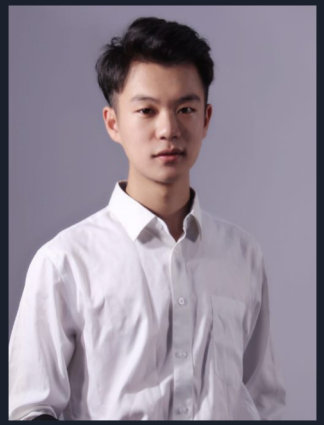
Will Badart



Wenhao Yu

Luke Duane

Determining Predictors of **H-1B** Approval

Introduction

What is the H-1B Visa?

- A visa in the US under the INA
- Given out to foreign workers at US company
- In 2017, 350,000 individuals applied and 200,000 were approved.

Problem

- How can we predict the approval status of a given H-1B visa application?
- What tangential analyses provide tangible business value for companies sponsoring H-1B visas?

Data Set

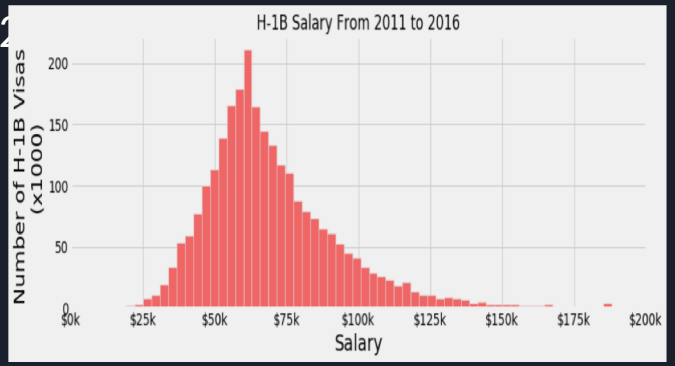
The data we used is from Kaggle. It contains over 3 million records and tracks 10 different features per application. https://www.kaggle.com/asavla/h1-visa/data

Description

- Count: 3,002,458 & Year: 2011-2016
- Attribute: Status (Y or N) / Employer name(Nominal) / Job title (Nominal) / Full time(Y or N) / Wage(Numerical) / Year(Numerical) / Worksite(Numerical)
- Salary : Mean 63,658 / Median 62,52

Data Cleaning

- Delete incomplete data
- Move outlier (wage)



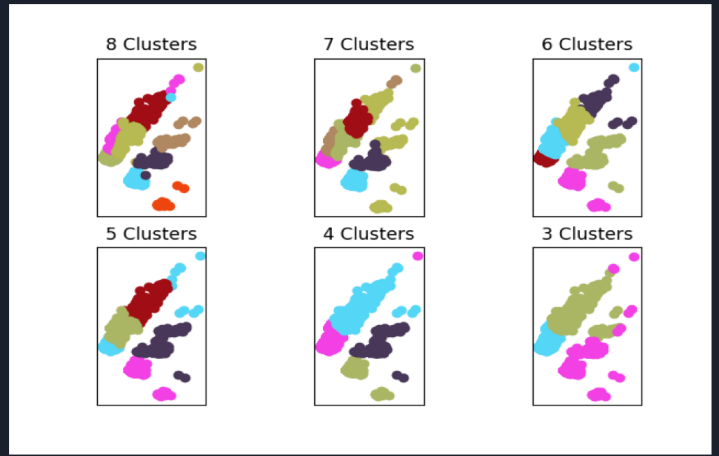
The Road Map

Before Milestone :

1. Described and visualized the data
2. Cleaned the data
3. Implemented simple NB and DT model
4. Evaluated the results

After Milestone:

1. Figured out the model issues (Overfitting / Zero Probability ...)
1. Clustered the Job Title
2. Tested varying parameters
3. Re-implemented NB, DT and MLP model
4. Evaluated the results



