ANALYZING JOB MARKET WITH NAUKRI.COM

GUIDED BY,
DR. C C CHAN

TEAM MEMBERS,
BINI ELSA PAUL
GEETHIKA LAKSHMI YARRA
SRI LAKSHMI PRIYANKA SADINENI

CONTENTS

- Objective and Problem Statement
- Data Set Used
- Preprocessing the data
- Data Mining and Result
- Discussion / Future work
- Conclusion
- References
- Questions

OBJECTIVE AND PROBLEM STATEMENT

- To implement KDD process
- To analyze the job market
  - In the view of employee
    - To predict the salary
    - To predict the job location
  - In the view of job portal
    - To group similar samples so that the new input can be easily grouped into an existing group. The new input can be
      - Resume/ job application from employees
      - Job advertisement from employers
- Data Mining Techniques used
  - Classification :- Prediction
  - Clustering :- Grouping

DATA SET USED

- Jobs on Naukri.com from Kaggle.com
- 22,000 job listings on Naukri.com
- Data size : 50 MB
- This dataset has following fields:

company	education	experience	industry	job description	jobid	joblocation_address
job title	number of positions	pay rate	postdate	site_name	skills	uniq_id

DATA PRE - PROCESSING

MAIN STEPS INCLUDES

1. Feature Reduction
2. Data Cleaning
3. Data Transformation
4. Quantification of features

FEATURE REDUCTION

company	education	experience	industry	job description	jobid	joblocation_address
job title	number of positions	pay rate	postdate	site_name	skills	uniq_id

DATA CLEANING

- Missing Values
  - Extracted missing values from job description with combined computer and human inspection
  - Filled number of positions depending on the number of job location
  - Ignored the tuples
- Remove Duplicates
  - Duplicates job locations removed by creating a tree structure

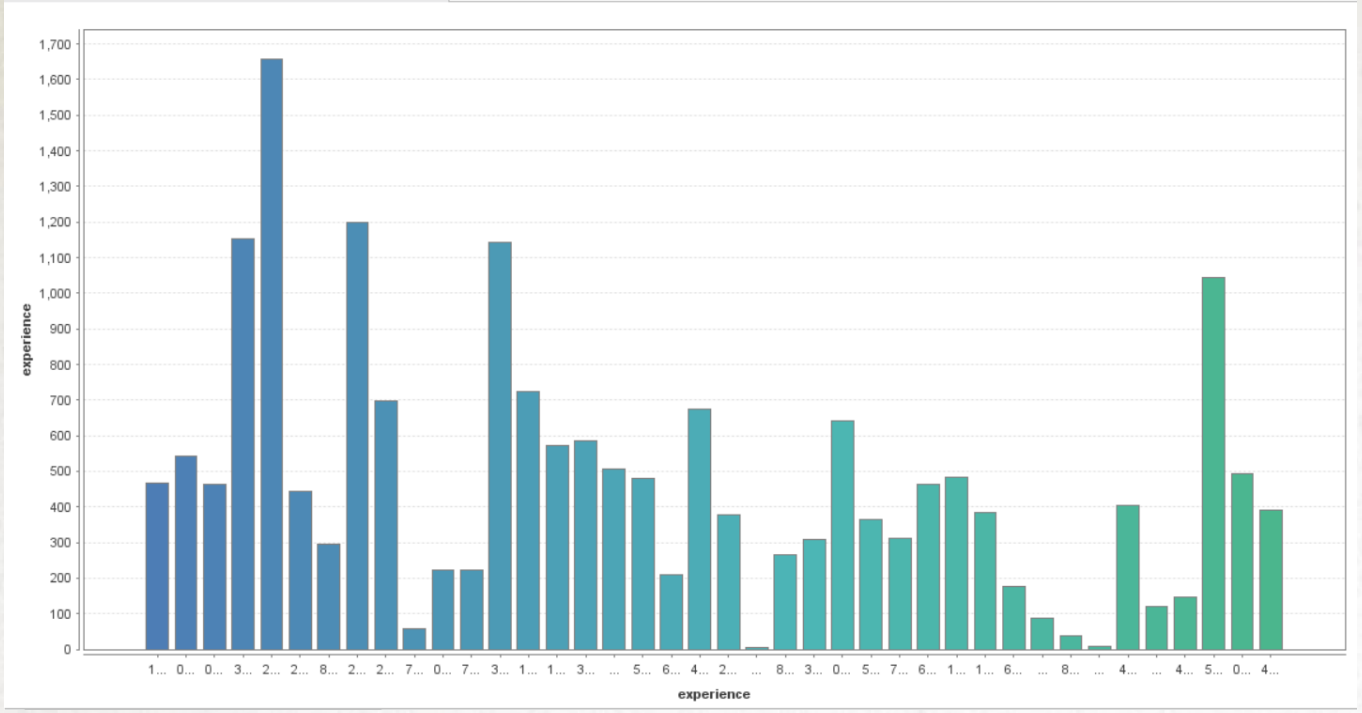
DATA TRANSFORMATION

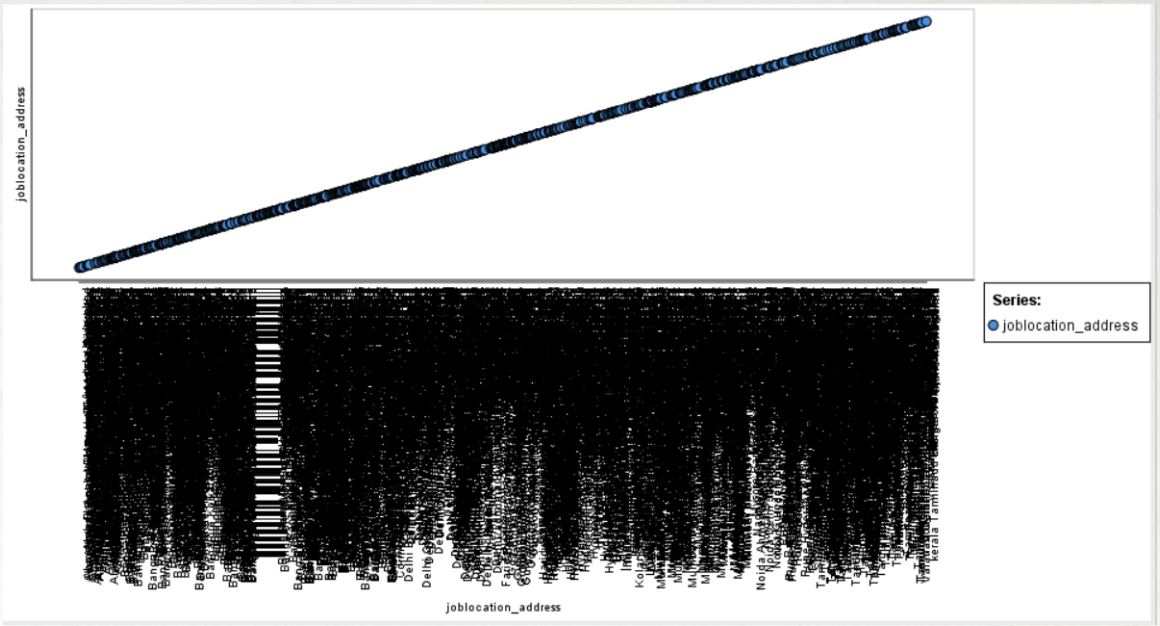
1. Tokenized Skills
    - Used iterative process
2. Used Python NLTK for all other text fields (especially job description)
3. Transformed pay rate values to only meaningful numbers with python

