SQL vs Python

Background

- In recent years, Data science has become one of the most popular fields that employees and companies are exploring.
    - The US Bureau of Labor Statistics predicts that mathematician and statistician roles, including data scientist jobs, will experience **36 percent growth between 2021 and 2031**, which is much faster than the average 8 percent for all occupations.*

Employer's Point of View:

- Proven to drive innovation process and increase efficiency

Employee's Point of View:

- Increased employment opportunities
- Salary commanded is growing

*https://www.coursera.org/articles/data-scientist-jobs

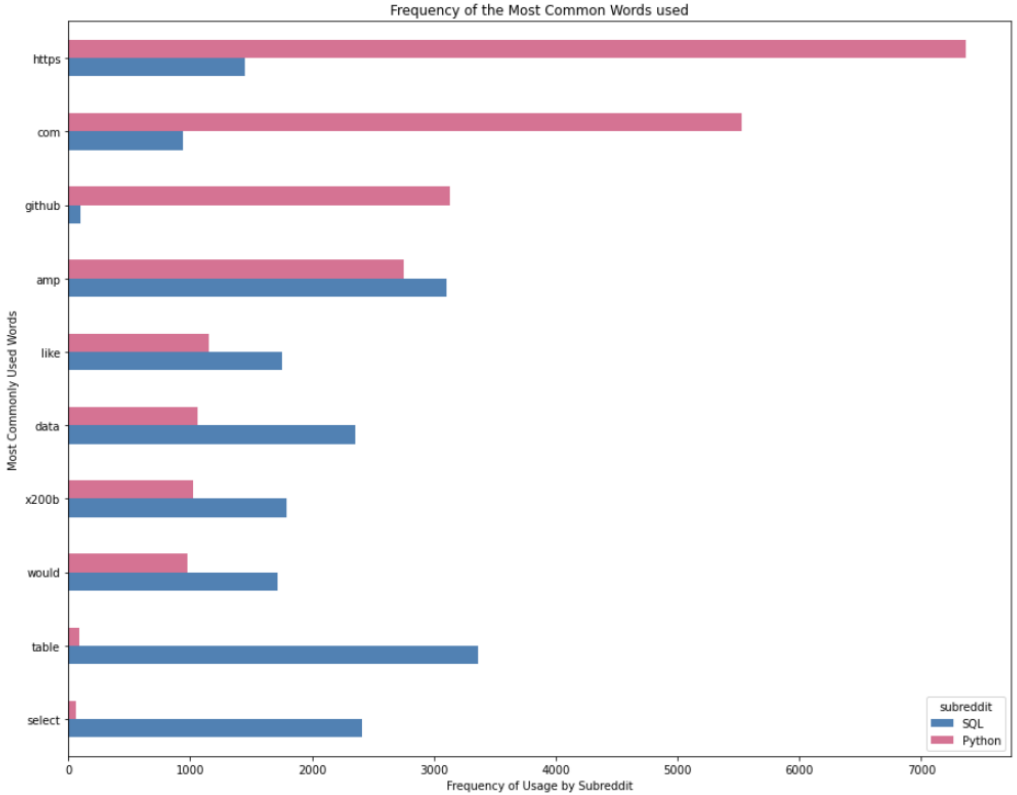
Background

- This led to a growing community who are upskilling themselves to transit into Data Science roles.
- With Technology, self learning resources are more accessible and communities for sharing are growing too.
- Specifically, Reddit is a popular platform for communities to post and share their questions and resolutions.