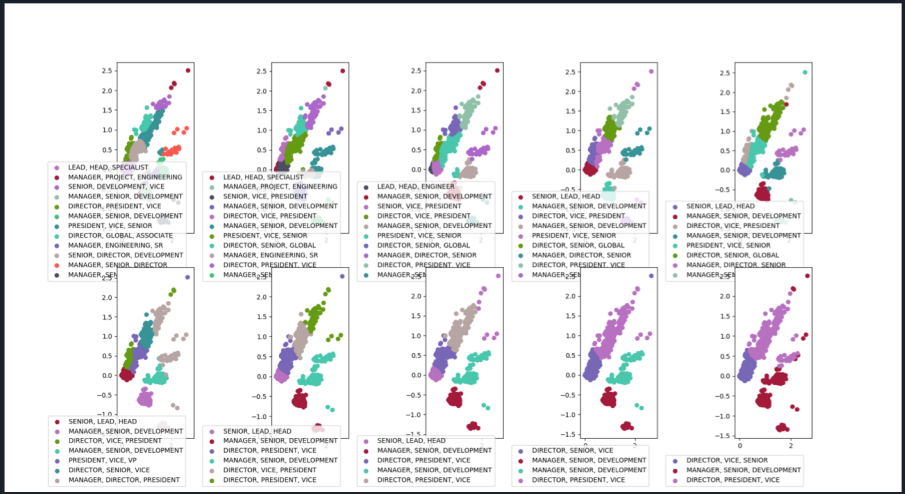
Clustering Job Titles

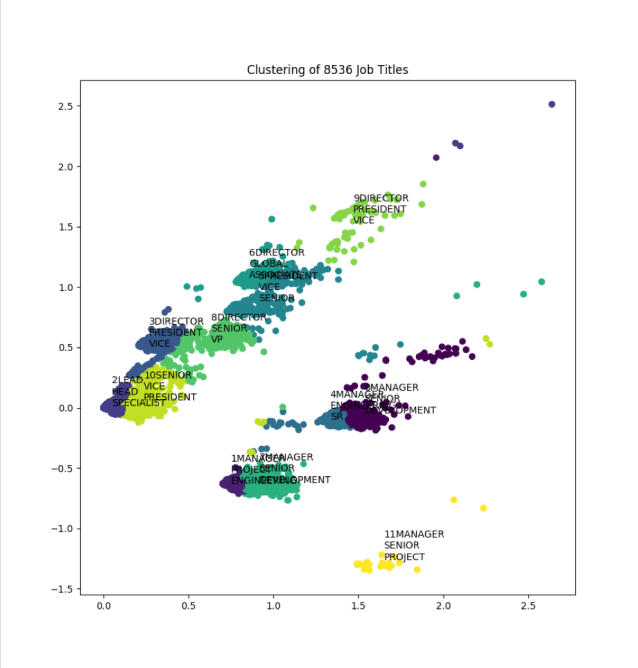
Reason:

1. 8,536 of different job titles, giving us no true information.
2. Overfitting problems in Decision Tree model.

Difficult:

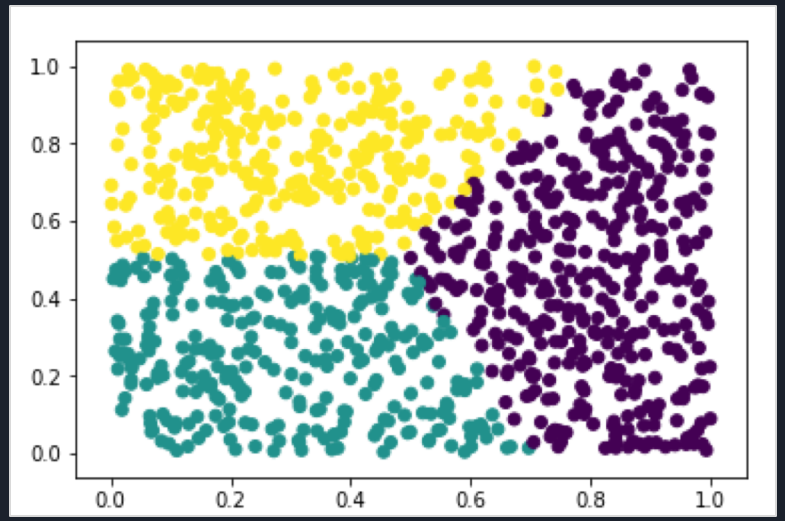
1. All the job titles were in strings

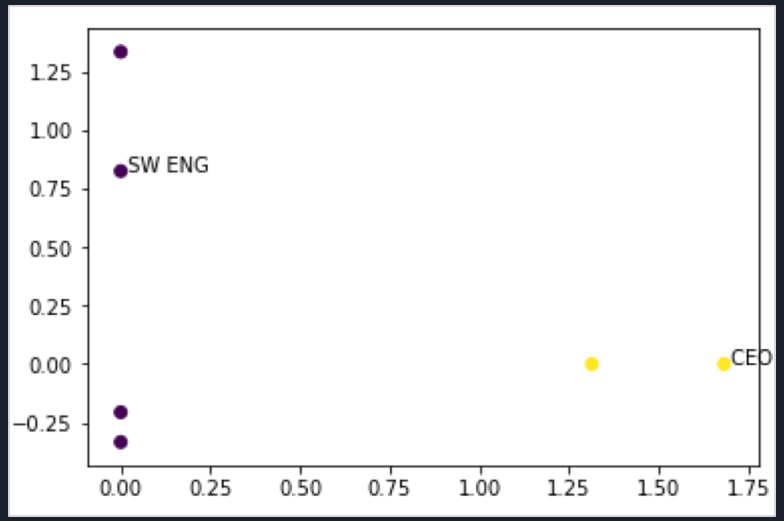




Methodology --Job Titles to clusters

1. A simple list of string is invalid
2. Vectorize the Job Title
3. Transform into sparse matrix of word counts
4. Structure is very high dimensional
5. Dimension reduction through SVD

