consideration. ** An employee can have only one manager at a time.
ROLE_ROLLUP_1	It refers to the company role grouping category ID 1 (eg US Engineering).
ROLE_ROLLUP_2	It refers to the Company Role grouping category ID 2 (eg US Retail).
ROLE_DEPTNAME	It refers to the description of the role of the department (eg Retail).
ROLE_TITLE	It refers to the company role business title description (eg Senior Engineering Retail Manager).
ROLE_FAMILY_DESC	It refers to the Company role family extended description (eg Retail manager, Software Engineering).
ROLE_FAMILY	It refers to the Company role family description (eg Retail Manager).
ROLE_CODE	It refers to the Company role code (eg manager). ** ROLE_CODE is unique to each role.

3) **The submission set**:
The submission set is a .csv file which contains the ID of the Employee and the Action associated with it.
Action=1 refers to the approval of the request for resource.
Action =0 refers to the denial of request of resource.
It has 58921 test sets for which we need to predict the outcomes.

Values various attributes can take:

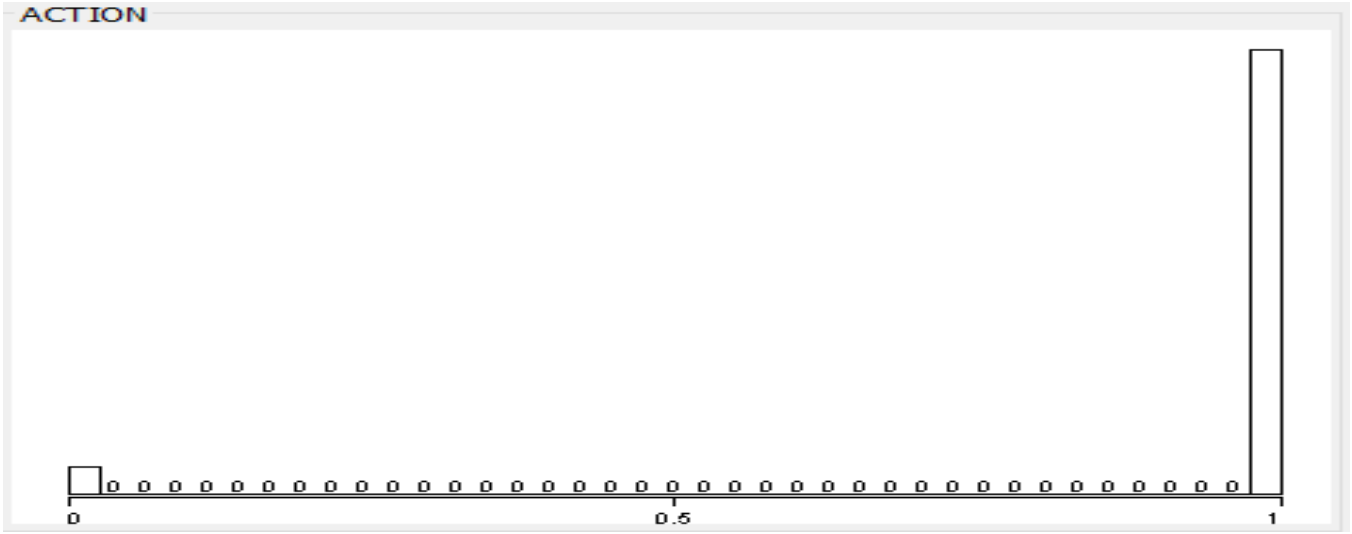
Table 3: Values the attributes can take

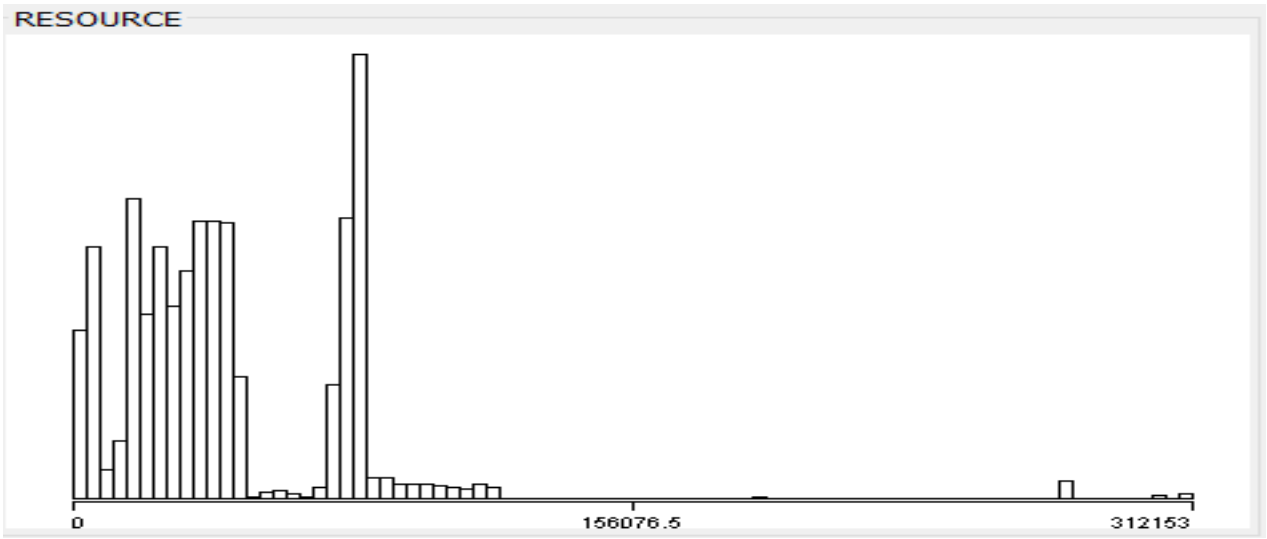
Attribute name	Value
ID	It can have values from 1 to infinity (For a company growing infinitely). The value cannot be negative.
RESOURCE	It can have values from 1 to infinity (For a company growing

