Job Hunt – Finding a Desirable Job based on LinkedIn Data

Cummings, Groulx, Meng, Werner

SEIS 763 Final Project

Objectives

- **Hypothesis**: there are differences in salaries among all factors
  - Hard skills
  - Locations (eg. regions, states & economic level)
  - Industries
  - Companies

- What are the **trends** in data-related jobs?
  - Where I should move to?
  - What skills is demanding?

INTRODUCTION- Dataset & Data Cleaning

ML final project data cleaning PART 1.ipynb
ML final project data cleaning PART 2.ipynb
ML final project data cleaning PART 3 nlp.ipynb
ML final project data cleaning PART 4.ipynb
ML final project data cleaning PART 5.ipynb

'LinkedIn Job Postings (2023 - 2024)'

1 **124,000 x 33**

2 **32,556 x 19**

3 **32,556 x 40**

4 **5,958 x 44**

5 **3908 x 22**

Basic Data Cleaning

- Only full-time jobs
- Drop inrellevent columns
- Calculate all hourly salarty into yearly

Data Merging

- Merging company profiles data
- Merging economic level (gdp& mean income)
- Merging Sector&industry

NLP Keywords Extracting

- Extracting data related job by extract data related Keywords from job job description and job title.
- Calculate score for each skillset.
- Industry sector simplification.
- Keep only data_related job postings

Futher Cleaning

- Drop rows with too many missing values
- Select columns that could be used in our project

NLP Keywords Extractions

["data"]

is_data_related

programming_languages_score

["python", "java", "c++", "r", "sql", "javascript","matlab"]

["bachelor", "undergraduate","4 year", "master", "postgraduate", "phd", "doctorate","mba"]

degree_level

analysis_statistics_score

["excel", "tableau", "statistics", "sas", "spss","regression","decision tree","predictive analysis", "statistics","statistical",'models']

Life & Health Insurance	40	Financials
Multi-line Insurance	40	Financials
Property & Casualty Insuranc	40	Financials
Reinsurance	[illegible]	[illegible]
Information Technology	45	Information Technology
IT Services	[illegible]	[illegible]
Internet Services & Infrastru	45	Information Technology
Software	45	Information Technology

165 Industry classifications

industry_Sector

machine_learning_score

["tensorflow", "keras", "scikit-learn", "pytorch", "deep learning","nlp","svm", "clustering","natural language","machine learning", "deeep learning","ai"]

["dataset","mysql", "postgresql","mongodb", "oracle","sql","database","query"]

cloud_platform_score

data_visualization_score

["power bi", "matplotlib", "seaborn", "dash","tableau","cognos",'bi']

New Columns

Basic Linear Regression

- Started by selecting a handful of predictors that we thought would make sense to help estimate potential salary.
    - Degree Level
    - Employee Count
    - Industry
    - Work Type

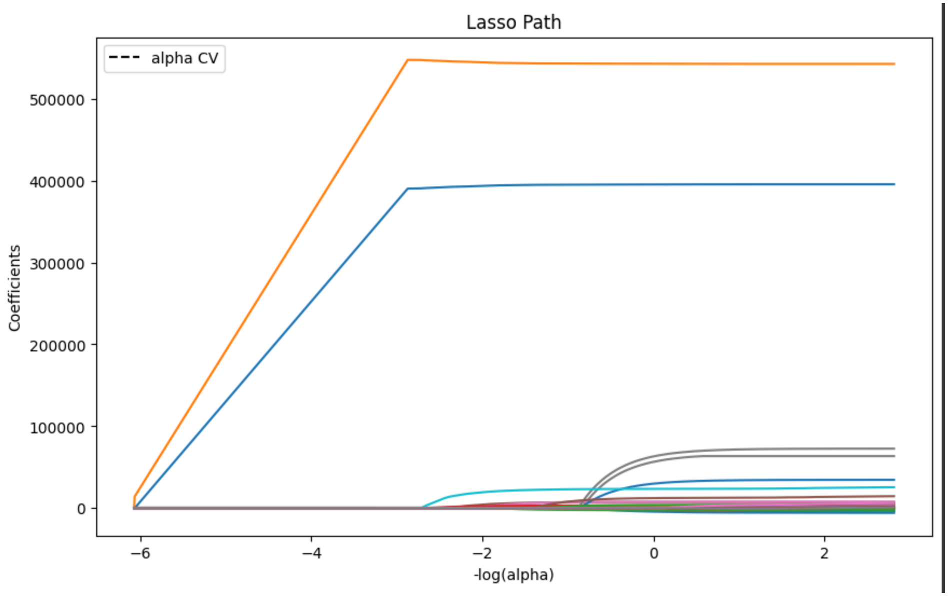
- Our inclinations were incorrect as displayed in the regression report

```
Regression Report:
R-squared: 0.0001
Adjusted R-squared: -0.0597
MSE: 4383564218238.2197
RMSE: 2093696.3052
Intercept: 99573.2902
```

Regression With Lasso

- Next we conducted Lasso Regression analysis.
- Our R2 values for both the train and test were excellent.
- We can quickly see that two predictors stand above the rest.

```
Lasso
Train:  0.9991809994956069
Test:  0.9881788513055522
Alpha:  0.0
```



Further Inspection

- Our two variables were min/max salary.

- This is possibly problematic as our dependent variable is essentially created from these two variables.

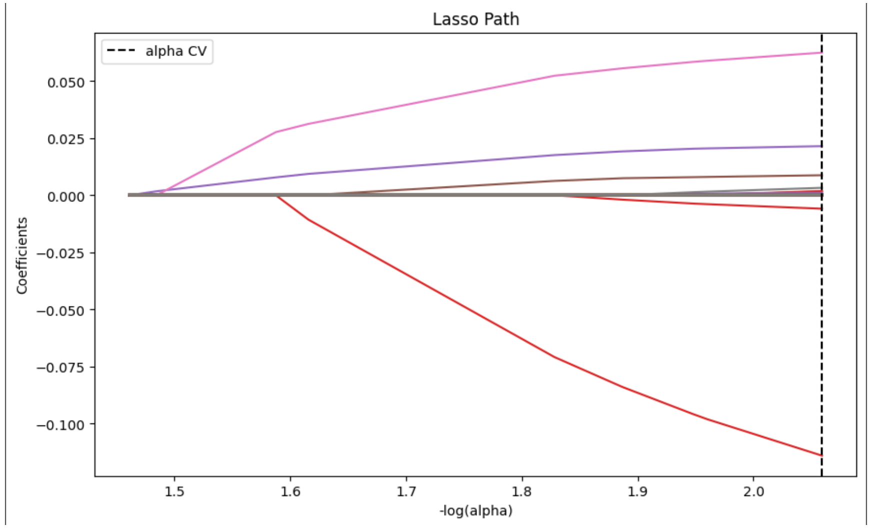
- Once we dropped those values our model fell apart.

```
Lasso
Train:  0.0
Test:  -0.019587133417183544
Alpha:  47962.1667503729
```

Logistic Analysis

- Since the model fell apart if we excluded min/max salary, we wanted to try a different approach.