Problem Statement

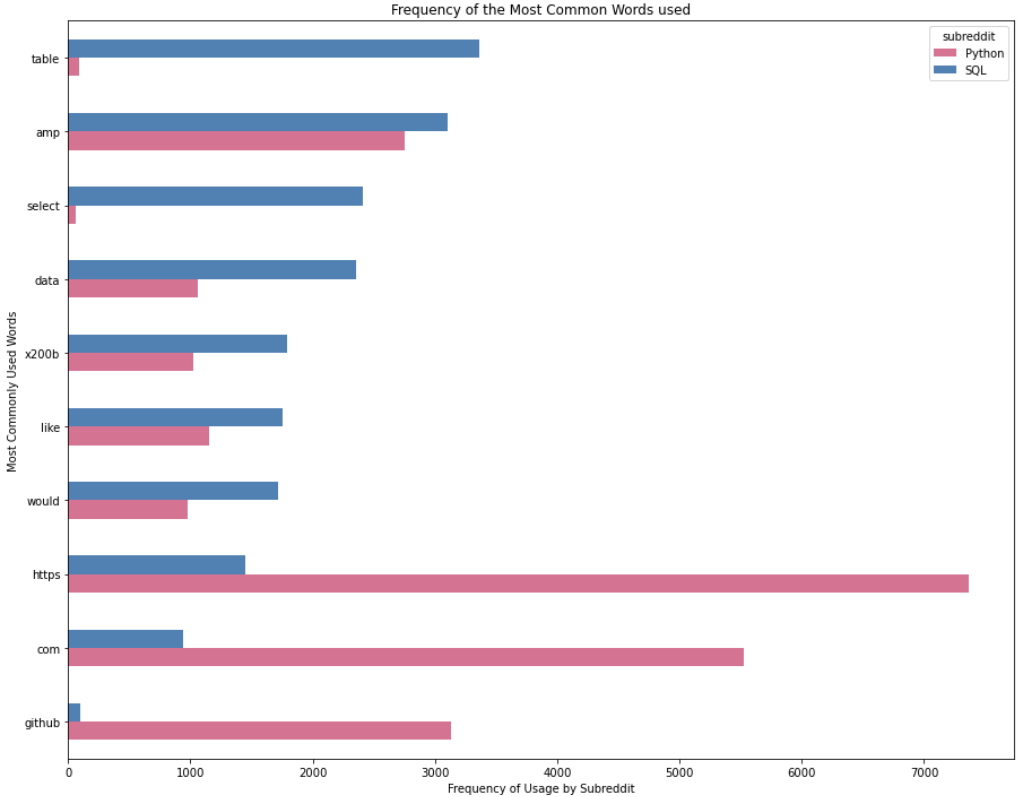
- Although the internet is an efficient and convenient place for self learning, there is just too much content that is available.
- The resources that have been searched and found, might not what you are looking for.
    - For example, while looking for resources on Python, what is found could be for other languages (for example, SQL). Time would be wasted to filter out the wrong contents.
- This project attempts to classify posts from Reddit, and identify if it was posted in subreddit Python or SQL.

Top Words used in Python Posts



Words	Python	SQL
https	7,370	1,448
com	5,527	942
github	3,131	97
amp	2,750	3,106
like	1,152	1,750
data	1,061	2,355
x200b	1,023	1,789
would	973	1,717
table	92	3,365
select	60	2,406

Top Words used in SQL Posts



Words	Python	SQL
table	92	3,365
amp	2,750	3,106
select	60	2,406
data	1,061	2,355
x200b	1,023	1,789
like	1,152	1,750
would	973	1,717
https	7,370	1,448
com	5,527	942
github	3,131	97

Distinct Words that are only used in either Python or SQL

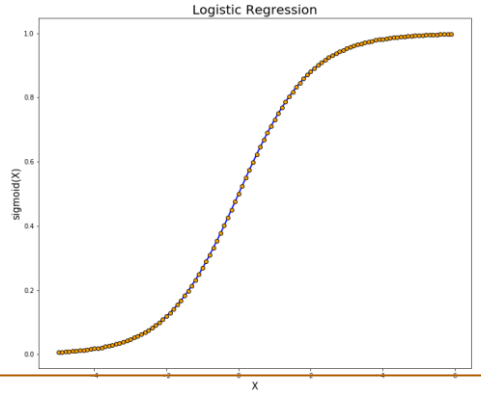
Words used in Python that are not used in SQL

Words	Python	SQL
pypi	254	0
pygame	152	0
tkinter	138	0
selenium	101	0

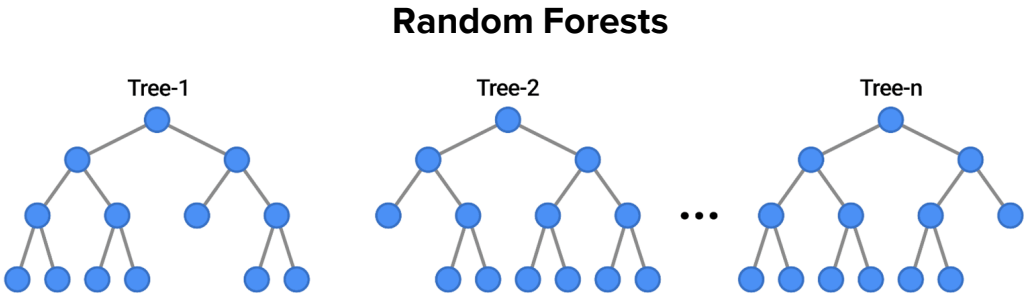
Top words used in SQL that are not used in Python