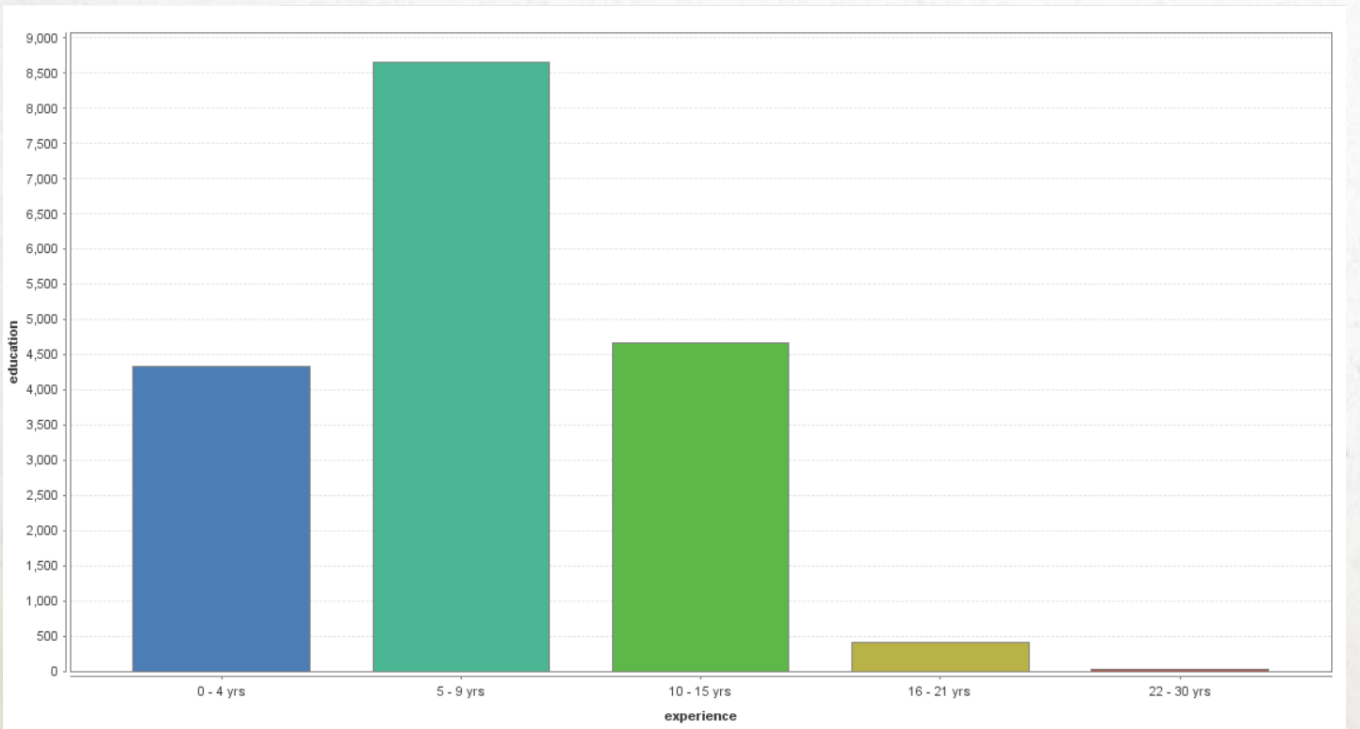
QUANTIFICATION OF FEATURES

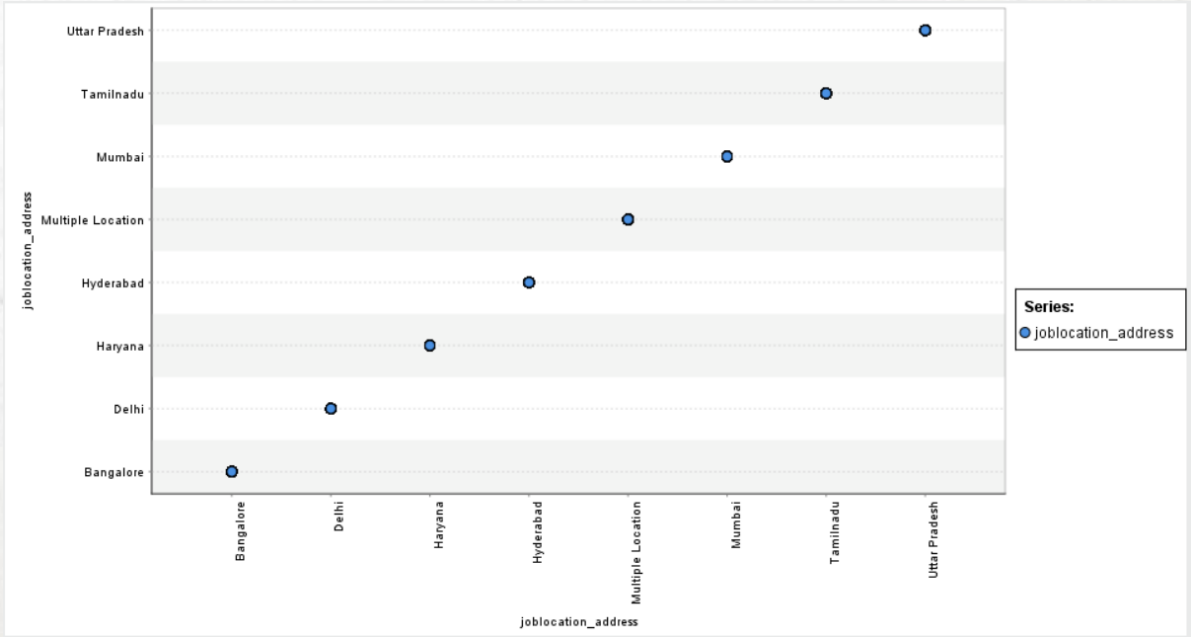
- The values in attributes grouped together (eg: pay rate, job location ...)
- Introduced a new attribute job_title_group by grouping similar titles (developer, lead ...)

Post Graduation Any Specialization 5 - 9 yrs integrated code Specialization Expert technologies BM Proficient existing manages including solutions innovative Candidate Doctorate Demonstrated developer Git web We Jobs knowledge application implementation debug Company Assist writing platform bases York samples communication Services suite transit first Salary customers good commerce Mobile Industry libraries iOS possible preferred functional framework know performance using ticket years scrolling secure maintaining compiler timely like settings skills versions fixes Functional Dispatch teams large team small Role security refer Ability Graduation work Any UIKit Provide participate Send group translating Application titles static UI Grand Platform blocks investment best techniques merchants ticketing consumers space open leader continually looking Desired Job optimization turnaround PG goes new Bangalore across Must C working Central creating business Programming interfaces maintains communicate little delivering implementing Shipped Contact Knowledge interface modeling understanding great talented launch Photo Xcode engineers allows confident experience times Cocoa etc range Computer speedy agency Download implement Inc support Keyskills Developer Category major solving features feel Qualifications app Please primary highly services Maintenance tools Store engineer use management networking For service APIs leading top Area apps system payment Required memory Bytemark Post cycle field testing infrastructure offer Details Science Object Experience delivery Not part understand related Allocations Coordinate worked bug Software professionals smartphone PPT developing past clients smooth frameworks future View problem providing App launched extracting debugging dedicated transition Design plugins delivers spectrum deep industry hardware Other INR New Objective You create tourism Duties Profile SQLite requirements Description provide improves responsible viewDidUnload also responsibilities complex role organizations January test party developers events development used multiple description comprehensive ponies IT release applications develops levels user dealloc improving The data UG Launched built building third mobile Instruments documentation Leaks customized As time Waterway software SDK Bangalore iOS Developer 1 760000 - 1000000 INR Application Programming IT Software 0104a9295f11b94eec393a13a7f6a792 Developer


experience
1,700
1,600
1,500
1,400
1,300
1,200
1,100
1,000
900
800
700
600
500
400
300
200
100
0
experience


joblocation_address
joblocation_address
Series:
joblocation_address


education
9,000
8,500
8,000
7,500
7,000
6,500
6,000
5,500
5,000
4,500
4,000
3,500
3,000
2,500
2,000
1,500
1,000
500
0
0 - 4 yrs
5 - 9 yrs
10 - 15 yrs
16 - 21 yrs
22 - 30 yrs
experience


joblocation_address
Uttar Pradesh
Tamilnadu
Mumbai
Multiple Location
Hyderabad
Haryana
Delhi
Bangalore
Bangalore
Delhi
Haryana
Hyderabad
Multiple Location
Mumbai
Tamilnadu
Uttar Pradesh
joblocation_address
Series:
joblocation_address

DATA MINING : CLUSTERING

```
Relation:     complete1
Instances:    18165
Attributes:   8
              education
              experience
              jobdescription
              joblocation_address
              numberofpositions
              payrate
              skills
Ignored:
              title_group
Test mode:    Classes to clusters evaluation on training data
```

```
Time taken to build model (full training data) : 78.09 seconds

=== Model and evaluation on training set ===

Clustered Instances

0         302 (  2%)
1        1709 (  9%)
2        3459 ( 19%)
3        1527 (  8%)
4         917 (  5%)
5       10251 ( 56%)
```

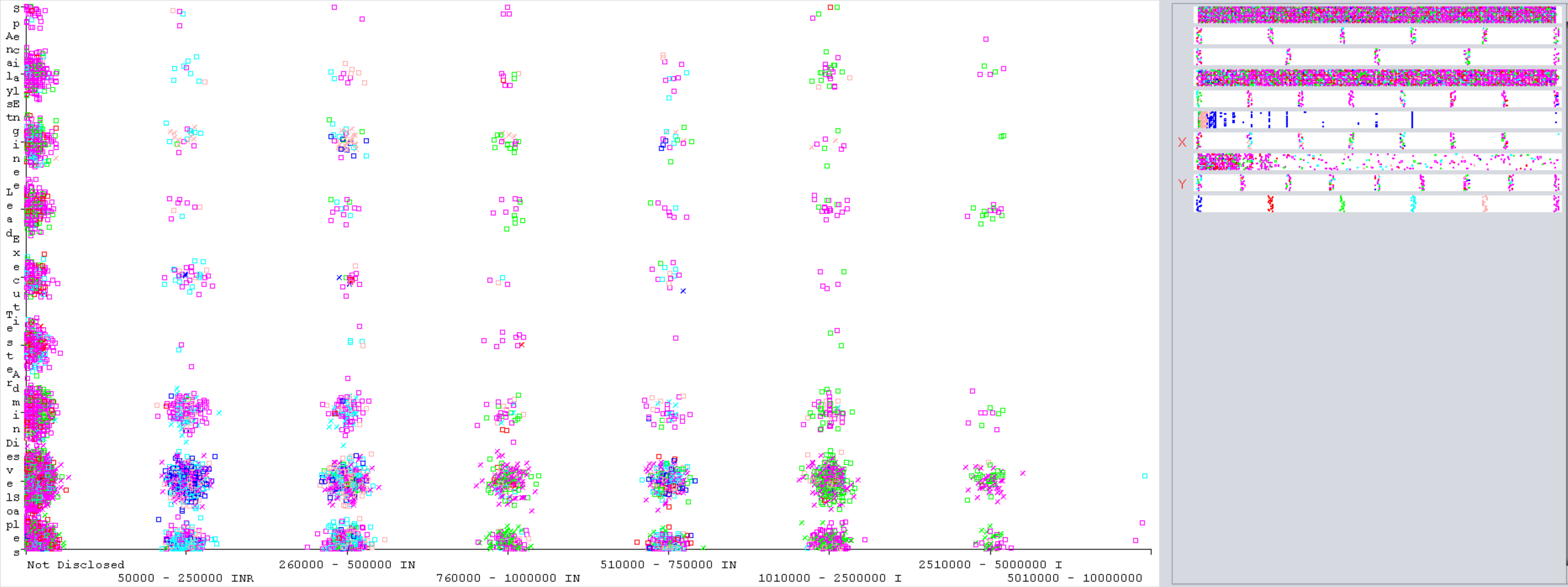
```
Number of clusters selected by cross validation: 6
Number of iterations performed: 8
```

```
Class attribute: title_group
Classes to Clusters:

    0    1    2    3    4    5  <-- assigned to cluster
   35  523 1105  455  243 2535 | Sales
  245  877 1783  740  508 5513 | Developer
    6  162  283  172   80 1166 | Administaor
    0   48   22   43    4  218 | Tester
    5   33   16   37   11  141 | Executive
    0   25   86   10    3  213 | Lead
   11   23  102   45   55  201 | Engineer
    0   16   59   23   12  231 | Analyst
    0    2    3    2    1   33 | Specialist

Cluster 0 <-- Executive
Cluster 1 <-- Tester
Cluster 2 <-- Sales
Cluster 3 <-- Administaor
Cluster 4 <-- Engineer
Cluster 5 <-- Developer
```