- Instead, we wanted to try and create a model that could tell us our likelihood of making over $60,000.



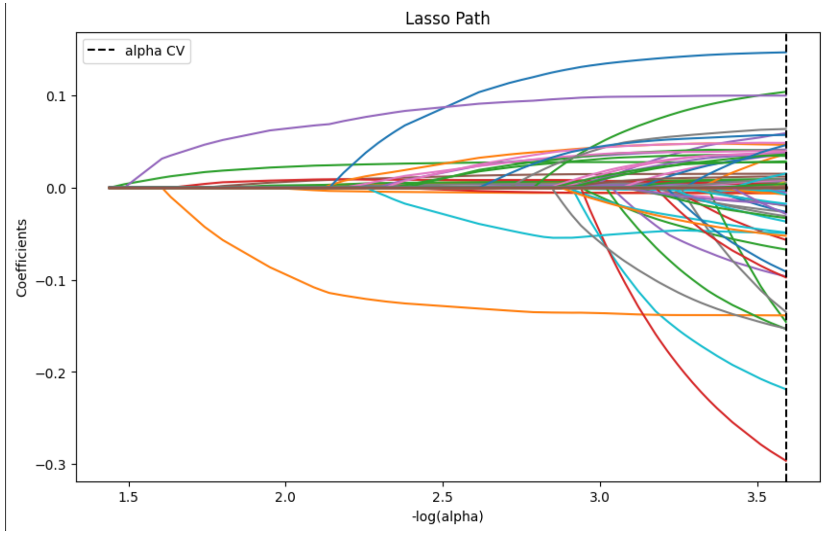
	precision	recall	f1-score	support
0	0.00	0.00	0.00	359
1	0.85	1.00	0.92	2033
accuracy			0.85	2392
macro avg	0.42	0.50	0.46	2392
weighted avg	0.72	0.85	0.78	2392

```
Lasso
Train:  0.06617149394884903
Test:  0.06376705625913592
Alpha:  0.008804043275673527
```

	feature	importance
73	formatted_experience_level_Entry level	0.114
76	formatted_experience_level_Mid-Senior level	0.062
4	programming_languages_score	0.022
15	follower_count	0.009
3	day_posting	0.006

Logistic Continued

- Interestingly, dropping min/max salary like we did before actually improved our model.



	precision	recall	f1-score	support
0	0.33	0.01	0.01	345
1	0.86	1.00	0.92	2047
accuracy			0.85	2392
macro avg	0.59	0.50	0.47	2392
weighted avg	0.78	0.85	0.79	2392

Lasso
Train: 0.12221512860534323
Test: 0.11908110748789347
Alpha: 0.0002421181323405178

	feature	importance
73	formatted_experience_level_Internship	0.296
19	location_AR	0.219
112	headquarter state_AL	0.153
127	headquarter state_IN	0.153
70	formatted_experience_level_Director	0.147

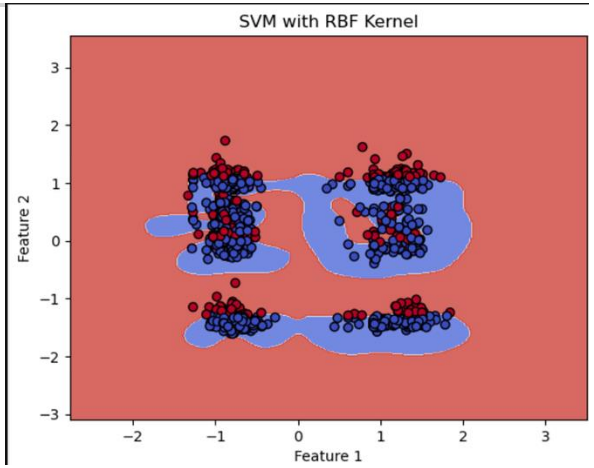
SVM Model 1 with PCA features

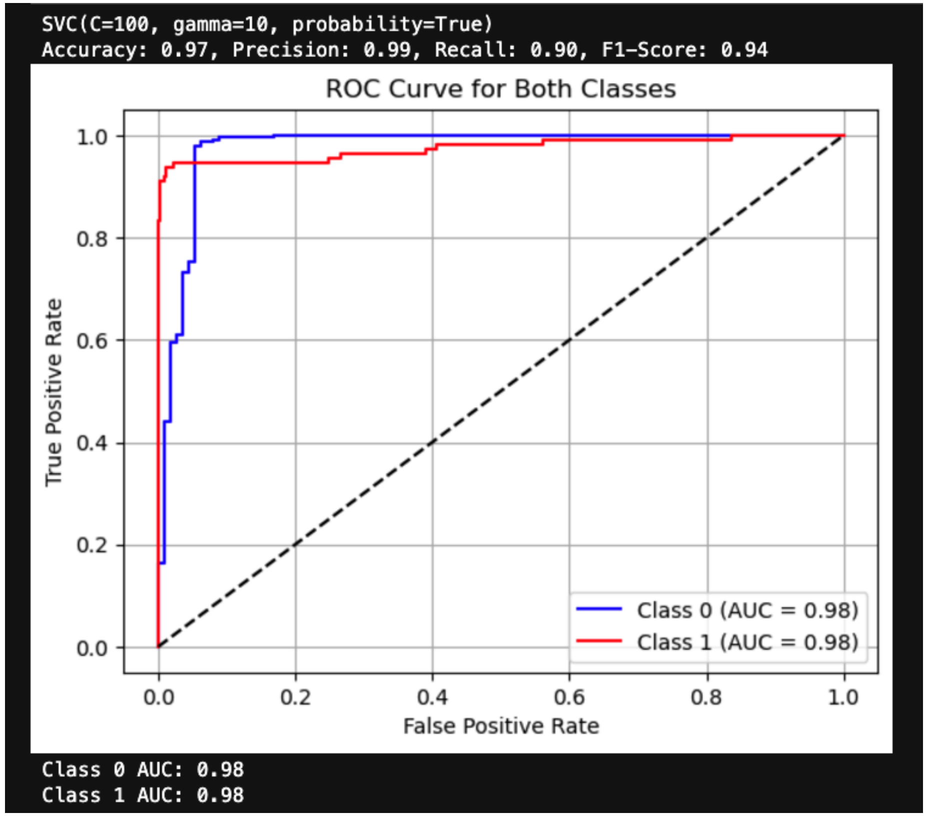
(best lambda& C tested)

```
Best Parameters: {'C': 100, 'gamma': 10, 'kernel': 'rbf'}
Classification Report:
              precision    recall  f1-score   support

           0       0.73      0.54      0.62       330
           1       0.23      0.41      0.30       113

    accuracy                           0.51       443
   macro avg       0.48      0.47      0.46       443
weighted avg       0.60      0.51      0.54       443
```