Words	Python	SQL
clause	0	154
cte	0	134
subquery	0	104
table1	0	95
ssms	0	77

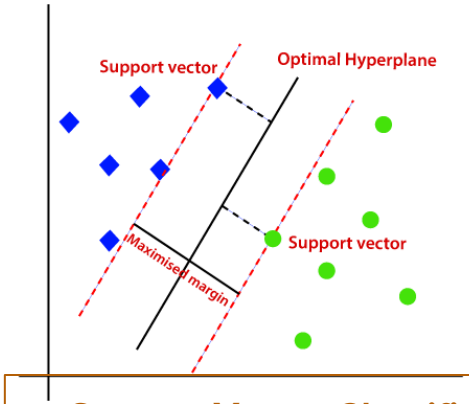
Models used



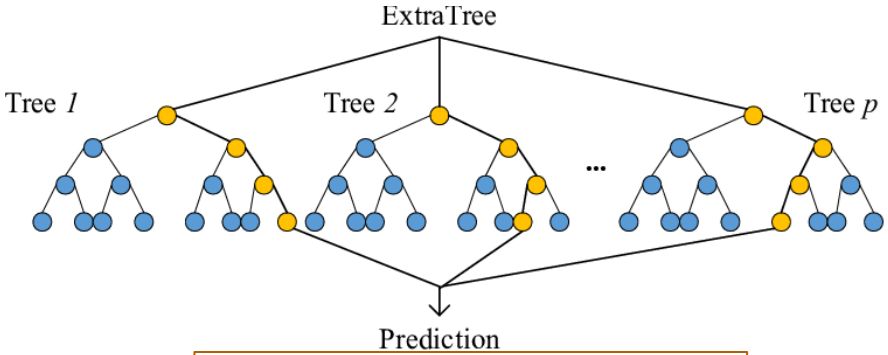
Logistic Regression (w/ L2 reg)