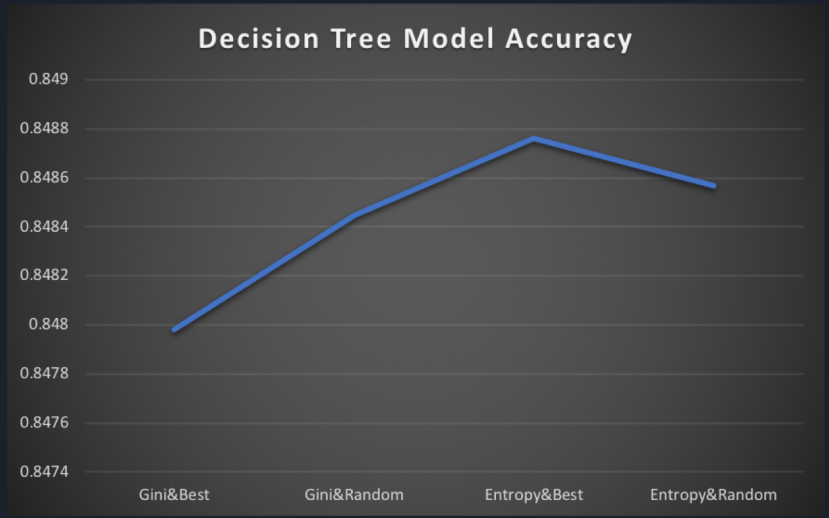




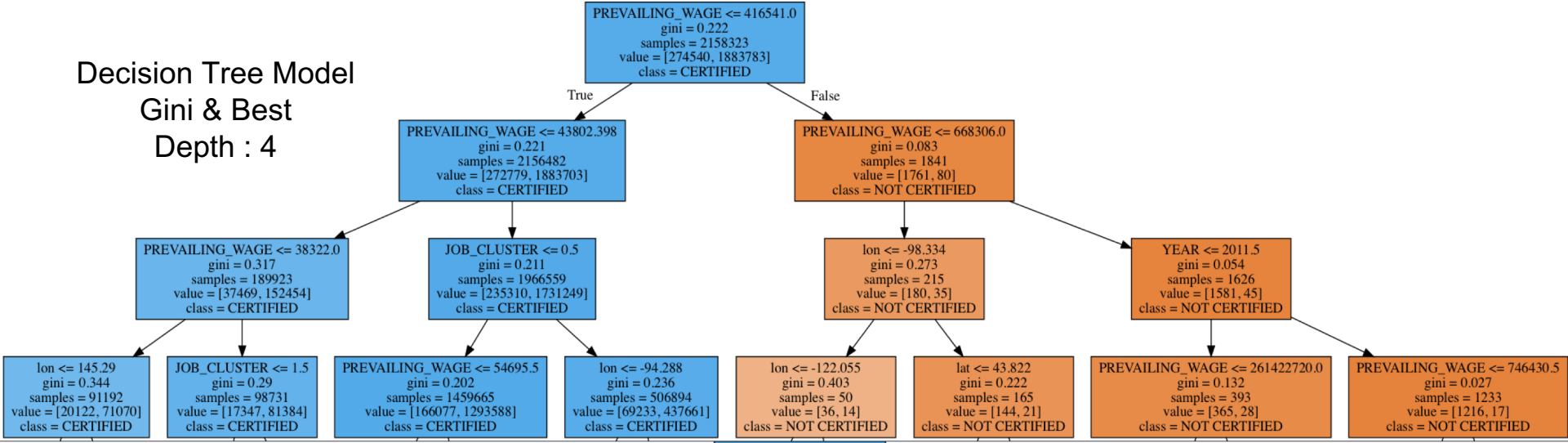
Decision Tree Analysis

1. Attribute: **Job Cluster (Ordinal)** / Full time(Y or N) / Wage(Numerical) / Year(Numerical) / Worksite(Numerical)
2. Label: Status (Yes/CERTIFIED or No/NOT CERTIFIED)
3. Parameter: (1) Gini/ Entropy (2)Best/ Random (3) Depth=4

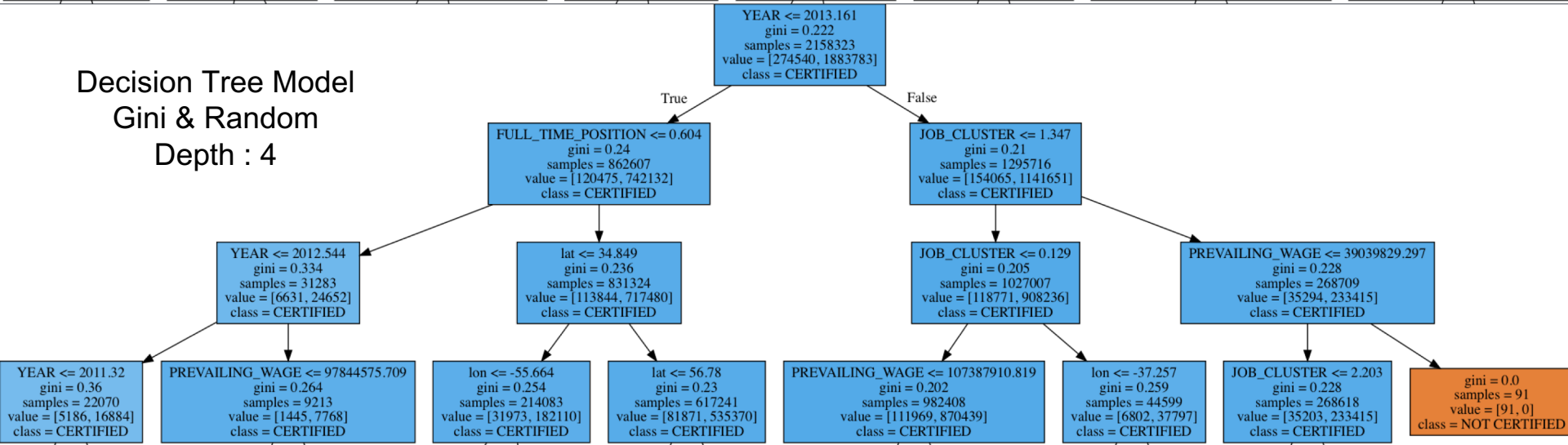
Best Feature: Wage
Worst Feature: Full Time Position