- Predict <140K vs. >140K

F.Y.I
140k is the 3rd quantile
100k is the median



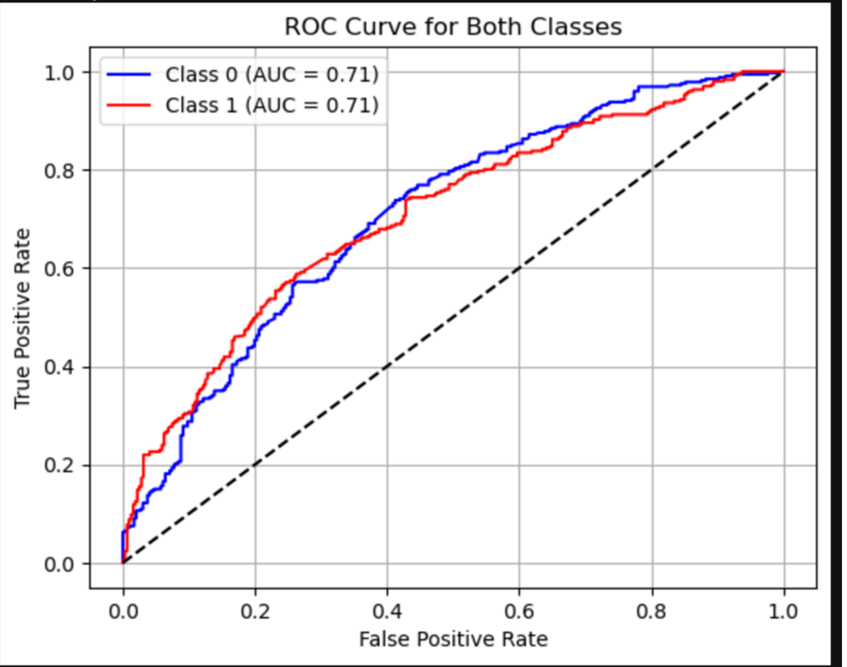


SVM Model 2 with Top 5 Predictors

- company_score
- economic_score
- company_size
- programming_languages_score
- job_classification_Engineering

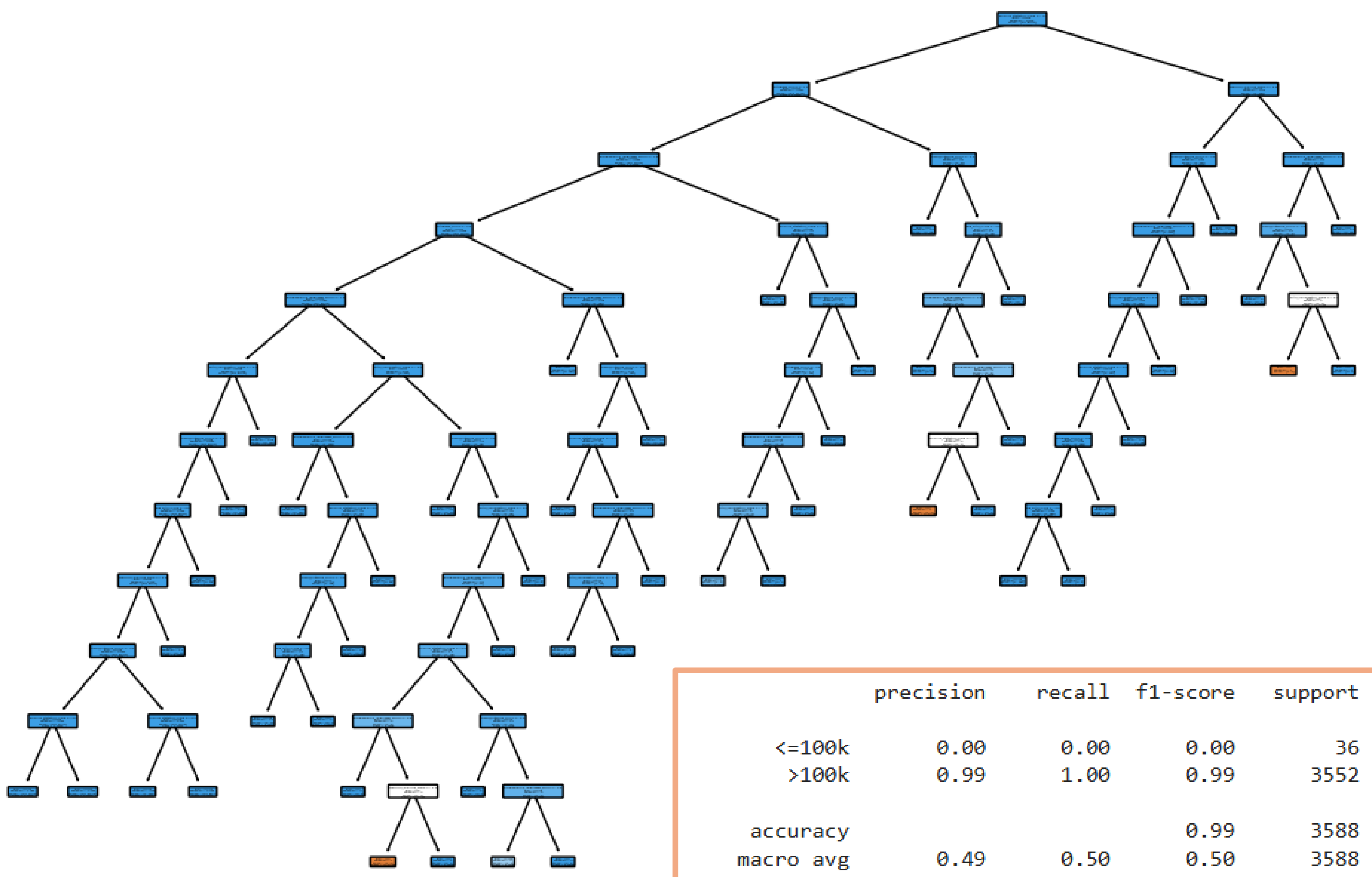
```
Best Parameters: {'C': 1, 'gamma': 0.1, 'kernel': 'rbf'}
Classification Report:
              precision    recall  f1-score   support

           0       0.62      0.73      0.67       320
           1       0.64      0.52      0.57       296

    accuracy                           0.63       616
   macro avg       0.63      0.63      0.62       616
weighted avg       0.63      0.63      0.63       616
```

```
SVC(C=1, gamma=0.1, probability=True)
Accuracy: 0.66, Precision: 0.67, Recall: 0.58, F1-Score: 0.62
```




```
Class 0 AUC: 0.71
Class 1 AUC: 0.71
```