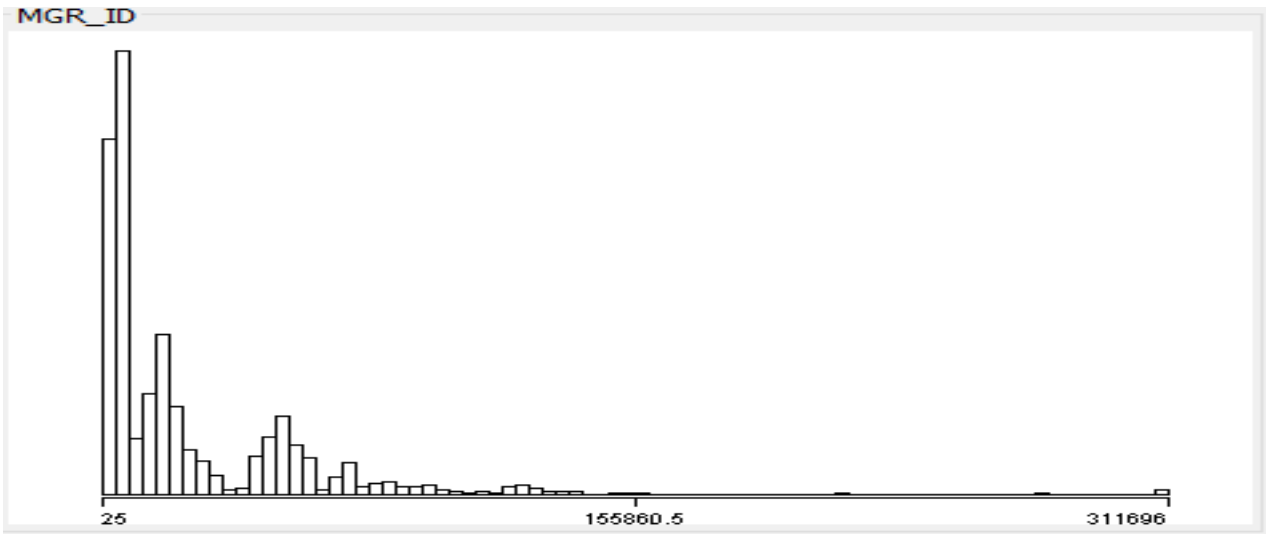
	infinitely). The value cannot be negative.
MGR_ID	It can have values from 1 to infinity (For a company growing infinitely). The value cannot be negative.
ROLE_ROLLUP_1	It can have values from 1 to infinity (For a company growing infinitely). The value cannot be negative.
ROLE_ROLLUP_2	It can have values from 1 to infinity (For a company growing infinitely). The value cannot be negative.
ROLE_DEPTNAME	It can have values from 1 to infinity (For a company growing infinitely). The value cannot be negative.
ROLE_TITLE	It can have values from 1 to infinity (For a company growing infinitely). The value cannot be negative.
ROLE_FAMILY_DESC	It can have values from 1 to infinity (For a company growing infinitely). The value cannot be negative.
ROLE_FAMILY	It can have values from 1 to infinity (For a company growing infinitely). The value cannot be negative.
ROLE_CODE	It can have values from 1 to infinity (For a company growing infinitely). The value cannot be negative.
ACTION	It can take up value 0 or 1.