Random Forests

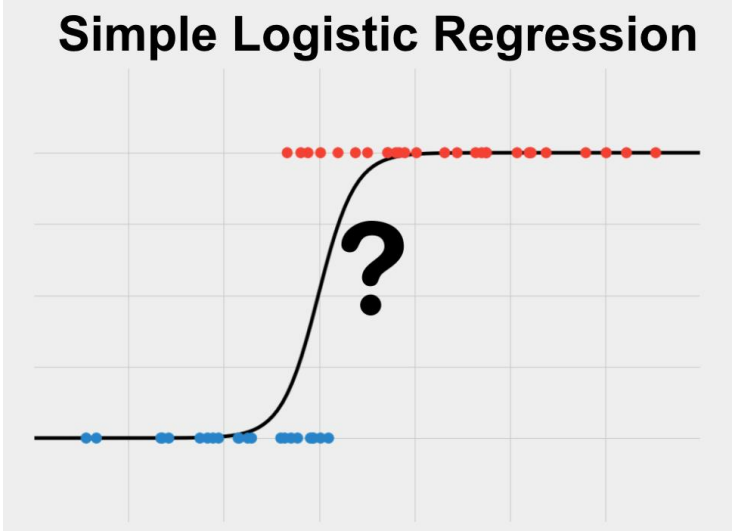


Support Vector Classifier



Extremely Randomized Trees

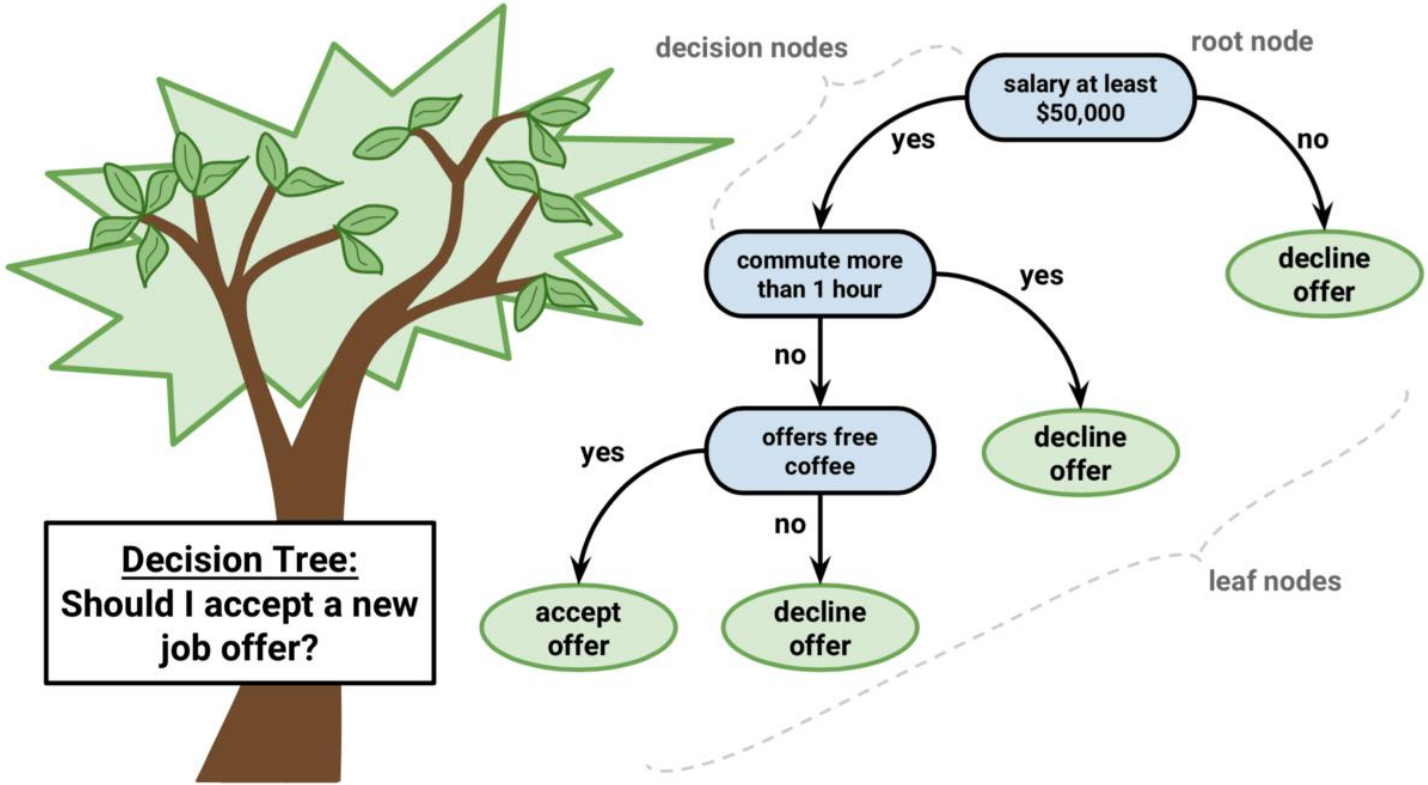
Logistic Regression



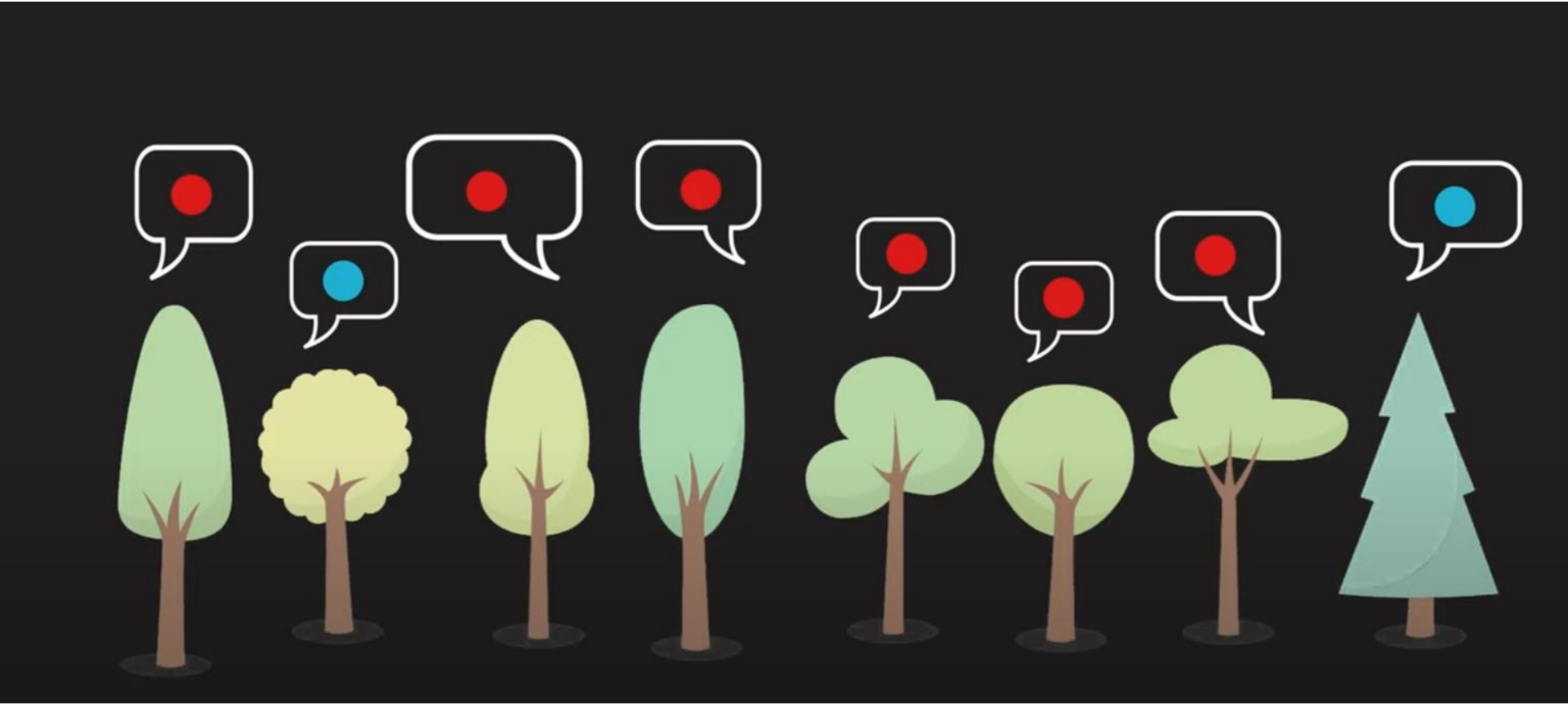
$$Ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