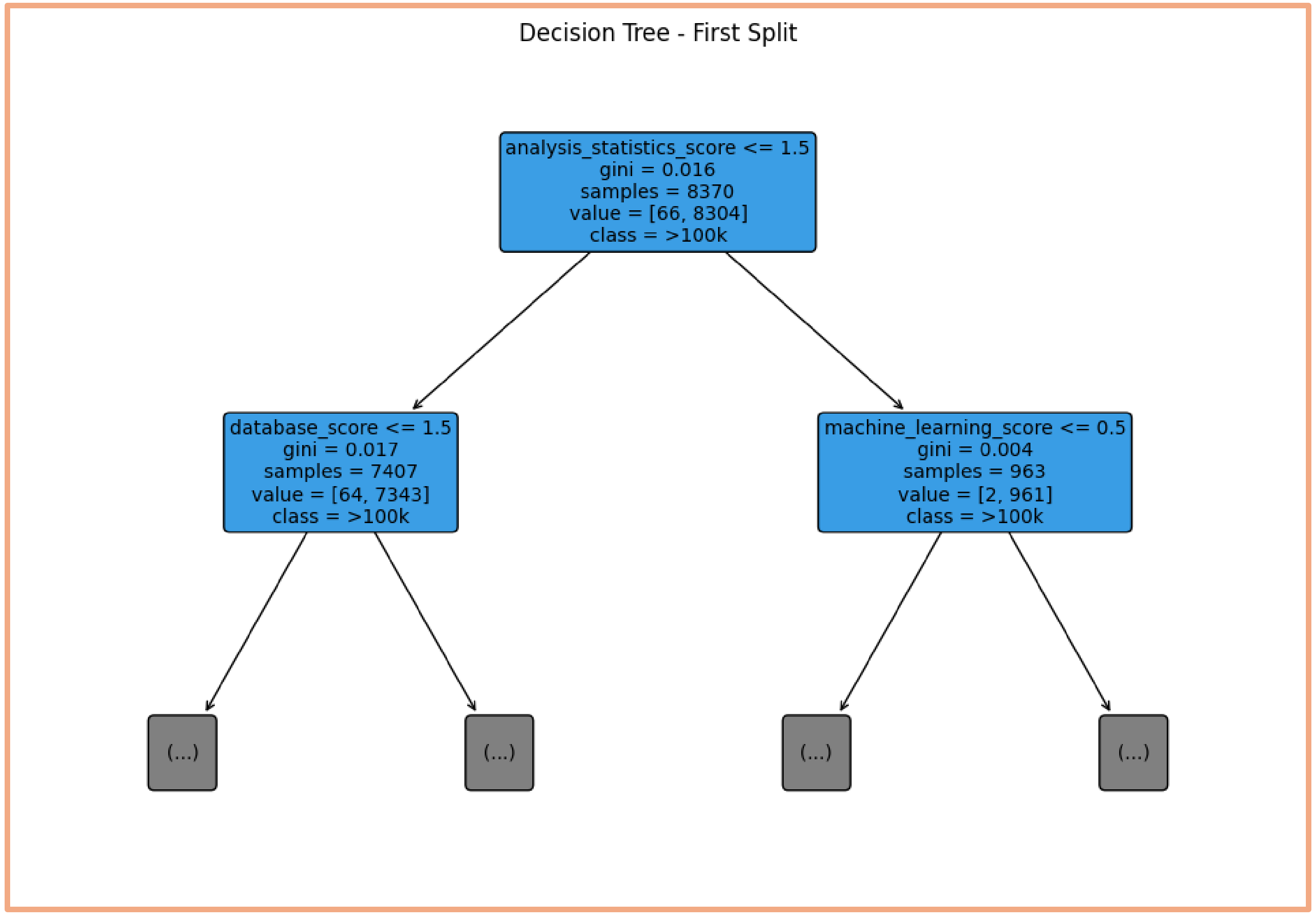
Decision Tree – Predicting Salary over $100k



	precision	recall	f1-score	support
<=100k	0.00	0.00	0.00	36
>100k	0.99	1.00	0.99	3552
accuracy			0.99	3588
macro avg	0.49	0.50	0.50	3588
weighted avg	0.98	0.99	0.98	3588

- **Precision, Recall, and f1-Score:** Great at predicting salaries over $100k
    - Less accurate for salaries under $100k, due to smaller sample size in dataset
- **Root Nodes:**
    - analysis_statistics_score
    - database_score
    - machine_learning_score