No.	1: education	2: experience	3: jobdescription	4: joblocation_address	5: jobtitle	6: numberofpositions	7: payrate	8: skills	9: **cluster**
	Nominal	Nominal	Nominal	Nominal	Nominal	Numeric	Nominal	Nominal	Nominal
1	Doctorate	0 - 4 yrs	concept Speci...	Multiple Location	Advert...	4.0	Not Di...	Mark...	cluster2
2	P G Fina...	0 - 4 yrs	sector BD sel...	Delhi	Requi...	20.0	50000 ...	Mark...	cluster3
3	Post Gra...	0 - 4 yrs	Web Specializ...	Bangalore	Digital...	2.0	50000 ...	Sales	cluster1
4	Doctorate	5 - 9 yrs	Specialization...	Bangalore	IOS A...	1.0	Not Di...	Appli...	cluster1
5	Post Gra...	5 - 9 yrs	Category Spe...	Multiple Location	Onlin...	10.0	26000...	Appli...	cluster3
6	Post Gra...	0 - 4 yrs	Requirement...	Multiple Location	PHP ...	2.0	Not Di...	Appli...	cluster3
7	Doctorate	5 - 9 yrs	Specialization...	Hyderabad	iPhon...	1.0	Not Di...	Appli...	cluster1
8	Under Gr...	10 - 15 yrs	Web Specializ...	Multiple Location	AEM A...	2.0	Not Di...	Appli...	cluster2
9	Under Gr...	5 - 9 yrs	Category dev...	Multiple Location	Sales ...	1.0	50000 ...	Sales	cluster3
10	Doctorate	0 - 4 yrs	atmosphere ...	Bangalore	Matron	2.0	Not Di...	HR	cluster1
11	Post Gra...	10 - 15 yrs	BA Candidate...	Multiple Location	Busin...	4.0	76000...	Appli...	cluster3
12	Doctorate	5 - 9 yrs	Specialization...	Bangalore	Core ...	1.0	Not Di...	Appli...	cluster1
13	Post Gra...	0 - 4 yrs	SIS Specializ...	Delhi	Logist...	1.0	50000 ...	Chai...	cluster1
14	Post Gra...	5 - 9 yrs	operations Sp...	Multiple Location	JAVA	2.0	Not Di...	Appli...	cluster3
15	Post Gra...	5 - 9 yrs	alm High Can...	Multiple Location	Assoc...	2.0	Not Di...	Testi...	cluster3
16	Post Gra...	0 - 4 yrs	Specialization...	Bangalore	MRD ...	3.0	Not Di...	Medi...	cluster1
17	Post Gra...	5 - 9 yrs	Reconciliatio...	Multiple Location	Sr Acc...	3.0	51000...	Acco...	cluster3
18	Diploma	0 - 4 yrs	control Categ...	Tamilnadu	Assist...	1.0	50000 ...	Sales	cluster1
19	Doctorate	10 - 15 yrs	Writing Categ...	Mumbai	Profe...	2.0	Not Di...	Appli...	cluster2
20	Doctorate	5 - 9 yrs	Category train...	Bangalore	SAS T...	1.0	Not Di...	Servi...	cluster1
21	Doctorate	5 - 9 yrs	Specialization...	Haryana	Secret...	1.0	Not Di...	Assi...	cluster1
22	Post Gra...	0 - 4 yrs	protocol Spec...	Uttar Pradesh	Assoc...	1.0	Not Di...	Appli...	cluster1
23	Post Gra...	5 - 9 yrs	code Speciali...	Bangalore	Softw...	1.0	Not Di...	Appli...	cluster1
24	Post Gra...	5 - 9 yrs	Specialization...	Haryana	Accou...	1.0	Not Di...	Sales	cluster1
25	Post Gra...	5 - 9 yrs	Specialization...	Multiple Location	Senio...	2.0	Not Di...	Acco...	cluster3
26	Doctorate	0 - 4 yrs	Specialization...	Multiple Location	Applic...	4.0	Not Di...	Mark...	cluster2
27	Post Gra...	5 - 9 yrs	paperless co...	Bangalore	Androi...	1.0	Not Di...	Mobi...	cluster1
28	MBBS Me...	0 - 4 yrs	Category Spe...	Delhi	Assoc...	1.0	10100...	Medi...	cluster1
29	Doctorate	5 - 9 yrs	warranty Spec...	Mumbai	Mana...	1.0	Not Di...	Prod...	cluster1

PAY RATE

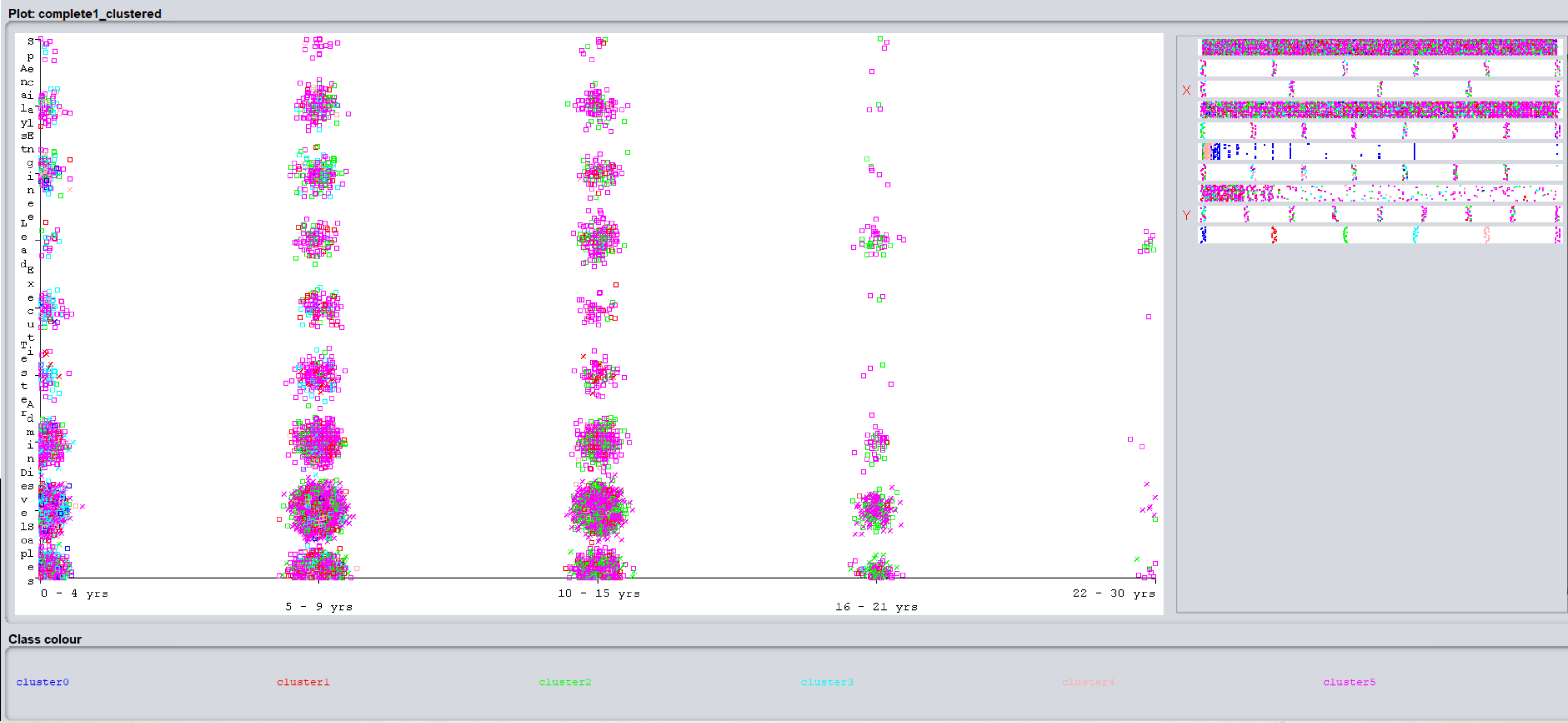
Plot: complete1_clustered



Class colour

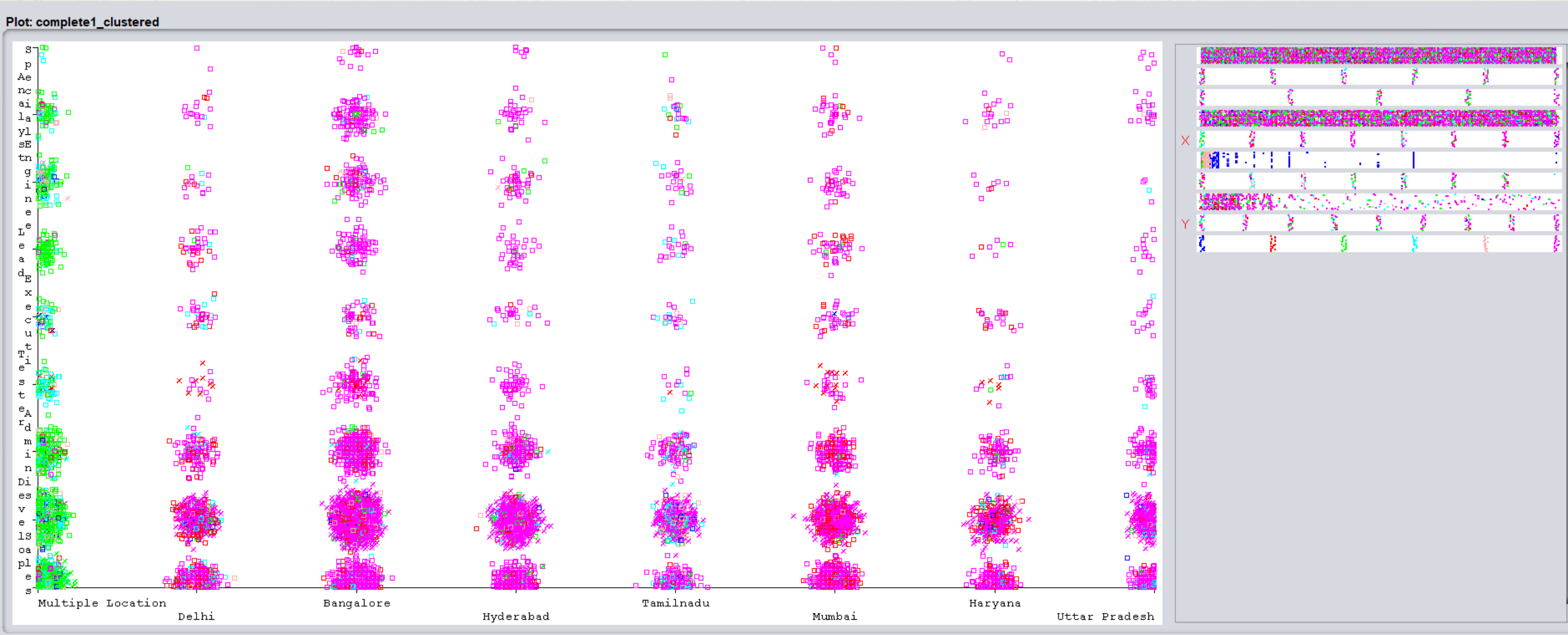
cluster0 cluster1 cluster2 cluster3 cluster4 cluster5

EXPERIENCE

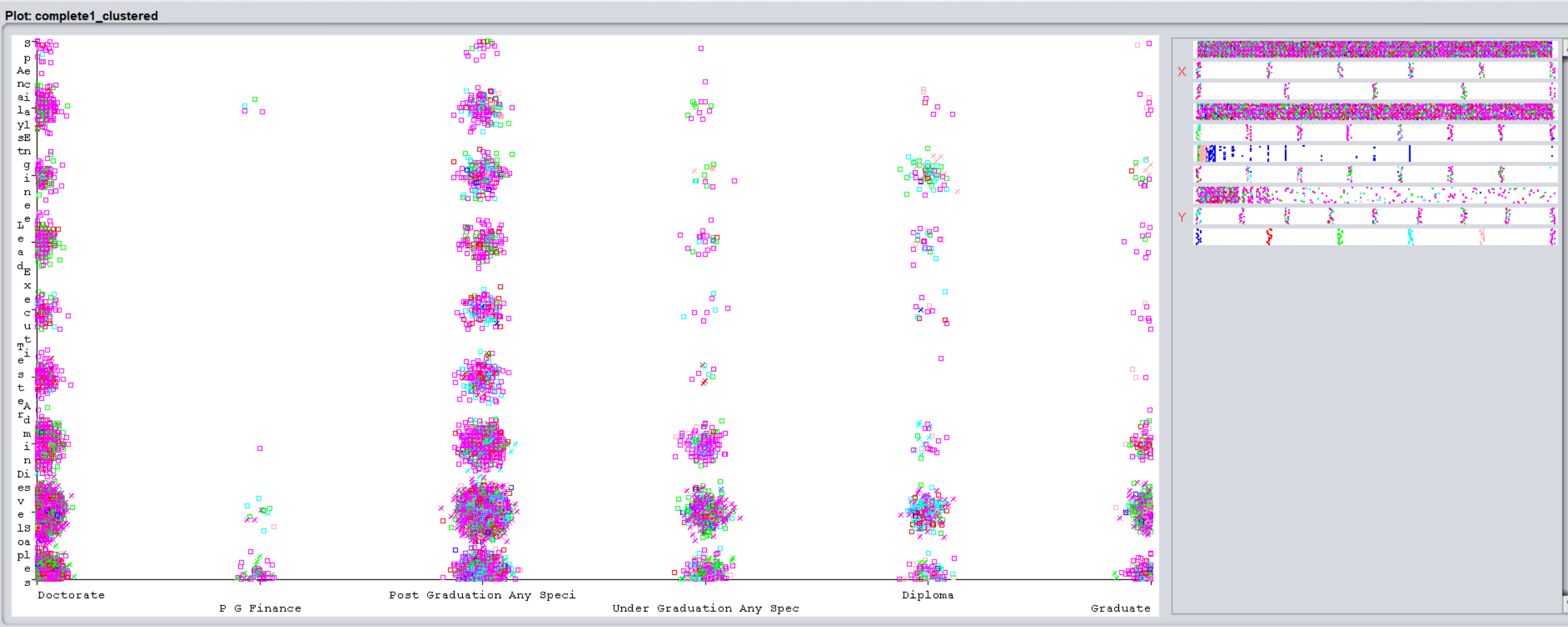


Plot: complete1_clustered
0 - 4 yrs
5 - 9 yrs
10 - 15 yrs
16 - 21 yrs
22 - 30 yrs
X
Y
Class colour
cluster0
cluster1
cluster2
cluster3
cluster4
cluster5