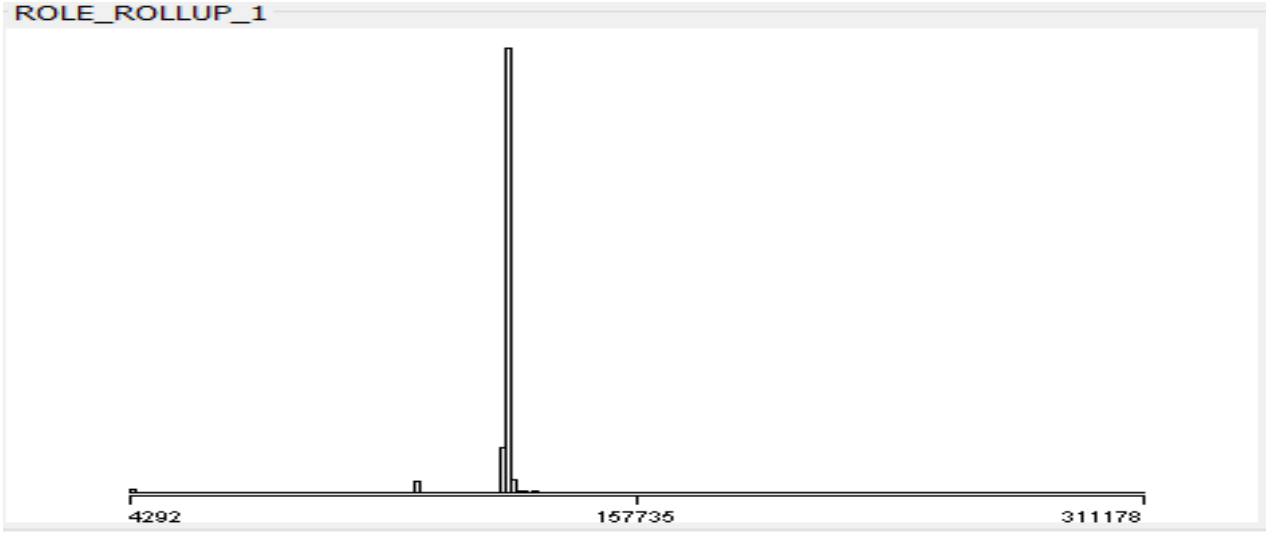
Attribute	**Minimum**	**Maximum**	**Mean**	**StdDev**
ACTION	0	1	0.942	0.234
RESOURCE	0	312153	42923.916	34173.893
MGR_ID	25	311696	25988.958	35928.032
ROLE_ROLLUP_1	4292	311178	116952.628	10875.564
ROLE_ROLLUP_2	23379	286791	118301.823	4551.589
ROLE_DEPTNAME	4674	286792	118912.78	18961.323
ROLE_TITLE	117879	311867	125916.153	31036.466
ROLE_FAMILY_DESC	4673	311867	170178.37	69509.462
ROLE_FAMILY	3130	308574	183703.409	100488.407
ROLE_CODE	117880	270691	119789.43	5784.276

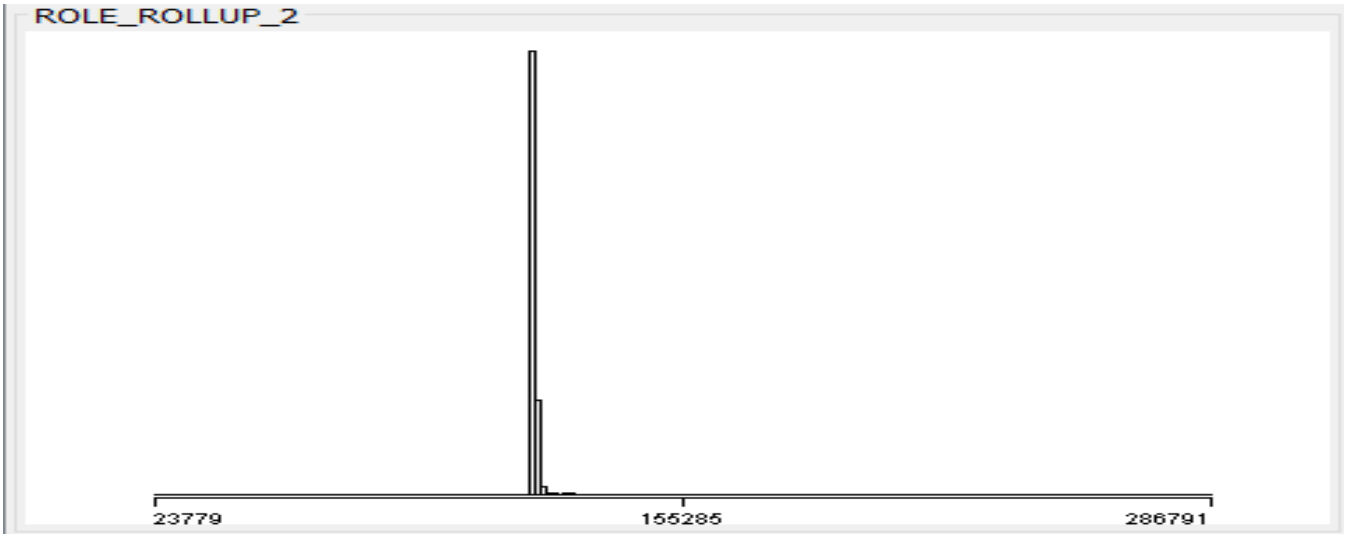
Analyzing the attributes using histograms:

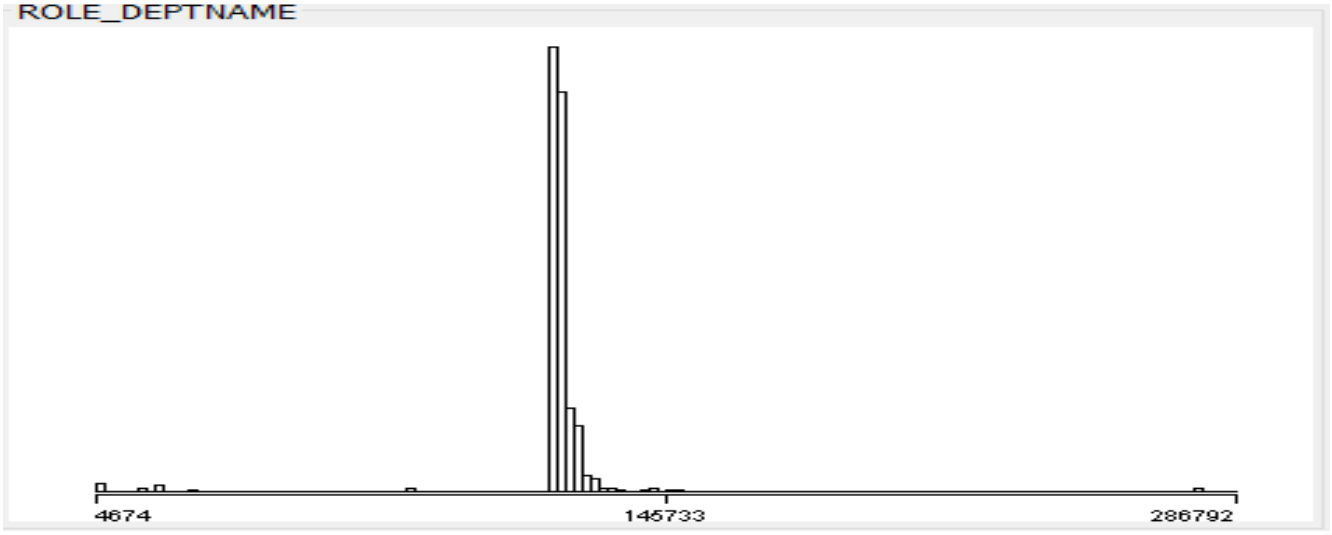

ACTION
0
0.5
1


RESOURCE
0
156076.5
312153

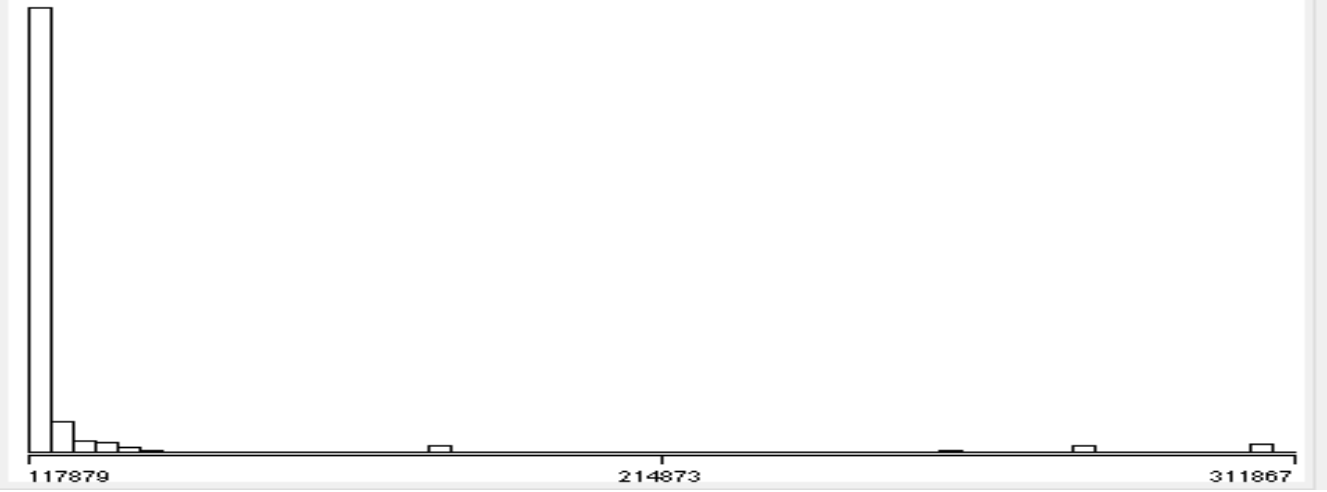

MGR_ID
25
155860.5
311696


ROLE_ROLLUP_1
4292
157735
311178

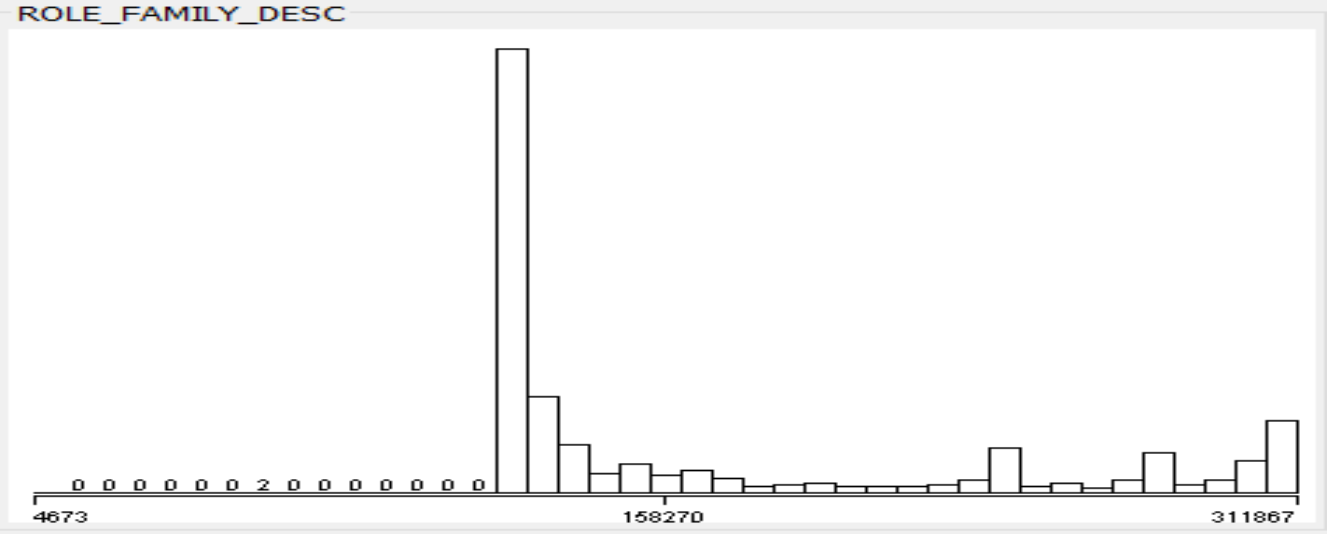

ROLE_ROLLUP_2
23779
155285