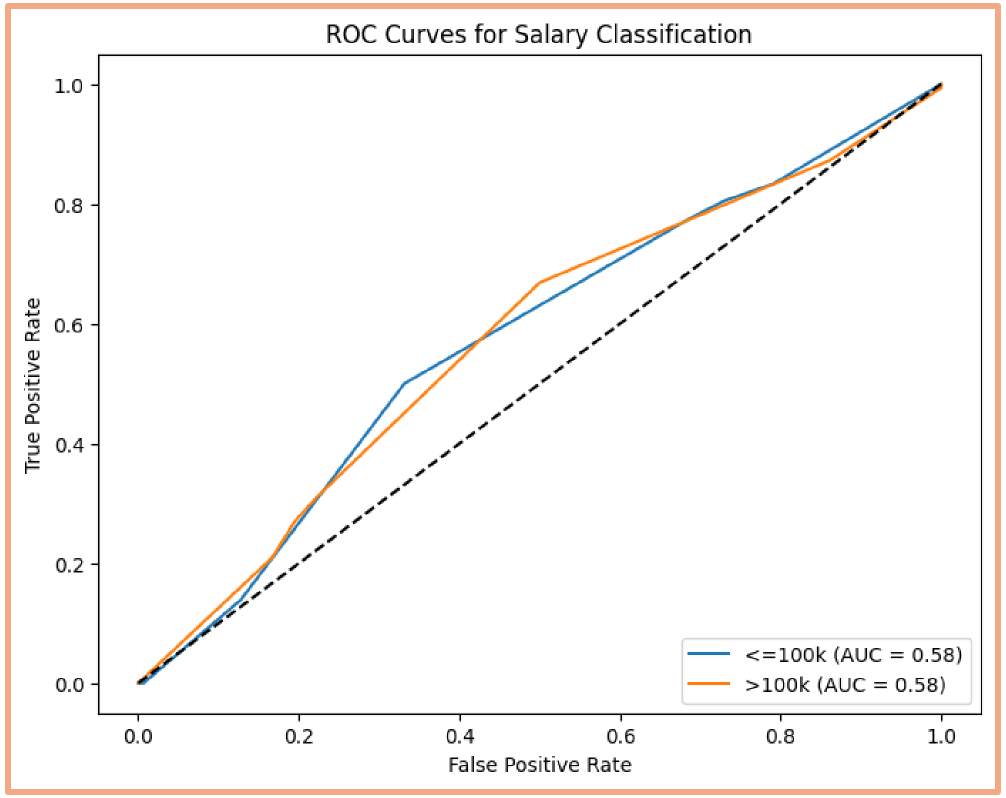
Decision Tree – Root Node Analysis



- **Root Node Decision:** Having an analysis_statistics_score less than or equal to 1.5
- **Higher Analysis Score:** Database score then needs to be less than or equal to 1.5
- **Lower Analysis Score:** Machine learning score then needs to be less than or equal to 0.5

Decision Tree – Validation

- **<=100k:** AUC of 0.58, performing slightly better than a coin flip
- **> 100k:** AUC of 0.58, performing slightly better than a coin flip
- **Accuracy:** (3549+0)/(3549+0+36+3) = 98.9% accuracy, with bias to class I