JOB LOCATION



EDUCATION


Plot: complete1_clustered
Doctorate
P G Finance
Post Graduation Any Speci
Under Graduation Any Spec
Diploma
Graduate
X
Y
Class colour
cluster0
cluster1
cluster2
cluster3
cluster4
cluster5

DATA MINING : CLASSIFICATION

1. Predict salary from
    1. Education
    2. Experience
    3. Job title
    4. Job location
    5. skills
2. Predict Job Location Address from
    1. Education
    2. Experience
    3. Job title
    4. Pay rate

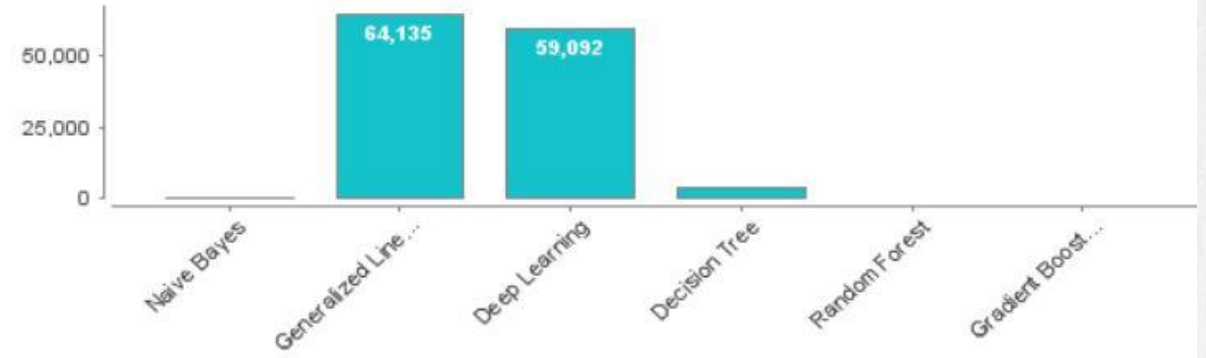
    Classification done using
    - Naïve Bayes Model
    - Decision Tree Model

PREDICT SALARY – STAGE 1

Overview

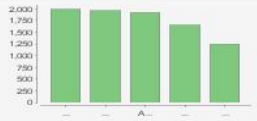
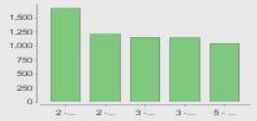
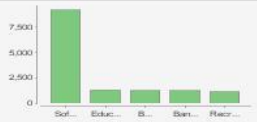
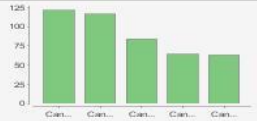
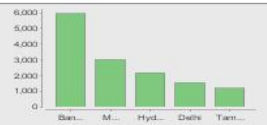
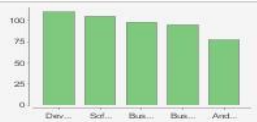
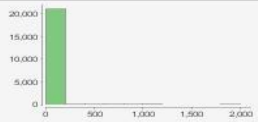
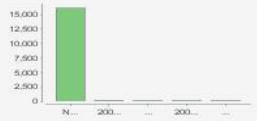
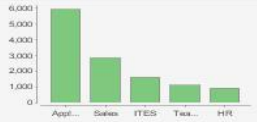
Accuracy

Runtime (ms)

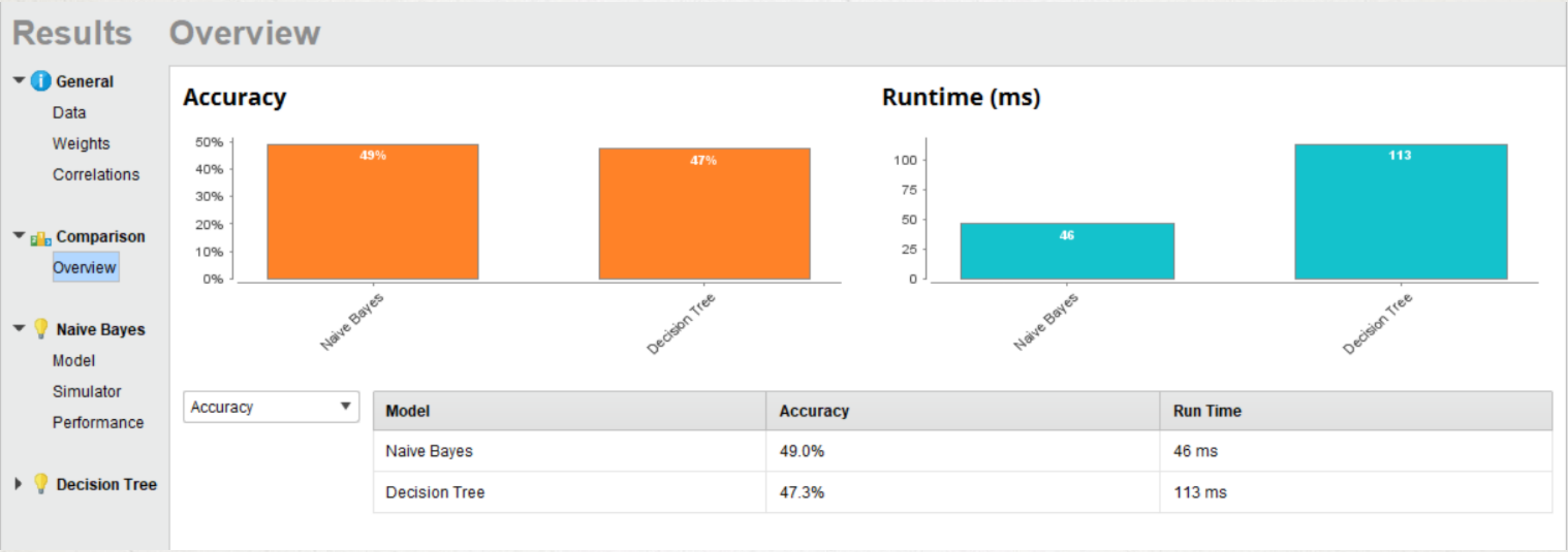


Accuracy

Model	Accuracy	Run Time
Naive Bayes	15.5%	368 ms
Generalized Linear Model	3.4%	1 min 4 s
Deep Learning	9.8%	59 s
Decision Tree	4.0%	4 s

education	Polynominal	0	 Open chart	Least UG Texti [...] uired (1)	Most UG PG Do [...] ed (1989)		Values UG PG Do [...] Required (1989), UG Any G [...] Required (1959), Any Gruaduate (1923), UG PG Po [...] Required (1652), ...[2049 more] Details...
experience	Polynominal	0	 Open chart	Least 7 - 16 yrs (1)	Most 2 - 7 yrs (1658)		Values 2 - 7 yrs (1658), 2 - 5 yrs (1199), 3 - 8 yrs (1151), 3 - 5 yrs (1142), ...[143 more] Details...
industry	Polynominal	0	 Open chart	Least Leather (1)	Most Software Services (9123)		Values Software Services (9123), Educatio [...] Training (1291), BPO Call Centre ITES (1234), Banking [...] s Broking (1223), ...[58 more] Details...
jobdescription	Polynominal	0	 Open chart	Least ynrgobra [...] on If (1)	Most Candidat [...] nal (121)		Values Candidat [...] fessional (121), Candidat [...] fessional (116), Candidat [...] mic Apply (84), Candidat [...] unctional (64), ...[20174 more] Details...
joblocation_address	Polynominal	0	 Open chart	Least thane maharashtra (1)	Most Bangalore (5946)		Values Bangalore (5946), Mumbai (3005), Hyderabad (2170), Delhi (1538), ...[1732 more] Details...
jobtitle	Polynominal	0	 Open chart	Least year Contract yrs (1)	Most Developer (110)		Values Developer (110), Software Engineer (105), Business [...] t Manager (98), Business [...] Executive (95), ...[15920 more] Details...
numberofpositions	Integer	0	 Open chart	Min 1	Max 2000	Average 10.685	Deviation 86.787
payrate	Polynominal	0	 Open chart	Least 950000 - 1900000 (1)	Most Not Disclosed (16107)		Values Not Disclosed (16107), 200000 - 300000 (172), 4 3740067000 10500 (121), 200000 - 400000 (117), ...[511 more] Details...
skills	Polynominal	0	 Open chart	Least support [...] kills (1)	Most Applicat [...] re (5900)		Values Applicat [...] Software (5900), Sales (2829), ITES (1605), Teaching (1085), ...[191 more] Details...