286791


ROLE_DEPTNAME
4674
145733
286792

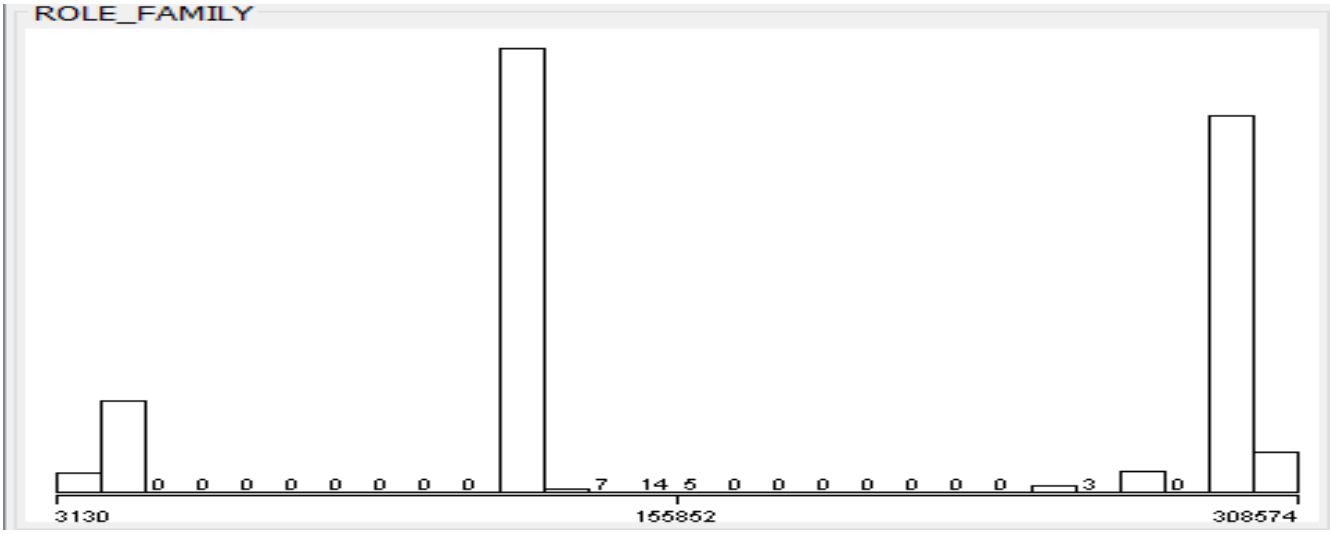
ROLE_TITLE

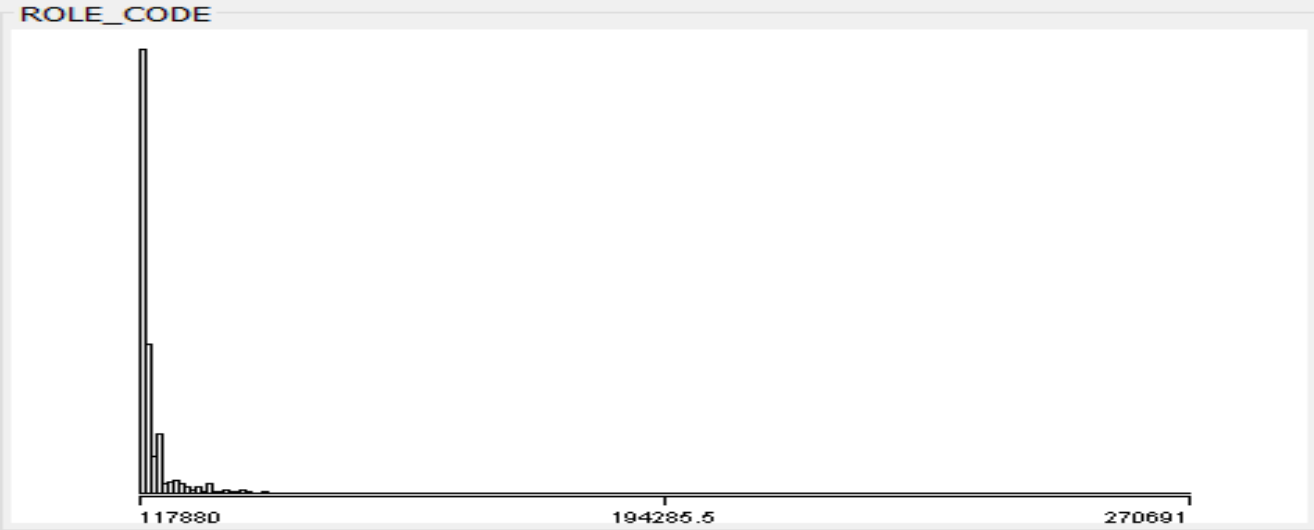


ROLE_FAMILY_DESC



ROLE_FAMILY





Is the data science solution formulated appropriately to solve this business problem?