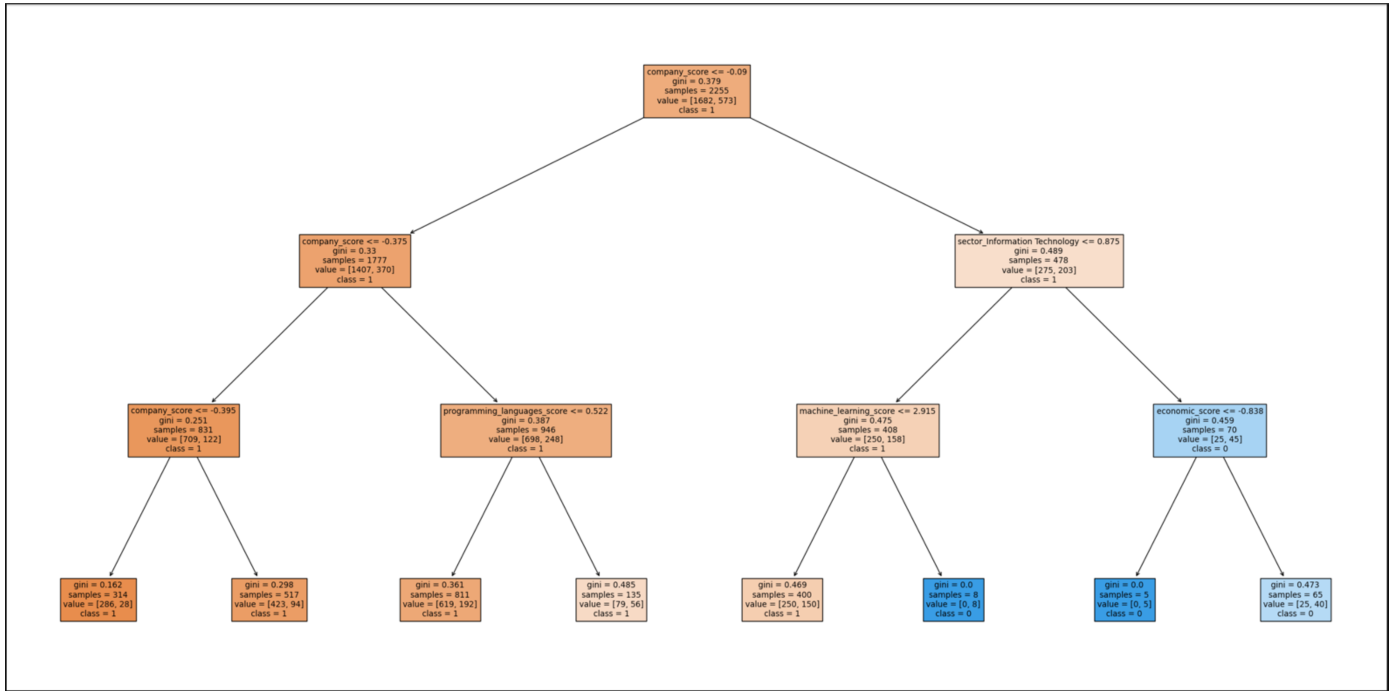


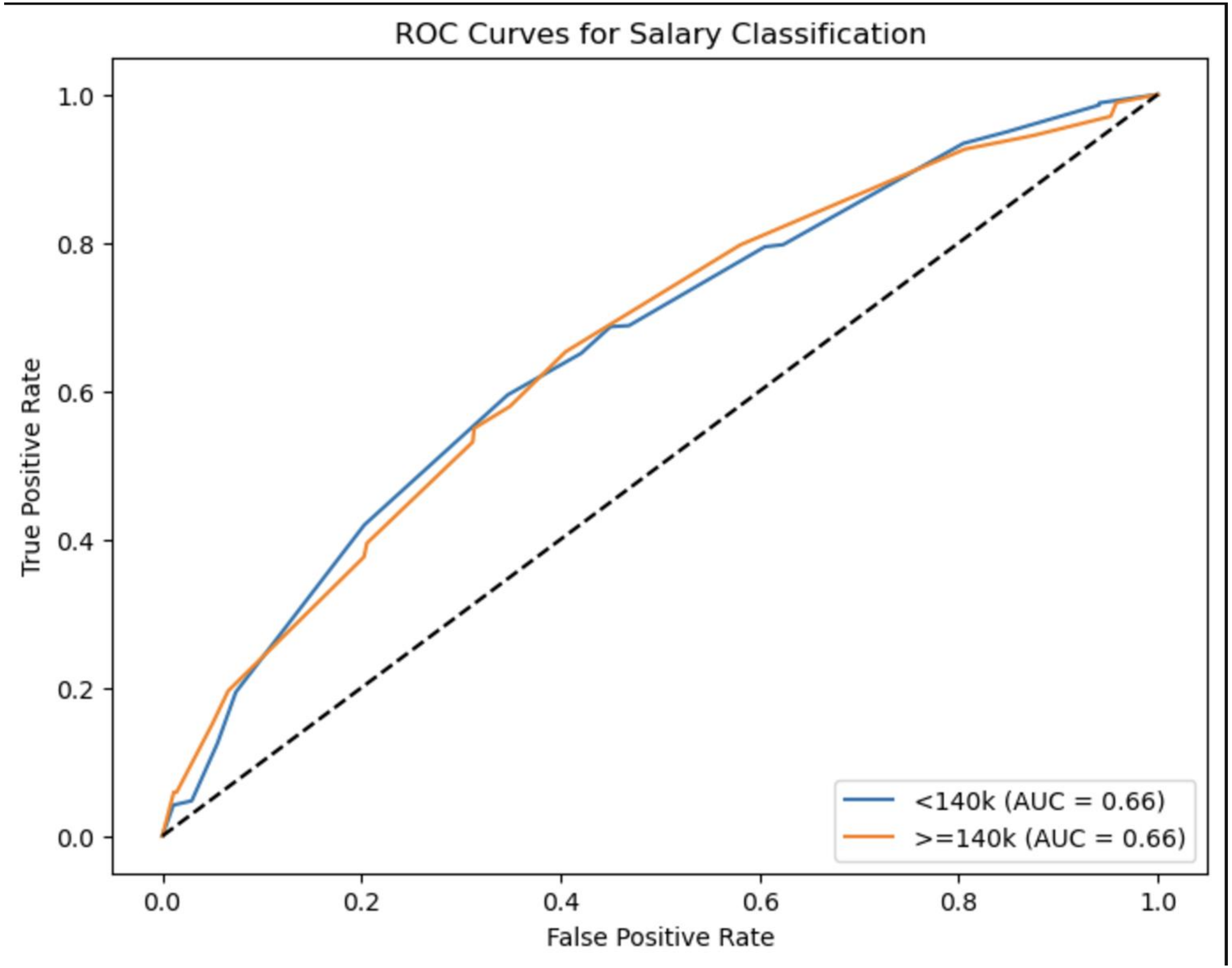
Confusion Matrix	False	True
False	0	36
True	3	3549

DT Model2- 140k Predicting

- company_score
- job_classification_Engineering
- programming_languages_score
- sector_Information Technology
- machine_learning_score
- economic_score

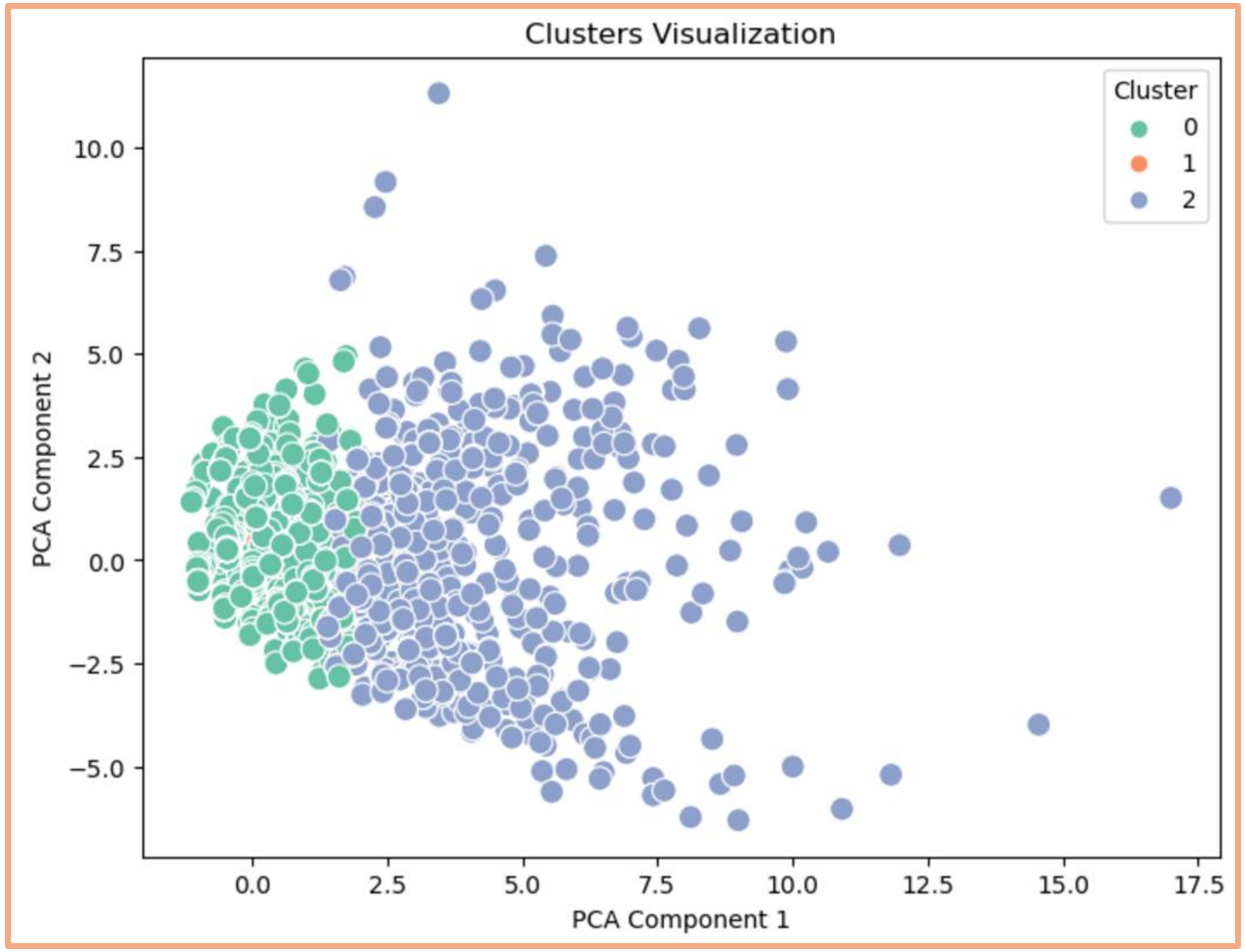
	precision	recall	f1-score	support
<=140k	0.74	0.98	0.84	696
>140k	0.62	0.10	0.17	271
accuracy			0.73	967
macro avg	0.68	0.54	0.50	967
weighted avg	0.70	0.73	0.65	967




ROC Curves for Salary Classification
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
<140k (AUC = 0.66)
>=140k (AUC = 0.66)

Clustering Analysis – Overview

Clustering Analysis of Job Market Trends for Data Majors



Clustering Methods Used:

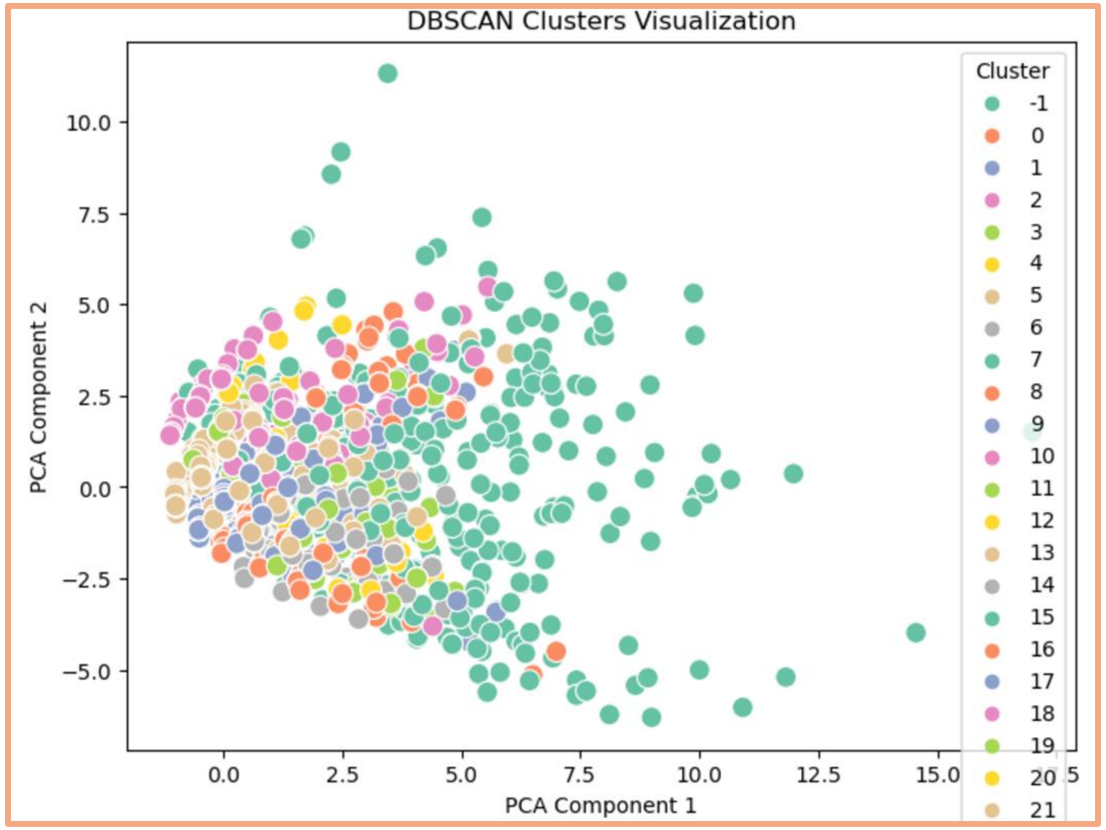
- **K-Means**: 3 clusters based on salary, views, skills, and job duration.
- **Hierarchical Clustering**: Explored relationships between clusters using a dendrogram.
- **DBSCAN**: Identified dense regions, noise points, and smaller clusters.

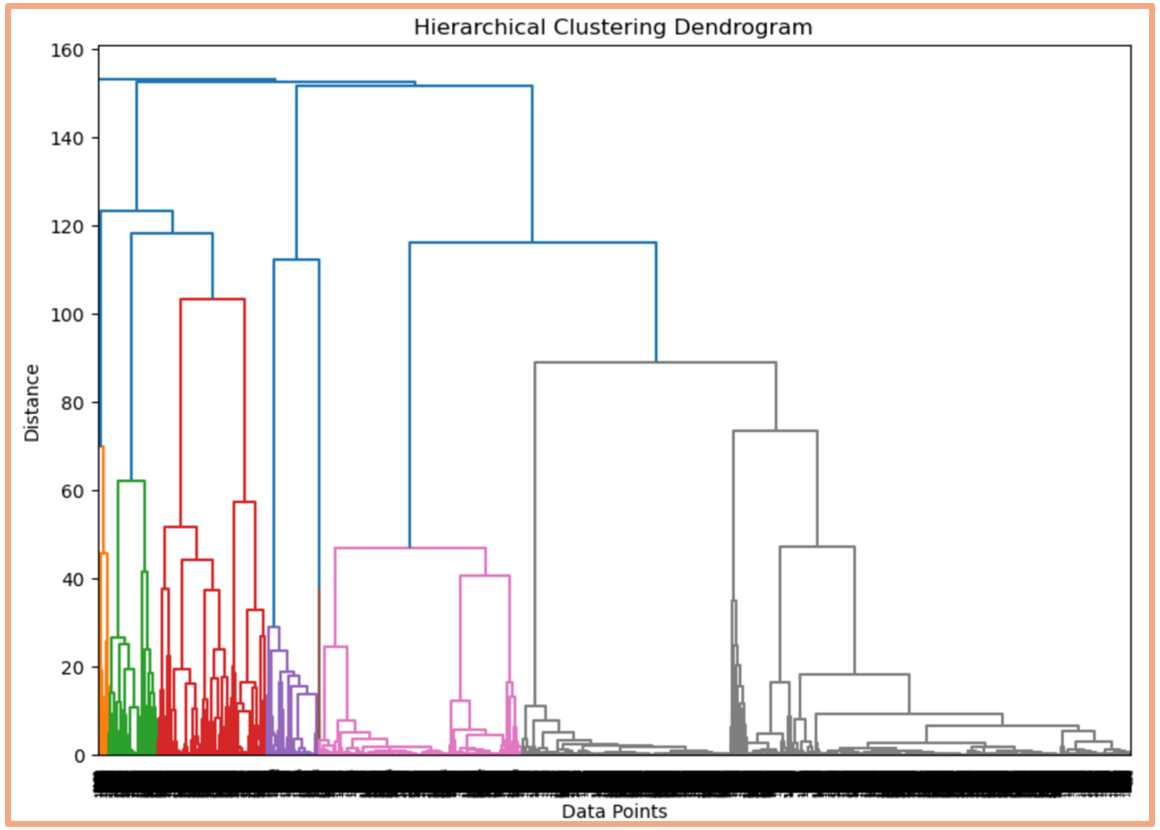
Key Insights:

- **K-Means**:
  - Cluster 0: Moderate salary, balanced skills.
  - Cluster 1: Extremely high salary, minimal skills.
  - Cluster 2: Low salary, high advanced skills (e.g., ML, data visualization).
- **Silhouette Score (K-Means)**: 0.601 (indicates good separation).