$$\min_{x \in \mathbb{R}^d} f(x) \triangleq \sum_{i=1}^{n} \left(\ln\left(1 + e^{\langle a_i, x \rangle}\right) - b_i \langle a_i, x \rangle \right) + \frac{\lambda}{2} \|x\|^2$$

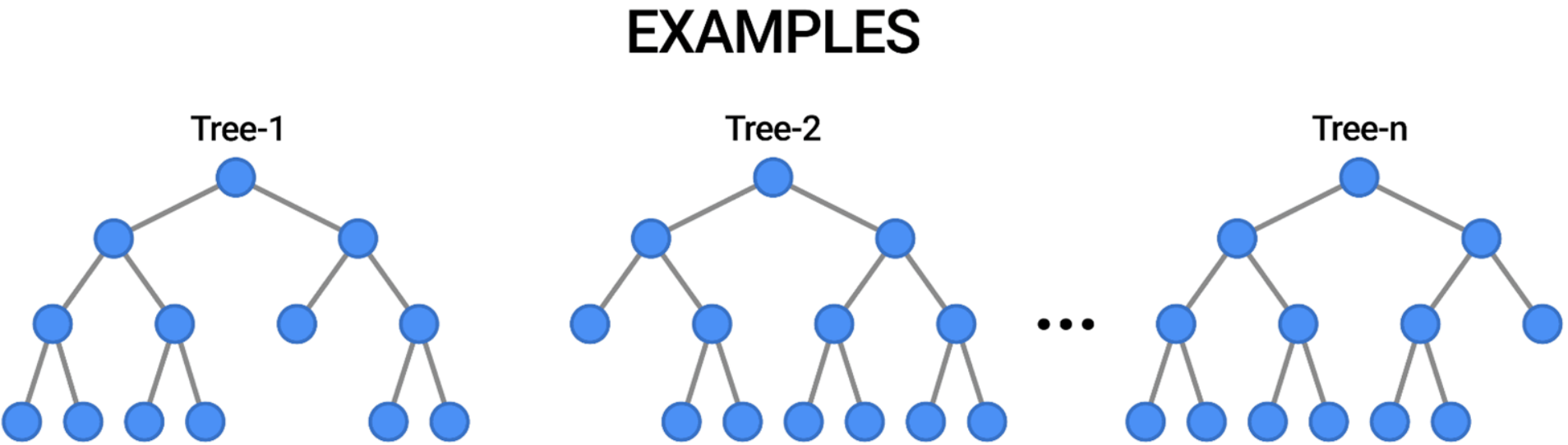
Decision Tree - A tree that thinks



Voting is compulsory



Each tree thinks like an individual



Random Forest vs ExtraTrees[1]