PREDICT SALARY – STAGE 2 (AFTER QUANTIFICATION)



Rapid Miner

Results Naive Bayes - Performance

- General
  - Data
  - Weights
  - Correlations
- Naive Bayes
  - Model
  - Simulator
  - Performance

Criterion
- accuracy
- classification error

Table View | Plot View

accuracy: 48.99%

	true 50000 - ...	true 260000 ...	true 760000 -...	true 510000 ...	true 1010000...	true 251000...	true 5010000...	class precisi...
pred. 50000 ...	167	83	8	20	3	1	0	59.22%
pred. 26000...	46	96	18	50	18	1	0	41.92%
pred. 76000...	0	6	10	5	5	1	0	37.04%
pred. 51000...	9	10	9	7	7	0	0	16.67%
pred. 10100...	6	15	36	17	101	18	1	52.06%
pred. 25100...	0	1	1	0	7	6	0	40.00%
pred. 50100...	0	0	0	0	0	1	0	0.00%
class recall	73.25%	45.50%	12.20%	7.07%	71.63%	21.43%	0.00%	

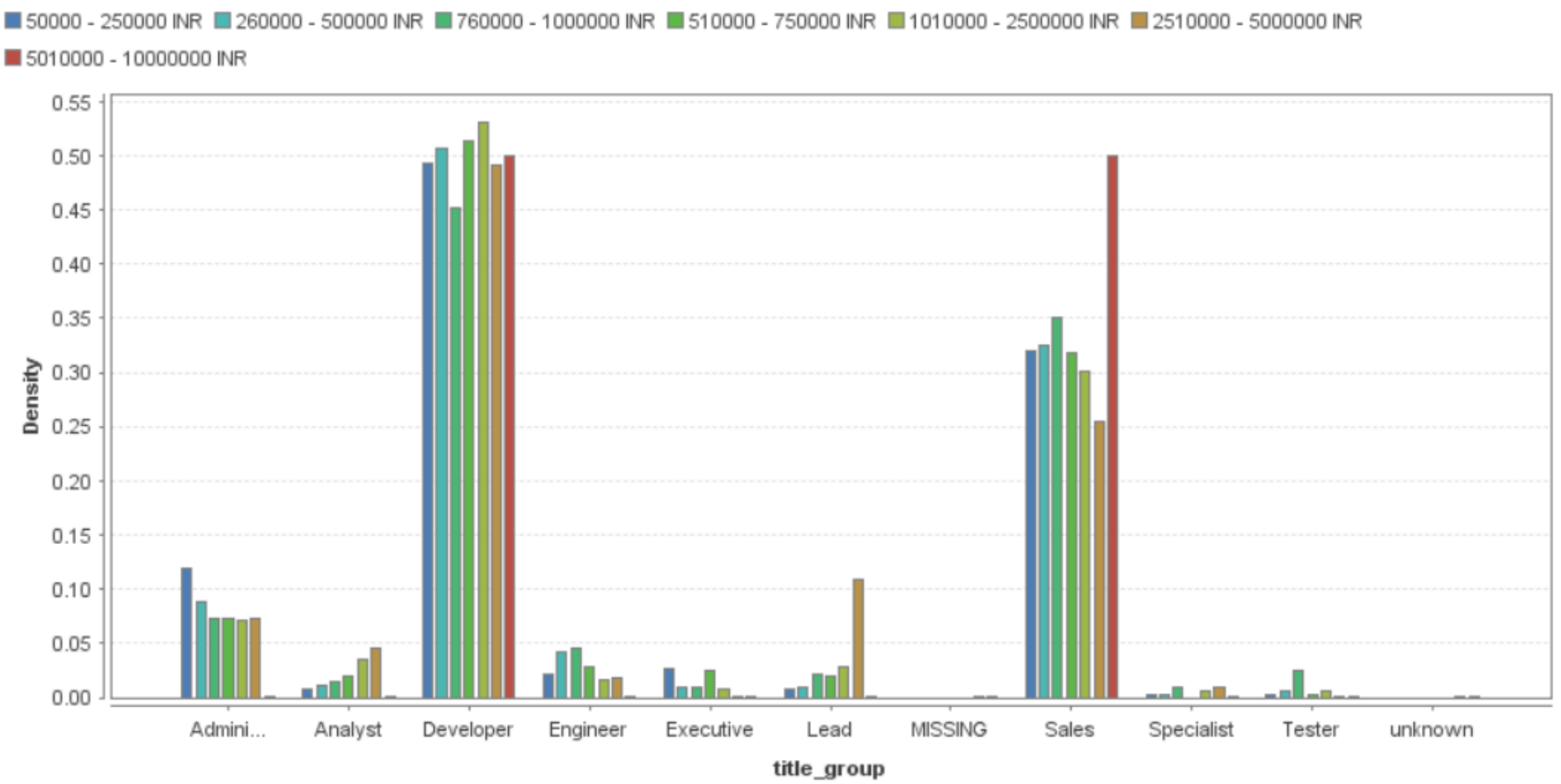
Results Naive Bayes - Model

- General
  - Data
  - Weights
  - Correlations
- Naive Bayes
  - Model
  - Simulator
  - Performance

Attribute:

title_group

Rotate Labels



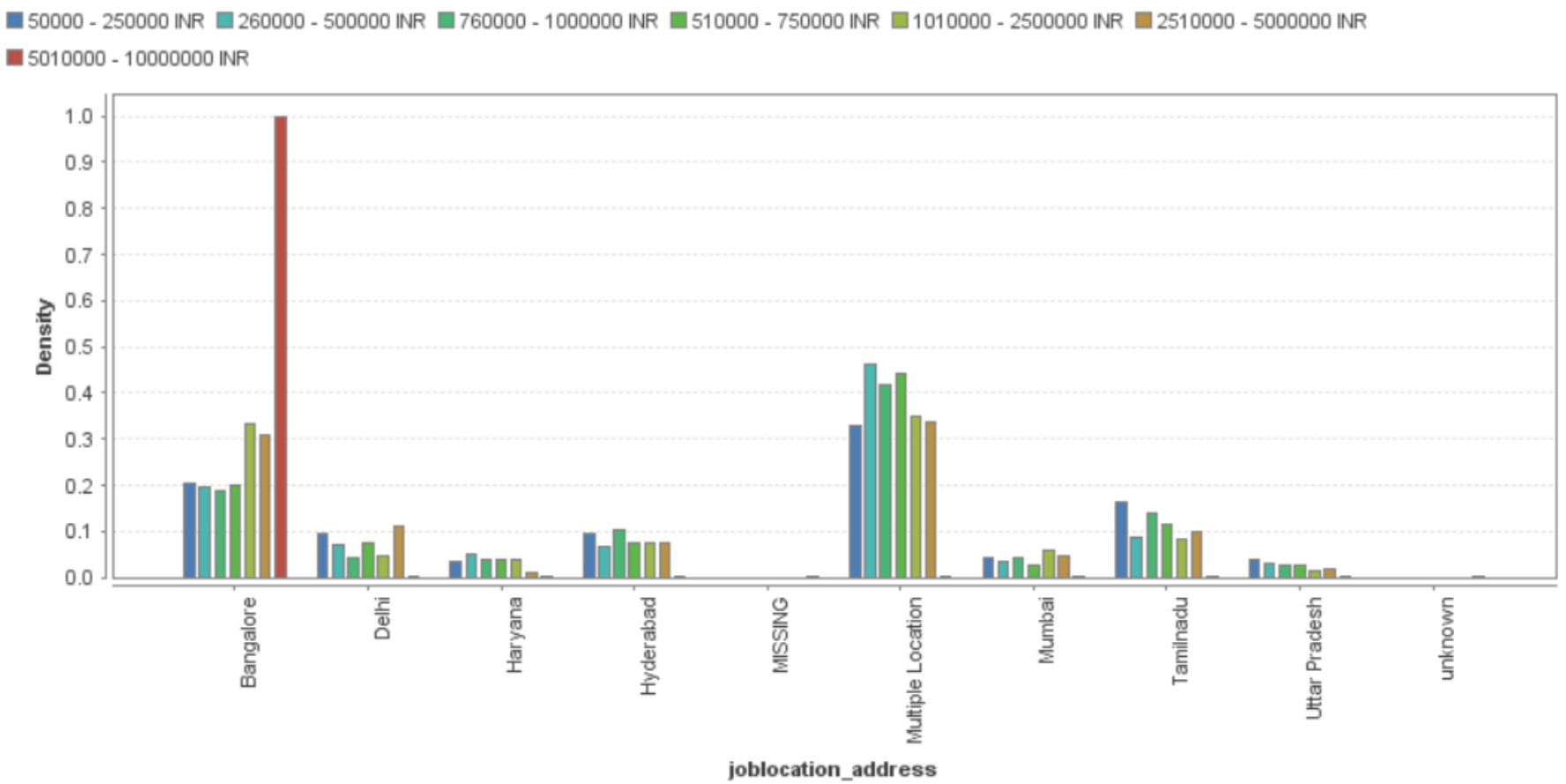
Results Naive Bayes - Model

- General
  - Data
  - Weights
  - Correlations
- Naive Bayes
  - Model
  - Simulator
  - Performance

Attribute:

joblocation_address

☑ Rotate Labels



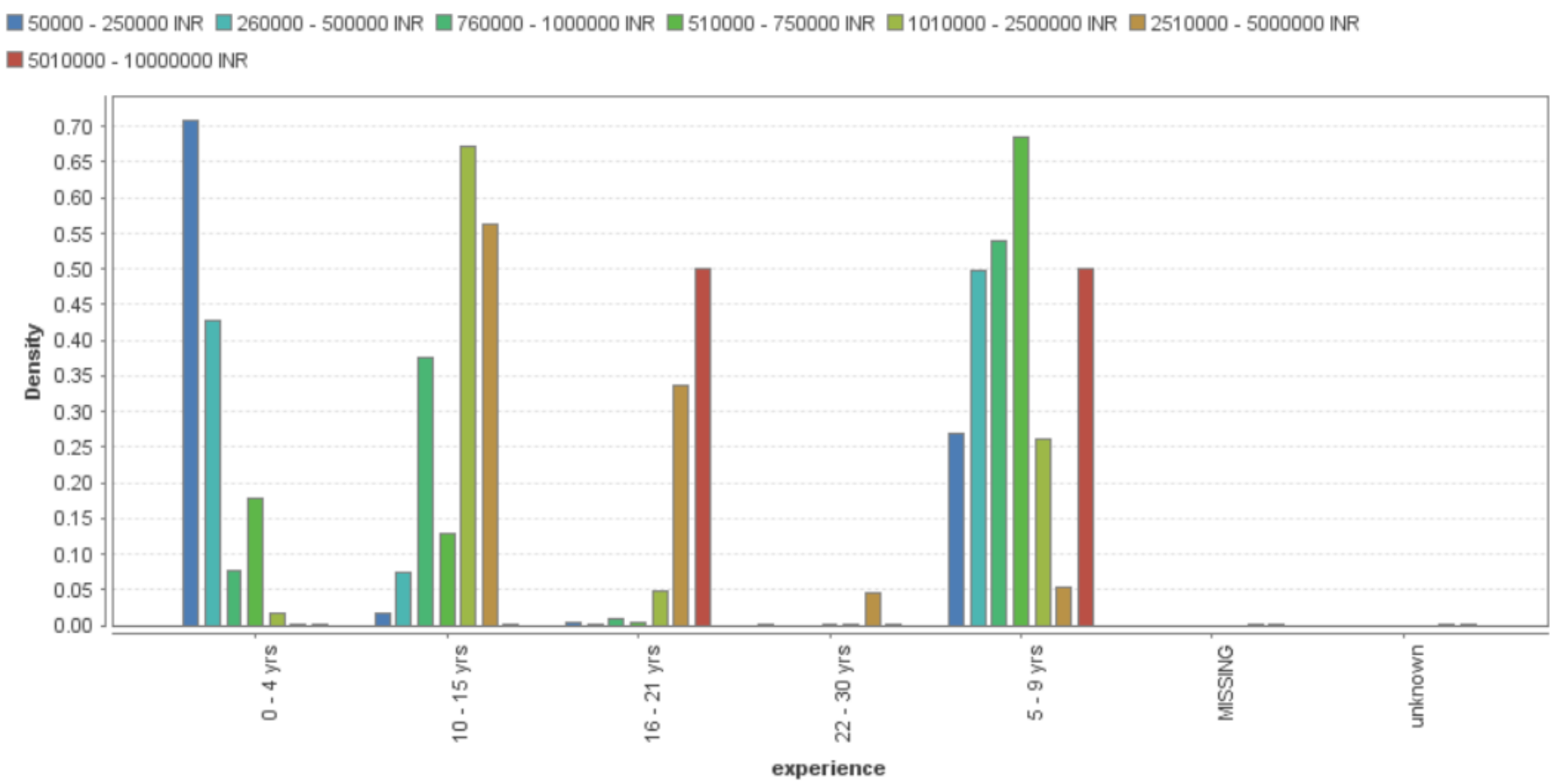
Results Naive Bayes - Model

- General
  - Data
  - Weights
  - Correlations
- Naive Bayes
  - Model
  - Simulator
  - Performance