Clustering Analysis – Results

Detailed Results of Clustering Approaches





Clustering Highlights:

K-Means:

- Distinct clusters with clear differences in salary and skill requirements.

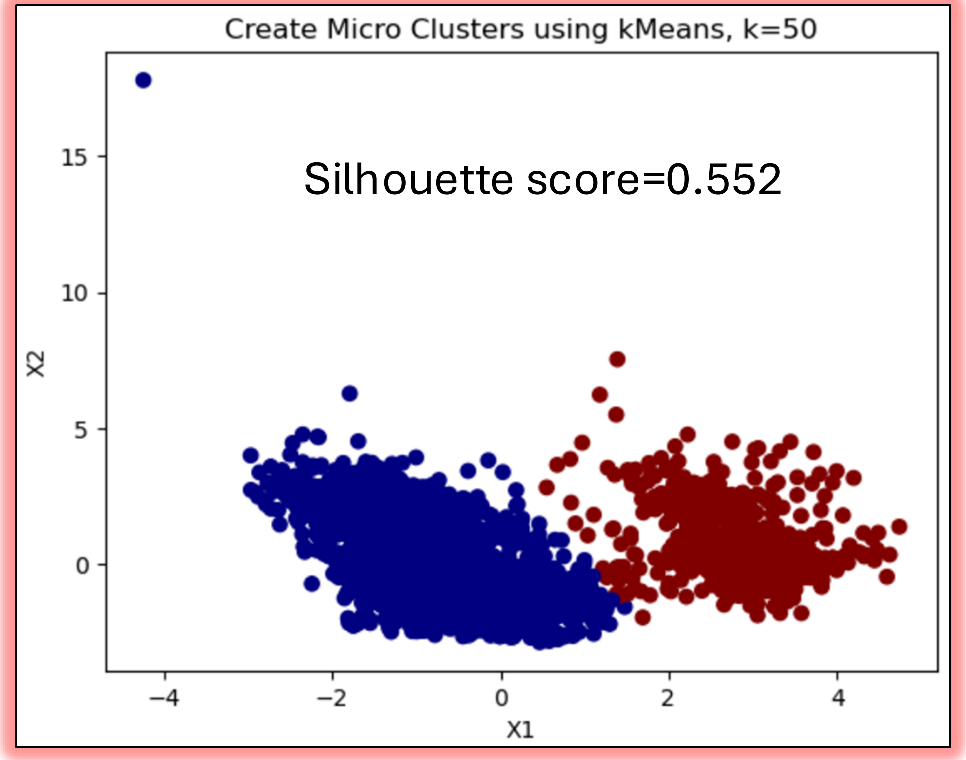
Hierarchical Clustering:

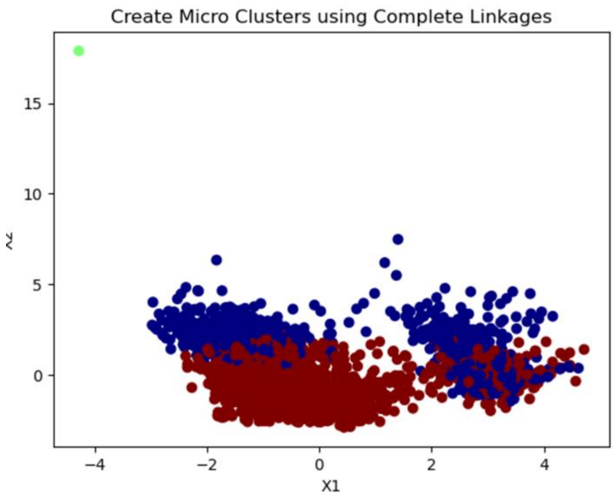
- Dendrogram shows relationships between clusters (e.g., Clusters 0 and 2 are more similar; Cluster 1 is an outlier).

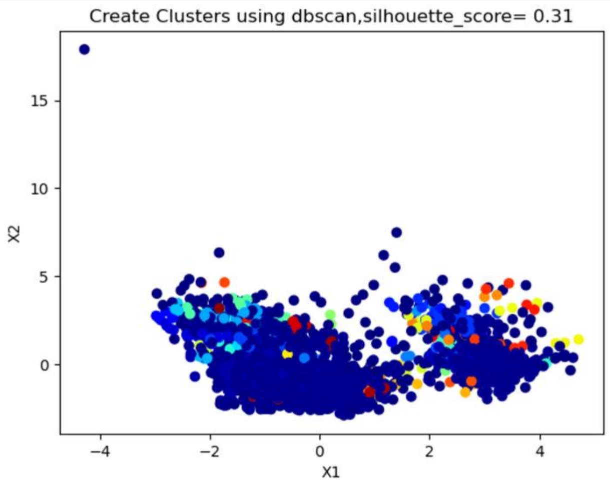
DBSCAN:

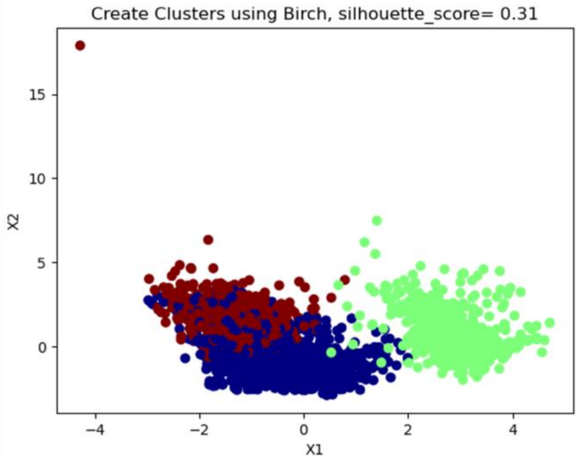
- 30 clusters and one noise group (-1).
- Detects niche job roles and unique postings.

Clustering Model 2: PCA & K-means









Summary

Linear Regression & logistic Models

Linear relationship is weak, which means the salary could not be predicted through regression models.

SVM & Decision Tree Models

Key predictors to predict salaries 140k or more:

- **Company**
- **Job location**
- **IT industry**
- **Engineering position**
- **Machine learning skills**

Decision Tree Models

Key predictors to salaries 100k or more

- **Analysis statistics score**
- **Database score**
- **Machine learning score**

Clustering Models

- Data Crew: Moderate salary, balanced skills.
- Ones who born in Rome: Extremely high salary, minimal skills.
- Wage Theft : Low salary, high advanced skills (e.g., ML, data visualization).

Thank You!
Q&A