Random Forest	**ExtraTrees**
Uses bootstrapped dataset	Full dataset
Uses randomized subset of features at each split	Full set of features at each split
Optimally splits the tree based on Gini impurity (or other metrics)	Randomly split the tree based on any feature
Works better when some features are irrelevant	Works better when all features are relevant
Less memory intensive, but slower	Faster, but requires more memory

[1] P. Guerts et al. “Extremely randomized trees”, 2005

Support Vector Machine

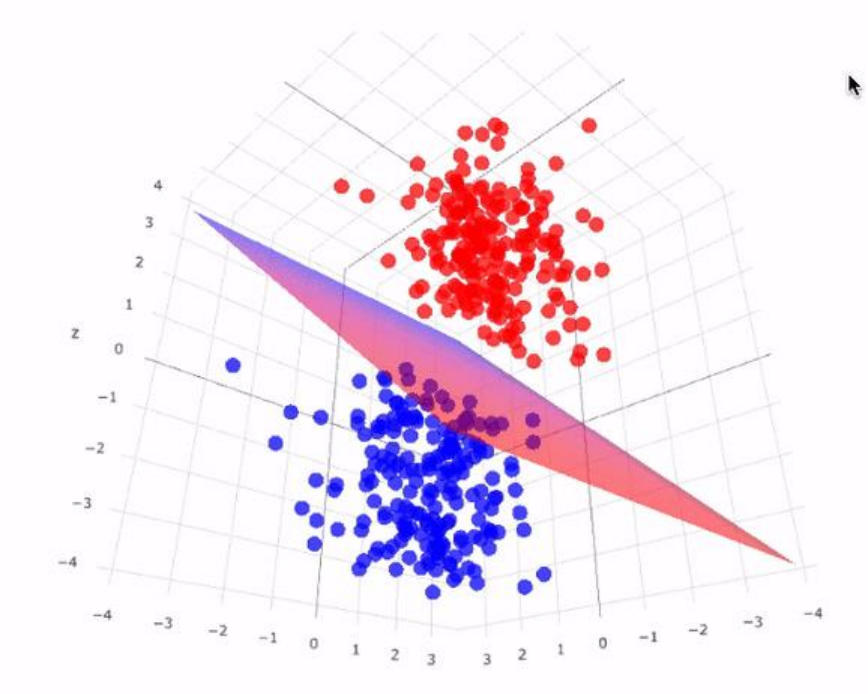
Looks cool. Sounds cool.

Performs rather well.

But slow when dealing with high dimensions

Complexity = $O(nd^2)$ [2]

And it is a black box



[2] V. Sreekanth et al. “Generalized RBF feature maps for Efficient Detection”, 2010

Models used

Coin flip (Baseline)	**Logistic Regression**	**Random Forests**	**Extra Trees**	**Support Vector Machines**
0.5	Train = 0.9897 Test = 0.9216	Train = 0.8869 Test = 0.8695	Train = 0.9124 Test = 0.9128	Train = 0.9616 Test = 0.922

Recommendation and conclusion

From the models, Logistic Regression has the high accuracy of 92.2%. Although there is significant overfitting, we would go ahead and use Logistic Regression for the ease of explanation and efficiency.

With this model, it can help to better identify if the resolution that was searched was for Python or SQL. This could make learning more efficient and useful. Specifically, we have also identified top 'words' that signals stronger.

Recommendation and conclusion

The top 5 words that signals that a resolution is 'Python' are:

Words	Coefficients
thread	5.090441
pandas	4.607804
automate	3.775485
def	3.540431
text	3.505633

The bottom 5 words that signals that a resolution is 'Python' are:

	Coefficients
queries	0.100162
query	0.113452
select	0.148223
snippets	0.173029
database	0.173148

Recommendation and conclusion

Moving forward, the below actions are recommended for better classification of the posts:

Collection of more data

Using more data columns for modelling (instead of only using 'selftext')

Run more models to check if there are other models with better performance