The given data set is unbalanced with the number of instances in one class being much more than that of the other. This is an instance of unbalanced classification problem which can be solved using various techniques. The main approach is to try to balance the distribution between the two classes.

In the training set data, we have ACTION=0 has less instances as compared to ACTION =1.

In such a case, where most learning algorithms have a principle to minimize the error rate to which minority class contributes less, would result in a problem and error. So balancing the data set is of utmost importance.
We can think of using techniques like oversampling and undersampling to balance data.

Is the problem a supervised or unsupervised problem?

The problem is a supervised problem with the target variable clearly defined. It is **ACTION.**

The target variable can take two values 0= resource access denied

1= resource access granted

For supervised problems, will modelling this target variable improve the stated business problem? An important subproblem? If the latter, is the rest of the business problem addressed?

Yes modelling of this target variable will help solve the business problem. It will help save time and money and also allow easy allocation of resources to employees during their transition time (as they join in as new employees or they leave the company).

The information gain on each variable will be really enough to come to the predictive nature of attributes. Cost of opportunity is higher for the system as it helps reduce human intervention in the resource allocation and diverting human intervention to a different sphere will help improve outputs and thus profits.

If unsupervised, is there an "explanatory data analysis " path well defined?

The data we are working on in this problem is supervised data. We could further extend the scope to derive other useful techniques from data where we could use unsupervised learning.

USE CASES:

Use Case 1: New Employee hire

Objective: Hire of a new employee

Primary Actor: New recruit

Dependencies: Hiring of a new recruit leads to creating of an entry in database. His name should be included in the database.

Trigger: New recruit joins the company

Secondary Actors: Resource manager

Preconditions:

1. There should be a new employee who is hired.
2. The records of the employee should not be there previously on the database.
3. The employee should be under some manager and should have value of all the attributes needed.

Post Condition(s):

1. The employee data is successfully added into the system.
2. The employee is allocated resources depending on predictions from the system.

MAIN SUCCESS SCENARIO

1. On a new employee hire, there is a new entry created in the database.
2. All the employee records are taken and added into the database.(Previously there should not be a matching entry).

3. Depending on the outcome of the resource allocation model, appropriate resources are allocated to the system.

VARIATIONS

Variation ID: *UC1 VAR 1.1*

1. *If the employee data is not available, the employee is asked to get the data first and then it is stored in the system.*

FAILURE VARIATIONS

Variation ID: *UC 1 VAR 1.2*

1. 1. If the employee data is incomplete, the data will not be accepted by the system and an exception will be raised.
2. If there is malicious data, an exception can be raised by the system.

1.

BUSINESS RULES

1. An employee has specific access rights which are provided in the database separately.
2. An employee has specific access rights which are provided in the database separately.

NOTES: This use case defines how to incorporate a new entry in the database on a new employee joining.

Use Case 2: Resource request

Objective: An employee requests for a resource

Primary Actor: Employee

Dependencies: Employee requests a resource, database needs to be checked and appropriate decision made.

Trigger: Resource request

Secondary Actors: Resource manager

Preconditions:

a) An employee has his details in the system.
b) He requests a resource.

Post Condition(s):

a) Decision is made whether to allow or deny the resource access.

MAIN SUCCESS SCENARIO

a) An employee asks for a resource.
b) His information is checked in the system and appropriate decision is made.

VARIATIONS

Variation ID: *UC2 VAR 2.1*

a) If the employee data is not available, the employee is asked to get the data first and then it is stored in the system. Only after inclusion of data in system will his details be checked and decision made whether to grant access or deny it.

FAILURE VARIATIONS

Variation ID: *UC 2 VAR 2.2*

a) If the employee data is incomplete/ unavailable, the data will not be accepted by the system and an exception will be raised. He cannot ask for resources in this scenario.
b) If the employee is not authorized to access the resource, his request is denied.

BUSINESS RULES

1) An employee has specific access rights which are provided in the database separately.

NOTES: This use case defines how to answer to user requests to access resources.

Use Case 3: Employee leaves company

Objective: Employee leaving company

Primary Actor: Employee

Dependencies: When an employee leaves the company, his records should be deleted from the system.

Trigger: Employee leaving the company.

Secondary Actors: Resource manager

Preconditions:

1) There should be an employee who is leaving.
2. The records of the employee should be there previously on the database.

Post Condition(s):

1) The employee data is successfully deleted from the system.

MAIN SUCCESS SCENARIO

a) When an employee decides to leave the company, his records need to be deleted from the system.
b) He initiates his resignation.
c) The system deletes his records from the database.

VARIATIONS

Variation ID: *UC3 VAR 3.1*

1) *If the employee data is not available, no changes are made.*

FAILURE VARIATIONS

Variation ID: *UC 3 VAR 3.2*

1) 1) If the employee has a change of department, his previous access rights can be taken and new allotted.