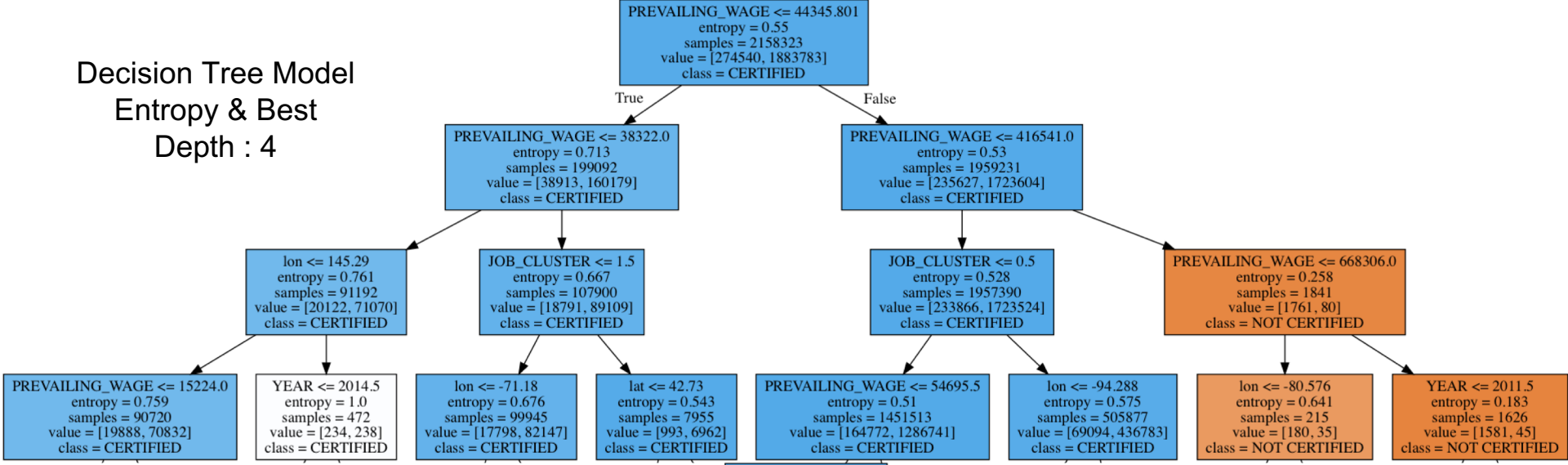
Decision Tree Model Gini & Best Depth : 4



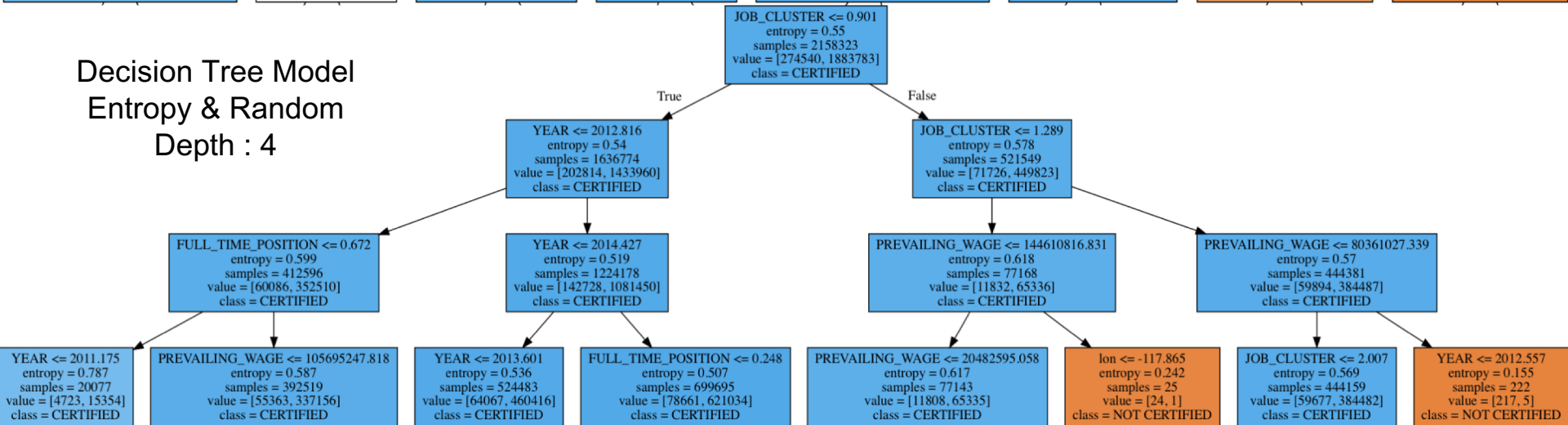
Decision Tree Model Gini & Random Depth : 4