Attribute:

experience

Rotate Labels



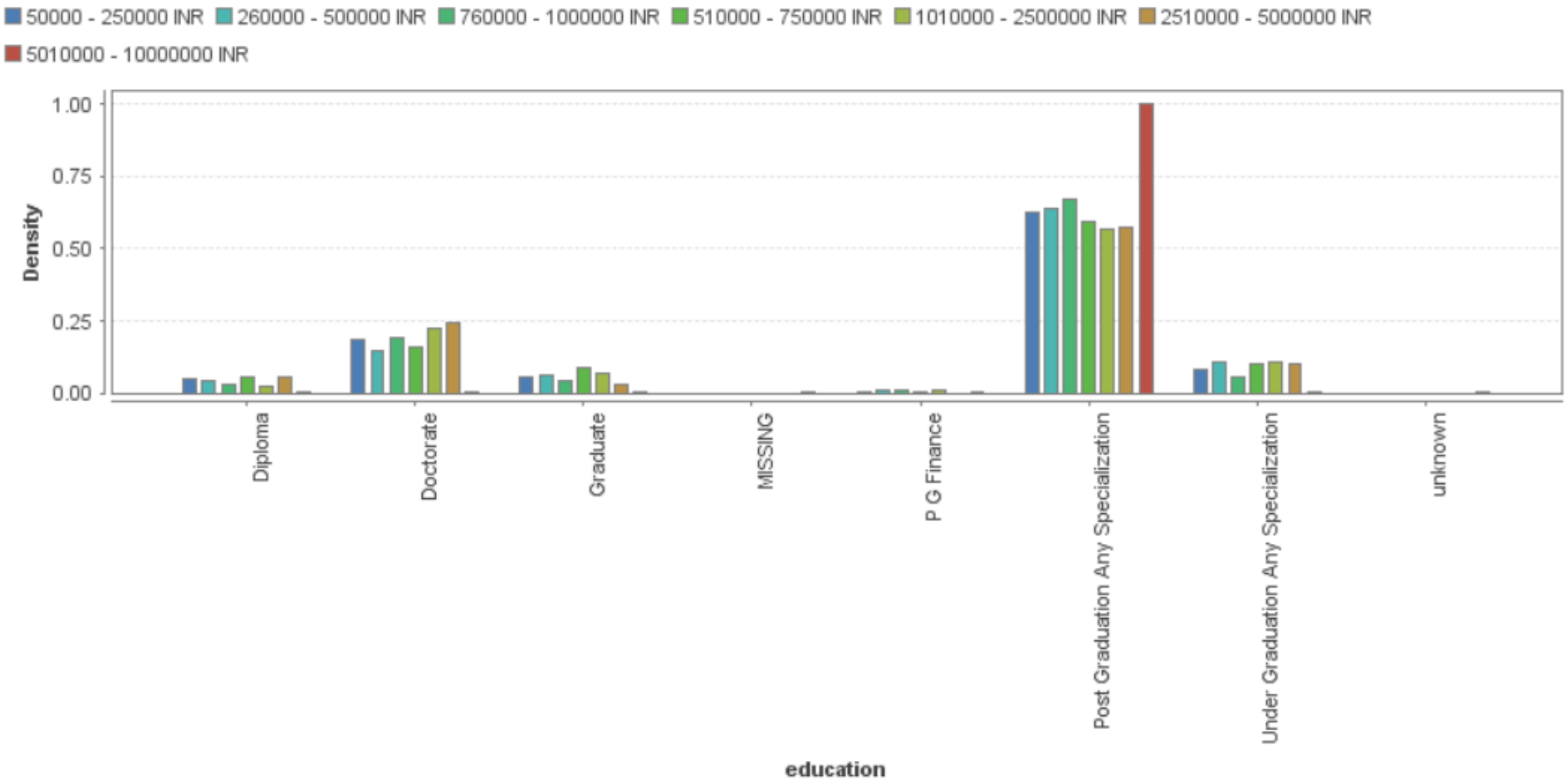
Results Naive Bayes - Model

- General
  - Data
  - Weights
  - Correlations
- Naive Bayes
  - Model
  - Simulator
  - Performance

Attribute:

education

Rotate Labels

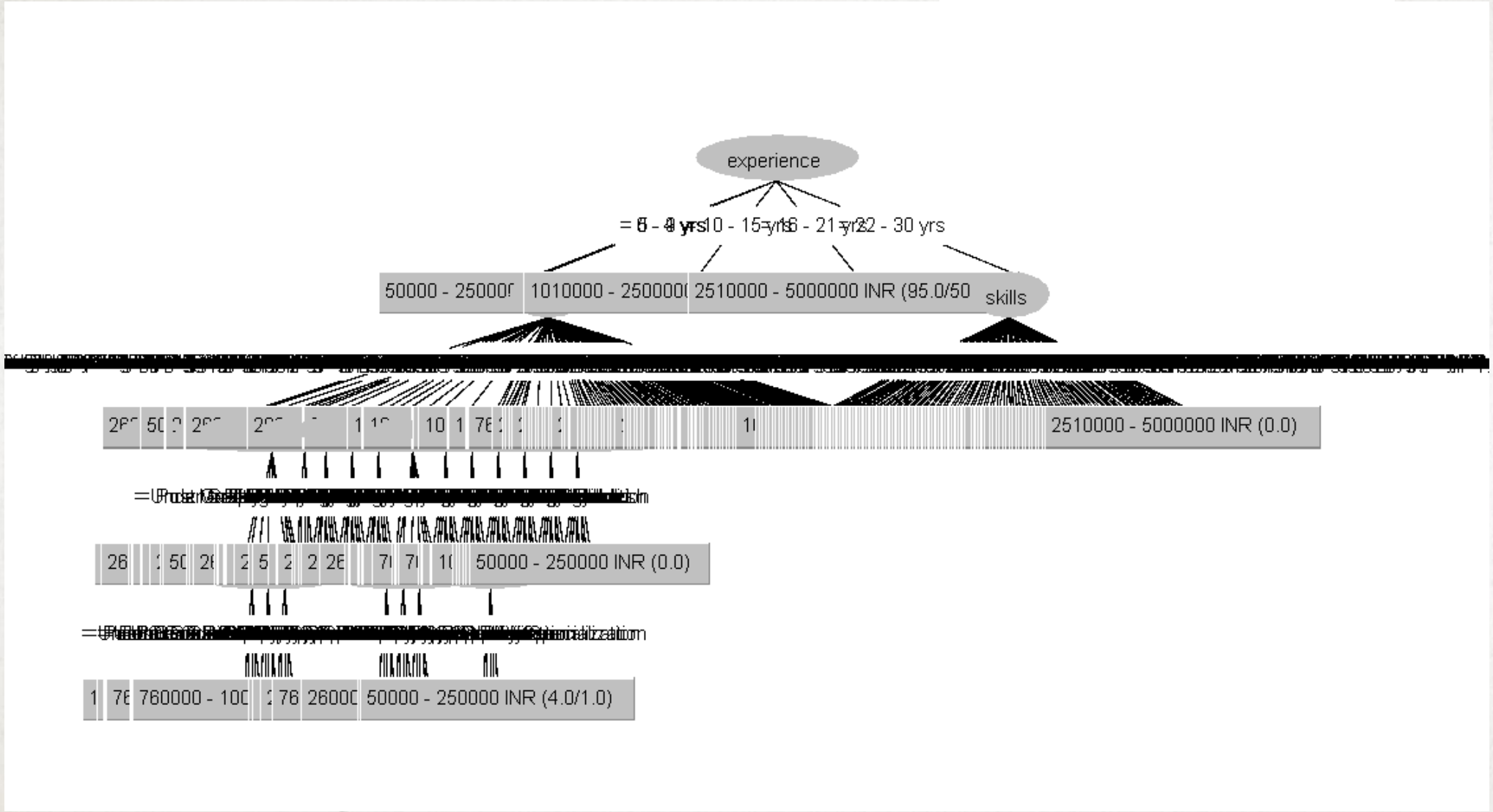


WEKA- 10-FOLD CROSS-VALIDATION

```
Time taken to build model: 0.02 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1912               47.992  %
Incorrectly Classified Instances      2072               52.008  %
Total Number of Instances             3984

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall   F-Measure   ROC Area  Class
                 0.763     0.258     0.546       0.763    0.637       0.795     50000 - 250000 INR
                 0.359     0.213     0.381       0.359    0.370       0.638     260000 - 500000 INR
                 0.077     0.025     0.267       0.077    0.120       0.707     760000 - 1000000 INR
                 0.084     0.027     0.311       0.084    0.132       0.639     510000 - 750000 INR
                 0.743     0.147     0.522       0.743    0.613       0.857     1010000 - 2500000 INR
                 0.355     0.015     0.458       0.355    0.400       0.874     2510000 - 5000000 INR
                 0.000     0.000     ?           0.000    ?           0.381     5010000 - 10000000 INR
Weighted Avg.    0.480     0.164     ?           0.480    ?           0.738

=== Confusion Matrix ===

   a   b   c   d   e   f   g   <-- classified as
 880 224   9   7  29   5   0 |   a = 50000 - 250000 INR
 515 383  30  35 102   3   0 |   b = 260000 - 500000 INR
  50 114  32  31 183   4   0 |   c = 760000 - 1000000 INR
 138 200  33  42  83   3   0 |   d = 510000 - 750000 INR
  26  81  14  19 526  42   0 |   e = 1010000 - 2500000 INR
   2   2   2   1  82  49   0 |   f = 2510000 - 5000000 INR
   0   0   0   0   2   1   0 |   g = 5010000 - 10000000 INR
```

DECISION TREE

```
Number of Leaves   :          324

Size of the tree :          346
```