3.

BUSINESS RULES

1) If a new user is leaving from the system, his details are deleted from the system.

NOTES: This use case defines how to delete an entry from the system for an employee who is leaving.

Conclusion:

The aim of the proposal is to develop the model whose features are defined above. Once the implementation of the system finishes, it will help save time and money and human skills can be used in a more progressive manner.

Phase II(Data Preparation)

1) **Will it be practical to get values for attributes and create feature vectors, and put them into a single table?**

Answer) My data set already has values for all the attributes, so this is not an issue for me. My data set is enough to apply data mining techniques and solve the problem.

2) **If not, is an alternative data format defined clearly and precisely? Is this taken into account in the later stages of the project? (Many of the later methods/techniques assume the dataset is in feature vector format.)**

Answer) My data is in .csv format. So I have a clear and precise format for data. This will suffice for rest all stages of the project.

3) **If the modeling will be supervised, is the target variable well defined? Is it clear how to get values for the target variable (for training and testing) and put them into the table?**

Answer) My data set is taken from the Kaggle website. It involves supervised learning and has a clearly defined target variable. Also the test and training sets are given separately. So this is not an issue for me and will suffice for my entire project.

4) **How exactly will the values for the target variable be acquired? Are there any costs involved? If so, are the costs taken into account in the proposal?**

Answer) My data set has clear interpretation of the target variable and also clearly provides its values in the training and test data set.

There were no costs involved in data collection, as I took the data from an online repository which provides free data set for use and research.

5) **Are the data being drawn from the similar population to which the model will be applied? If there are discrepancies, are the selection biases noted clearly? Is there a plan for how to compensate for them?**

Answer) I have taken the data set from an online repository. The test and training set have been clearly segregated. I have to use the data as it is specified and will work on the assumption that data provided on the website has no bias for the population in devising the test and training data set.

The data set, given on the website is an imbalanced data set (more of yes->grant of access to resource rather than no). But I will try to find the answer to utmost perfection and use various evaluation frameworks to analyze the results.

Phase III(Modeling)

1) **Is the choice of model appropriate for the choice of target variable? —Classification, class probability estimation, ranking, regression, clustering, etc.**

Answer) Yes the choice of problem is appropriate for the system. I have applied techniques for both supervised and unsupervised learning. In supervised learning I have applied classification, regression while in unsupervised I have used clustering.

In supervised learning my output is to predict whether a resource can be allocated to a person or not. This is predicted accurately by classification. Regression also provides good estimates about the access of resource.

For the second part of the problem I used clustering (unsupervised learning). I have first removed the target variable and then tried to cluster my data. I get clusters based on which I can identify people to be put under the same resource manager.

Clustering provides a good estimate of people of similar type and thus I can easily allot a single resource manager for them.

2) Does the model/modeling technique meet the other requirements of the task?

a. Generalization performance, comprehensibility, speed of learning, speed of application, amount of data required, type of data, missing values?

Answer) Yes, the model meets the requirements of the task.

The performance of the model is fairly good.

I do not need a 100 percent accuracy because there can always be human intervention in the system. But the accuracy here will be defined by least human intervention in the system. It gives a decent number of correct predictions in a fairly less amount of time. Speed of learning for all techniques is good, except the fact that since I have a slow processor, the model takes time in building over a large dataset (32 thousand records), so the speed of application decreases.

There were no missing values and so the data did not need pre-processing. The type of data is numeric and as a part of the first subtask needs to be divided into two categories allow or deny access of resource. Regression has also been used for it. For the second part, we need to group people to identify whether they can be allowed to work under the same resource manager. Clustering has been used for it.

b. Is the choice of modeling technique compatible with prior knowledge of problem (e.g., is a linear model being proposed for a definitely nonlinear problem)?

Answer) Yes the modeling technique used is compatible with the prior knowledge of problem. For numeric output, we have used regression (which gives output in binary). Also classification is used to get categorical output.

3) **Should various models be tried and compared (in evaluation)?**

Answer) Yes various models can be tried and compared. This basically gives us an estimate of the accuracy of prediction we can get. I have used Area under the curve as the basis for my accuracy. Also I have created table to tell the true positive, false positive, true negative, false negatives. At the end if we can select the model which is the most accurate on giving outputs.

Eg For decision tree, we have AUC as 0.57, 0.47 for KNN and 0.65 for SVM. So both of classification techniques give around same performance.

4) **For clustering, is there a similarity metric defined? Does it make sense for the business problem?**

Answer) Yes clustering has a similarity metric. We have used Jaccard similarity. It is making sense for the problem as we get a proper cluster formation so as to divide people under resource managers to handle similar types of requests.

Phase IV(Evaluation and Deployment)

● **Is there a plan for domain-knowledge validation?**

Answer) Yes there is a very well laid plan for domain knowledge validation. Before starting the business problem, a detailed research was carried out for the domain of the problem. So I am well aware of the needs of the problem. Thus whenever I got an output from my model, I have compared it to the need and usage in the resource management field.

a) Also I would like to mention that the entire output and my target variable are strictly in conformation with the exact needs of the domain.
b) For all the techniques, either one of the testing tools has been used such as Confusion Matrix, RUC, AUC or mean squared error. The usage of testing techniques enabled me to determine which model is best fit for the business problem.