Decision Tree Model
Entropy & Best
Depth : 4



Decision Tree Model
Entropy & Random
Depth : 4



Decision Tree Evaluation

Gini & Best	Predicted Approval	Predicted Denied
True Approval	585178	43153
True Denied	66215	24896
Acc: 84.7%	F1: 0.914	Pre: 0.898

Gini & Random	Predicted Approval	Predicted Denied
True Approval	585552	42779
True Denied	66255	24856
Acc: 84.8%	F1: 0.915	Pre: 0.898

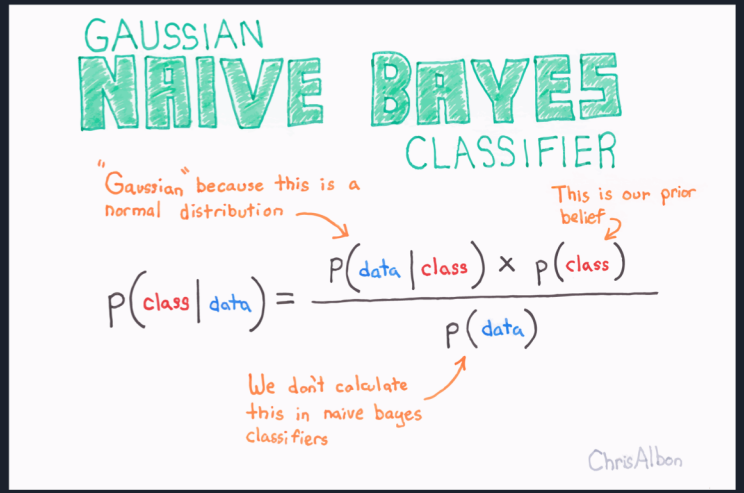
Best

Entropy & Best	**Predicted Approval**	**Predicted Denied**
True Approval	**585789**	**43542**
True Denied	**66266**	**24845**
Acc: 84.9%	**F1: 0.915**	**Pre: 0.898**

Entropy & Random	Predicted Approval	Predicted Denied
True Approval	585940	42391
True Denied	66555	24556
Acc: 84.7%	F1: 0.914	Pre: 0.898

Gaussian Naive Bayes

1. Attribute: **Job Cluster (Ordinal)** / Full time(Y or N) / Wage(Numerical) / Year(Numerical) / Worksite(Numerical)
2. Label: Status (Y or N)
3. Parameter: (1) Laplace Smoothing to avoid Zero Probability

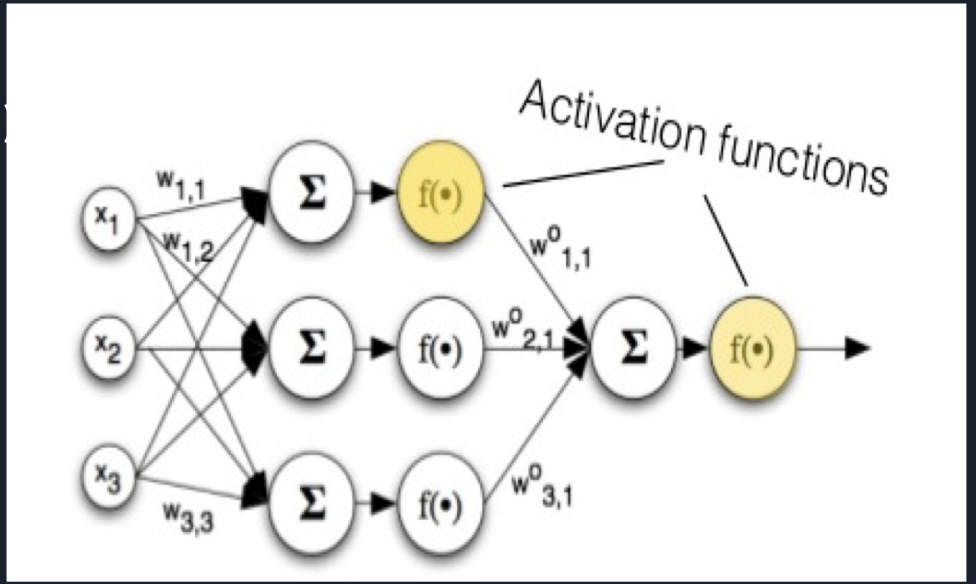


Evaluation	Predicted Approval	Predicted Denied
True Approval	627,845	13
True Denied	91,134	450
Acc: 87.33%	F1: 0.9333	Pre: 0.8732