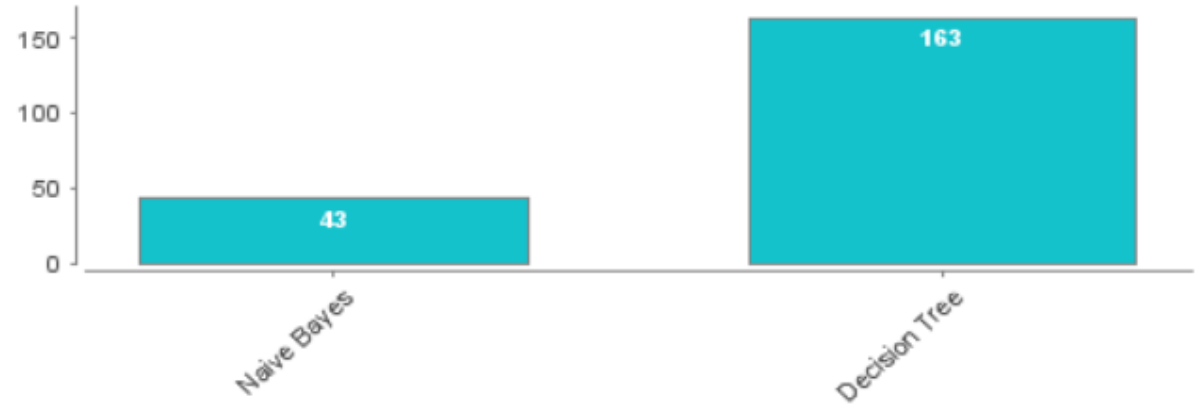


PREDICT JOB LOCATION ADDRESS

Overview

Accuracy

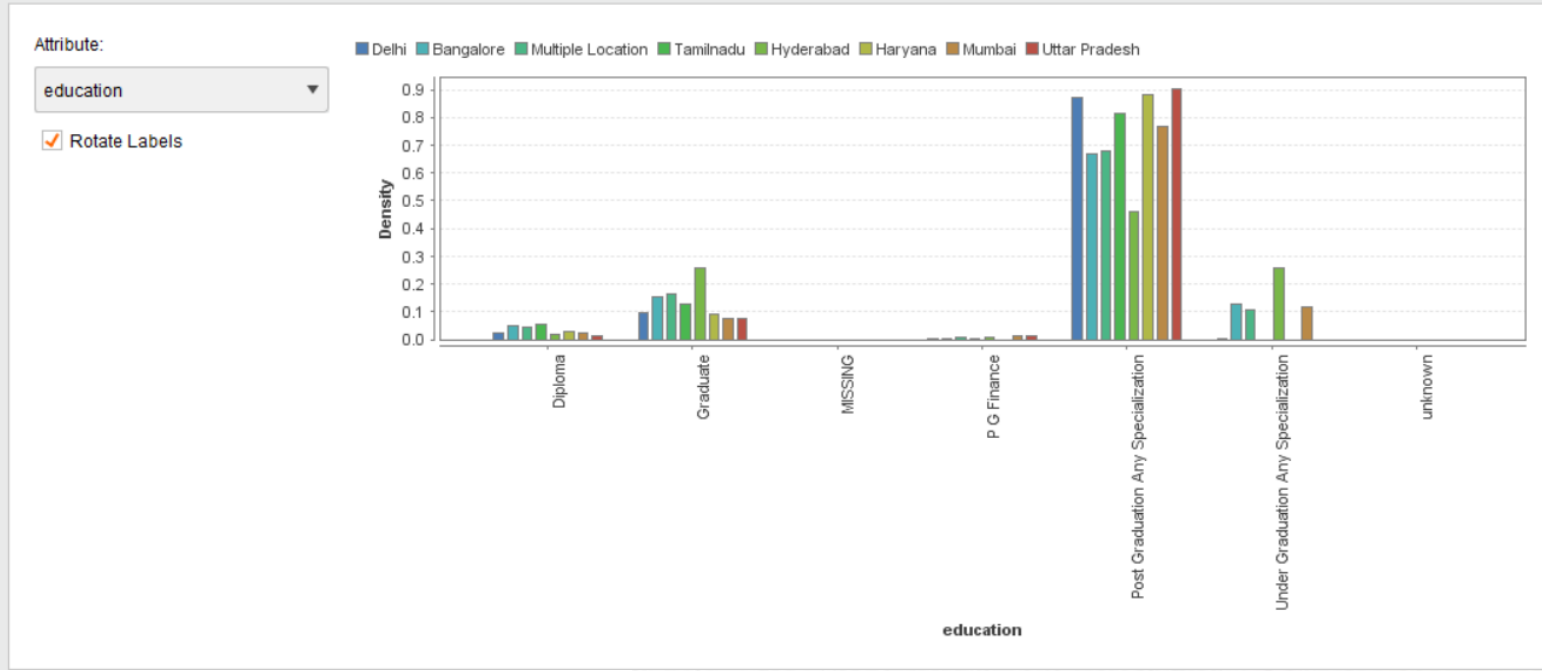
Runtime (ms)