—Will domain experts or stakeholders want to vet the model before deployment? If so, will the model be in a form they can understand?

Answer) Yes stakeholders or domain experts will want to vet the model before deployment. It is in the form that they can clearly understand.

a) The target variable (ACTION) clearly explains its meaning. ACTION refers to whether the resource will be allocated or not. If ACTION =0, it will not be allocated and if it is 1, it will be allocated.
b) Also the model made is of the form that it is easy to understand to all the stakeholders. Each line of code has proper comments. Additionally model gives clear output, as to what will the resource allocation result be depending on the attribute values.

● Is the evaluation setup and metric appropriate for the business task? Recall the original formulation.

Answer) Yes the evaluation step and metric is appropriate for the given business problem. For the classification model, AUC has been calculated. It gives a clear idea of the prediction correctness.

Also while allocating employees under resource managers; I have considered them to initially have 2 resource managers. They have then been clustered in two clusters. As and when the need of the business increases, it can also be increased.

The system developed is not a hard real time model and can tolerate little bit glitches of prediction in between.

—Are business costs and benefits taken into account?

Answer) Yes the business costs and benefits have been taken into account. If the model is deployed for the business, it will help in saving time and effort of the resource managers. The time and effort can be used in some more productive work.

Thus a successful deployment of the model will help in increasing all over productivity of the business.

—For classification, how is a classification threshold chosen?

Answer) For classification, the threshold is chosen according to the target variable (ACTION). It is a binomial variable and can take up 2 values, 0 if resource is not allocated and 1 if resource is allocated.

—Are probability estimates used directly?

Answer) No probability estimates are not used directly. Since my target variable (ACTION) is a binomial variable, the output has always been in the form of 0 or 1. So there has never been a direct use of probability estimates.

—Is ranking more appropriate (e.g., for a fixed budget)?

Answer) No ranking is not more appropriate here. This is because employees have different attributes. They are not on the same platform that ranking can be done for them.

Some of them have different department, different rollup, family description etc. So ranking in such a situation would be useless.

—For regression, how will you evaluate the quality of numeric predictions? Why is this the right way in the context of the problem?

Answer) Regression in terms of clear prediction of numeric output has not been used. Linear regression has been used as a classifier which predicts binomial output as 0(resource not allocated) or 1(resource allocated). All together a clear prediction of different numeric value was never needed all though-out the business problem.

This was right in context of the business problem because the system can be used even if there are minor glitches in the predicted output. This is because even if the prediction is wrong, an employee can always reappeal to the resource manager for allocation of the resource. However the success of the system can also be measured from the fact that there is minimum intervention from the resource manager for the entire system output.

● Does the evaluation use holdout data?

Answer) No, holdout data has not been used. I had separate test and training data set. So my testing has always been very robust and efficient.

—Cross-validation is one technique.

Answer) Since I have separate test and training data set, so I did not use cross validation as a testing technique.

Also testing data clearly provides a set for efficient testing of system.

● Against what baselines will the results be compared?

Answer) The result can be compared along the baseline that the Area under the curve should be maximum possible. Also the confusion matrix should have maximum value along its diagonals.

The system can afford glitches in between and the real success baseline of the system will be when it requires minimum human intervention and all the resource allocation done by it are correct.

—Why do these make sense in the context of the actual problem to be solved?

Answer) These make sense in terms of the actual problem because higher the area under curve, better is the prediction and less is the human intervention in the system. Also better is the prediction from the system, it will lead to less wastage of time of resource managers of the company. They can then do work more productive and thus lead to more profits of the company.

—Is there a plan to evaluate the baseline methods objectively as well?

Answer) Yes the baseline methods can be evaluated objectively as well. The model can predict the output and we can randomly test the particular employee with the resource manager as to whether he should be allocated the resource or not. This will provide a brief idea about the efficiency of the system

● For clustering, how will the clustering be understood?

Answer) For clustering, the target variable has been removed. Then I have started with two resource managers and thus divided all the employees in two clusters. Employees who have almost the same features come under one cluster. Thus one resource manager can be allocated to each cluster and this can help in better management of work.

● Will deployment as planned actually (best) address the stated business problem?

Answer) Yes deployment of the model will help address the business problem. This is because we have a supervised business problem. The target variable output will exactly predict whether the resource will be allocated to the employee or not. This will exactly address our business problem. If the model is deployed in the company, it will reduce intervention of the resource manager and his/her time and efforts can be used in something more productive in the company.

● If the project expense has to be justified to stakeholders, what is the plan to measure the final (deployed) business impact?

Answer) If the project cost has to be justified to the stakeholders, we can explain it to them in a way that the expected value framework in the project will give a very high positive output. Also the project deployment will help save time and effort of the resource manager of the company. The skills and efficiency of the resource manager can be used somewhere else in a better productive way and thus the company profit can increase in the long term.

References:

1) https://www.kaggle.com/c/amazon-employee-access-challenge
2) Citation of UCI Data set:
   @misc{Montanez:2011 ,
   author = "Ken Montanez",
   year = "2011",
   title = "{UCI} Machine Learning Repository",
   url = " https://archive.ics.uci.edu/ml/datasets/Amazon+Access+Samples",
   institution = "Information Security, Amazon Corp" }