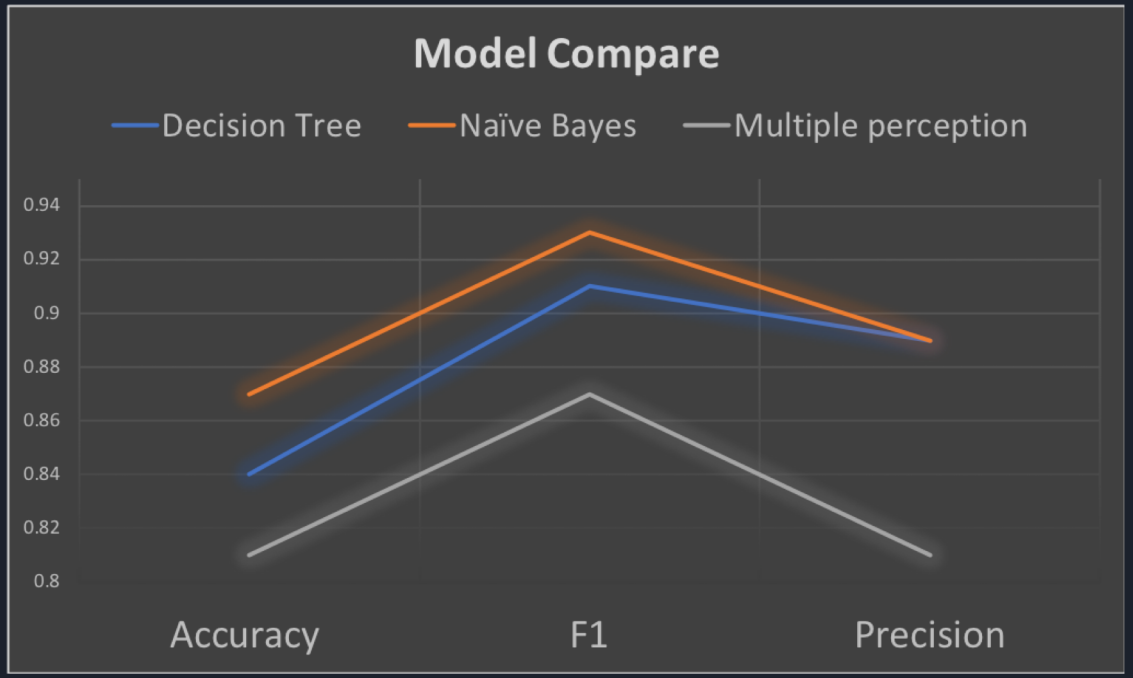
MLP Classifier

1. Attribute: **Job Cluster (Ordinal)** /
   Full time(Y or N) / Wage(Numerical)
   Year(Numerical) / Worksite(Numerical)
1. Label: Status (Y or N)
2. Parameter:

(1) hidden layer size
(2) activation (relu, identity, logistic, tanh)
(3) solver (adam, lbfgs, sgm)



Model Compare



Model	Time
DT	18.093s
NB	1.932s
MLP	46.78s

Conclusion

In this work, we proposed a series of process to cluster the Job Title, making it into an viable attribute. Then we used Job Clusters and other attributes building Decision Tree model, Naive Bayes model and MLP model to classify the approval status. The best model is Naive Bayes, which offers us a very high accuracy (87%). Therefore, that will be very useful to justify the status of H-1B visa of applicants.

Concluding Insights: what could make this better?

- More attributes
  - Where are these applicants from?
  - What's their education level?
  - How much work experience do they have?
- Considerations of American Foreign Policy
  - How have laws changed?
  - Should we give more/less weight to certain years that might be stricter/more lenient in acceptance? How would that be quantified?