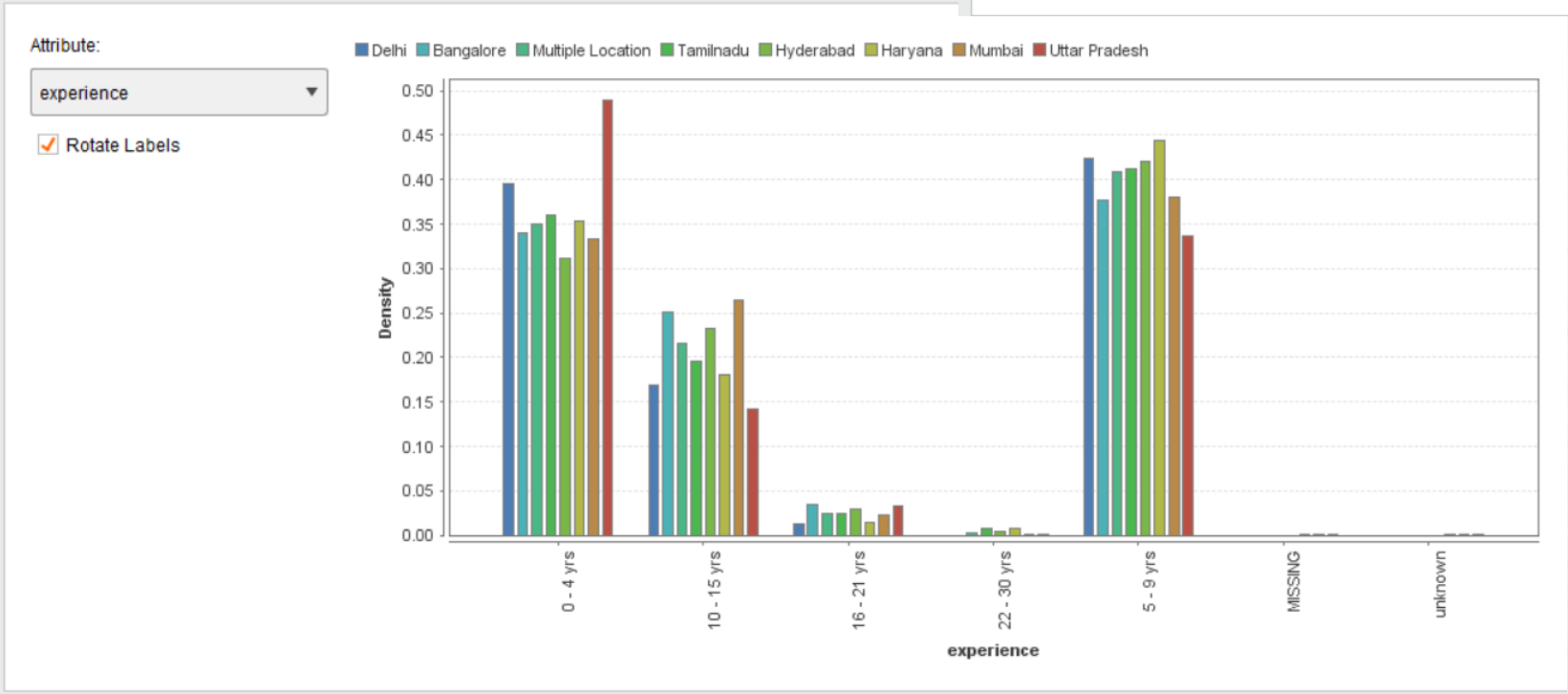
Accuracy

Model	Accuracy	Run Time
Naive Bayes	40.3%	43 ms
Decision Tree	38.7%	163 ms

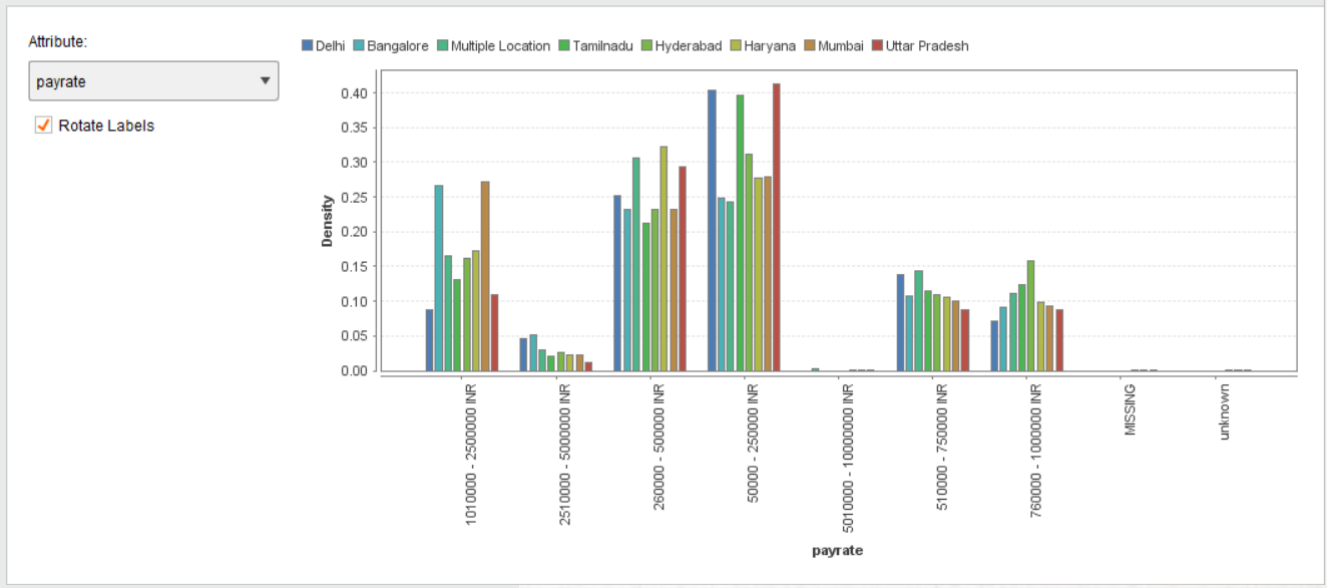
Naive Bayes - Model



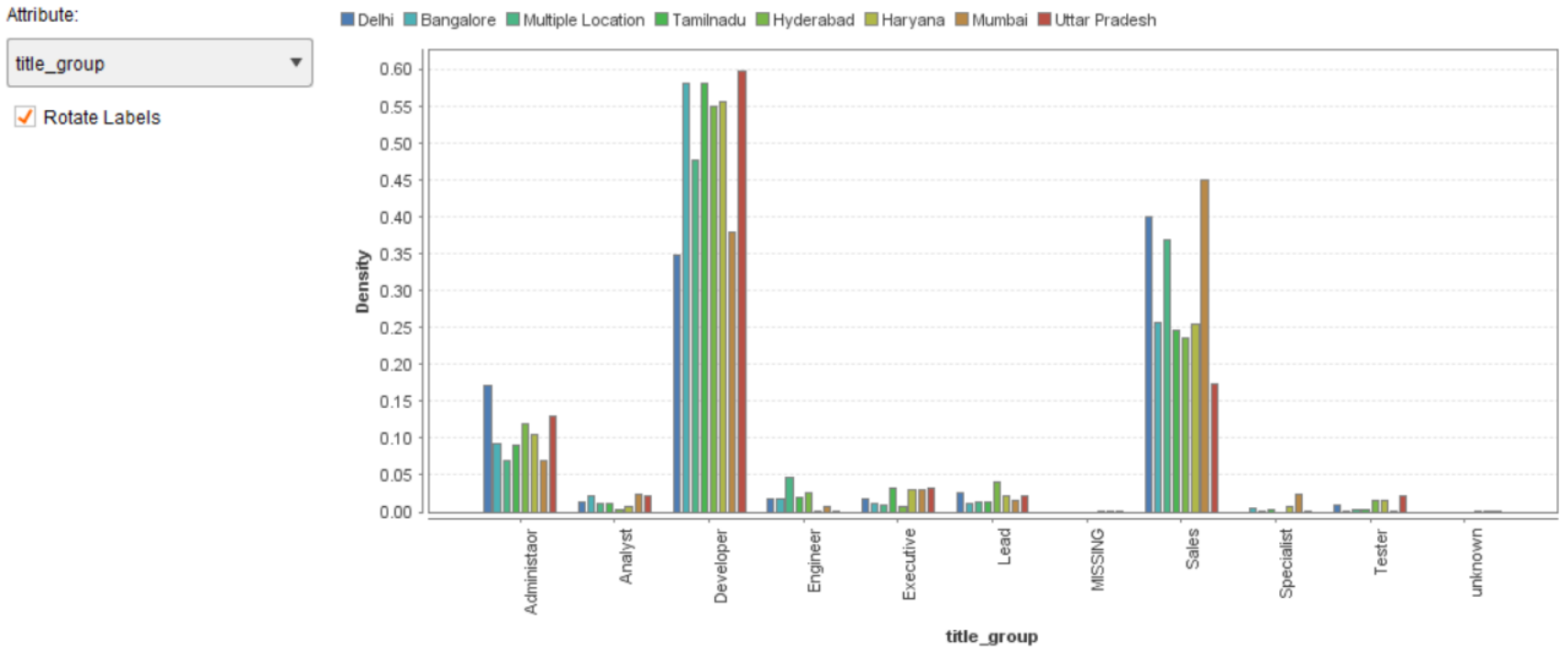
Naive Bayes - Model



Naive Bayes - Model



Naive Bayes - Model



DECISION TREE PERFORMANCE

	true Delhi	true Bang...	true Multipl...	true Hyder...	true Tamil...	true Mumbai	true Harya...	true Uttar ...	class prec...
pred. Delhi	21	0	0	10	13	9	4	1	36.21%
pred. Ban...	87	635	491	231	118	191	79	82	33.18%
pred. Multi...	0	0	0	0	0	0	0	0	0.00%
pred. Hyde...	33	0	0	73	27	50	29	23	31.06%
pred. Tami...	12	0	0	5	20	7	4	4	38.46%
pred. Mum...	98	0	0	55	31	169	36	32	40.14%
pred. Hary...	0	0	0	1	1	0	1	0	33.33%
pred. Uttar...	0	0	0	0	0	0	0	0	0.00%
class recall	8.37%	100.00%	0.00%	19.47%	9.52%	39.67%	0.65%	0.00%	

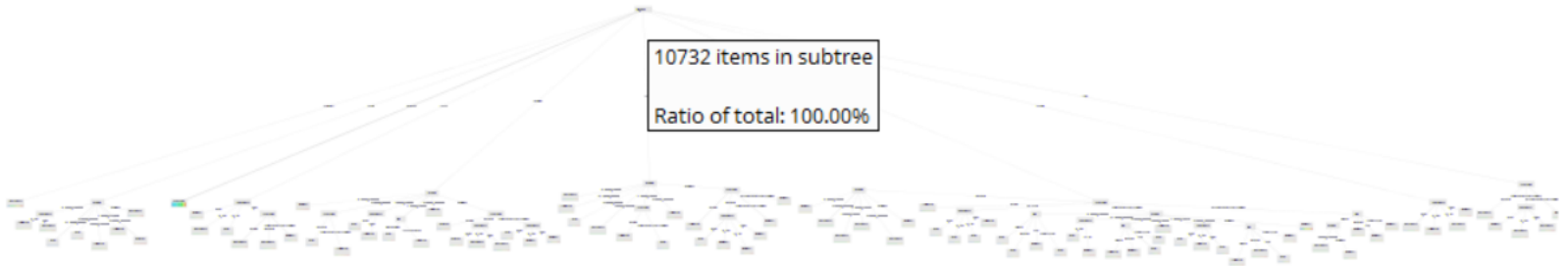
Auto Model

Select Data Select Task Prepare Target Select Inputs Model Types Results

Results

Decision Tree - Model

- General
  - Data
  - Weights
  - Correlations
- Comparison
  - Overview
- Naive Bayes
  - Model
  - Simulator
  - Performance
- Decision Tree
  - Model
  - Simulator
  - Performance
  - Optimal Parameters



Back Open Process

Information

Results: Classification

This is the final step of Auto Model, where you can inspect the generated models together with other results. The output depends on the data and the choices you made. For example, if you deactivated the calculation of correlations or decision trees, those results will not be displayed. Other results might be shown only for certain types of problems. Lift Charts, for example, are only available for two-class problems.

Please note that the results are calculated in the background. However, you can immediately start to inspect the results as they are completed. You can stop background execution by pressing the *Stop* button at the bottom. Calculations which are not completed when execution is stopped won't be available. You can go back and make changes after the execution is finished or after you stopped it.

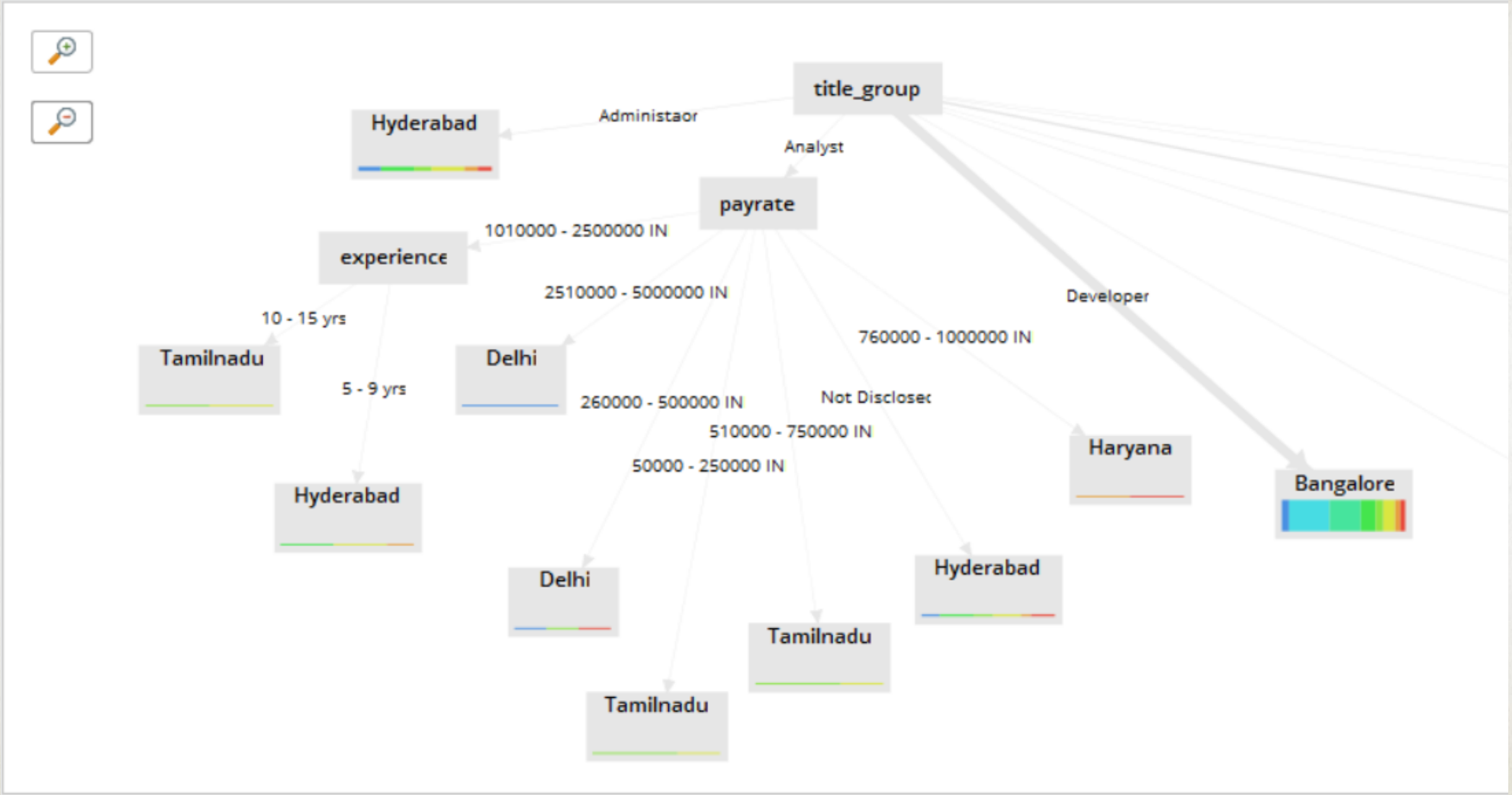
We at RapidMiner do not believe in black boxes. This is why you can always open the process which created the model and all related results. Simply click on a model result and on *Open Process* at the bottom of the screen. This will show you the process which performs all necessary data preprocessing and model optimization. You can use this process for deploying the model or as a starting point for further optimizations.

We will now discuss the possible results in detail below.

General

This section shows generic information which is independent of the models.

Decision Tree - Model



CONCLUSION

1. Implemented KDD steps
2. We have achieved our objective (Both classification and clustering)
    1. Classified pay rate and job location
    2. Clustering data into 5 clusters.
3. Learned how the data can affect the accuracy

DISCUSSION / FUTURE WORK

- The number of possible values in a column affects the accuracy.
- Imbalanced data set gives wrong result. (eg : pay rate with not disclosed)

Future Work

- Try to get a better accuracy rate
- Use these algorithm to develop a web application which can be used by employees to find out pay rate and job location
- Using the clusters we generated, develop an application to group the incoming resume.

REFERENCES

1. https://rapidminer.com/training/videos/#introductions!loading
2. https://www.kaggle.com/PromptCloudHQ/jobs-on-naukricom
3. https://www.cs.waikato.ac.nz/ml/weka/