categorizing fake and real jobs?

Instructor: Dr. Jay Qi
By Mohammad Hossein Movahedi

Introduction

- In this project, the question is, can we use job posting attributes to train a machine learning model to successfully categorize jobs into two categories of fake and real jobs?
- By comparing the overall performance of Naive buyers and SVM algorithms and comparing them to standard GLM the best algorithm will be found, and the question will be answered.

Exploratory Data Analysis

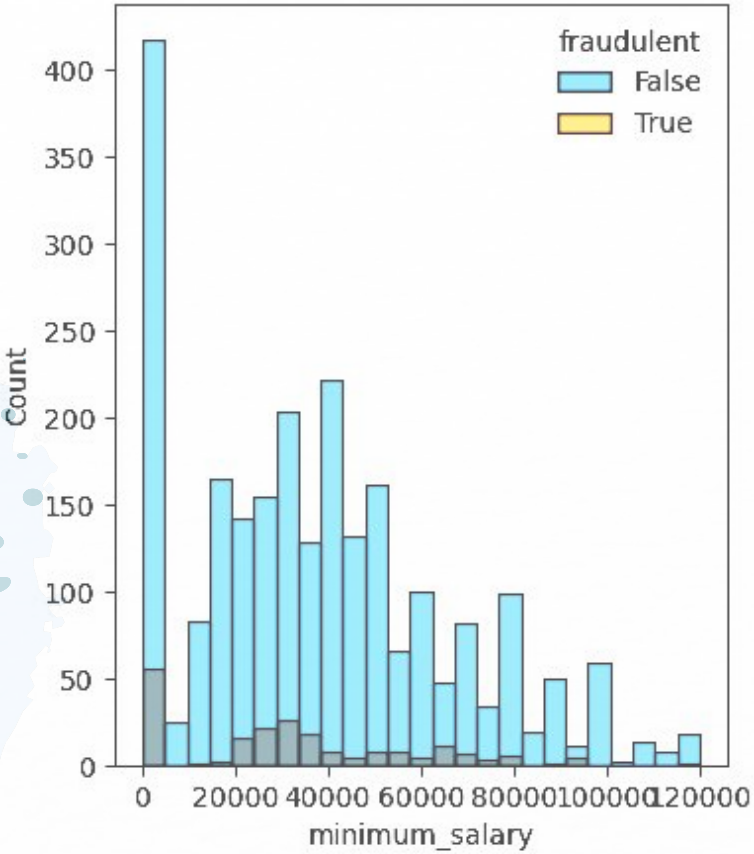
- The shape of the dataset is (17853, 20), and the final data types are listed below

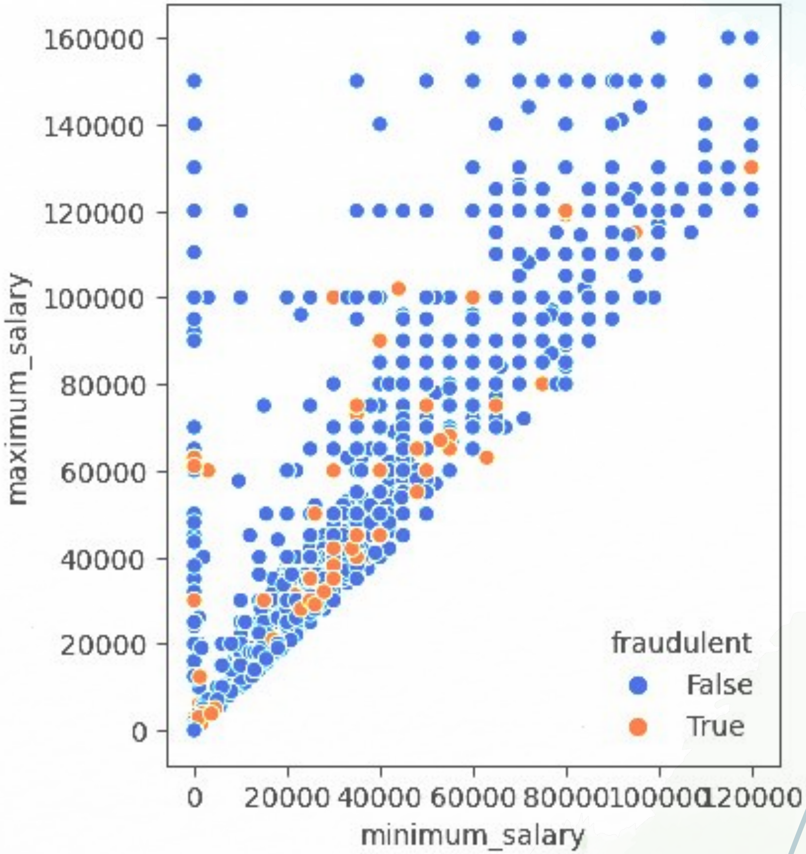
```
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   title                17853 non-null  object
 1   location             17509 non-null  object
 2   department           6330 non-null   object
 3   company_profile      17853 non-null  object
 4   description          17853 non-null  object
 5   requirements         17853 non-null  object
 6   benefits             17853 non-null  object
 7   telecommuting        17853 non-null  bool
 8   has_company_logo     17853 non-null  bool
 9   has_questions        17853 non-null  bool
 10  employment_type      17853 non-null  category
 11  required_experience  17853 non-null  category
 12  required_education   17853 non-null  category
 13  industry             17853 non-null  category
 14  function             17853 non-null  category
 15  fraudulent           17853 non-null  category
 16  minimum_salary       2841 non-null   float64
 17  maximum_salary       2841 non-null   float64
 18  country              17509 non-null  category
 19  keywords             17853 non-null  object
```

Data Extraction and Data Cleanup

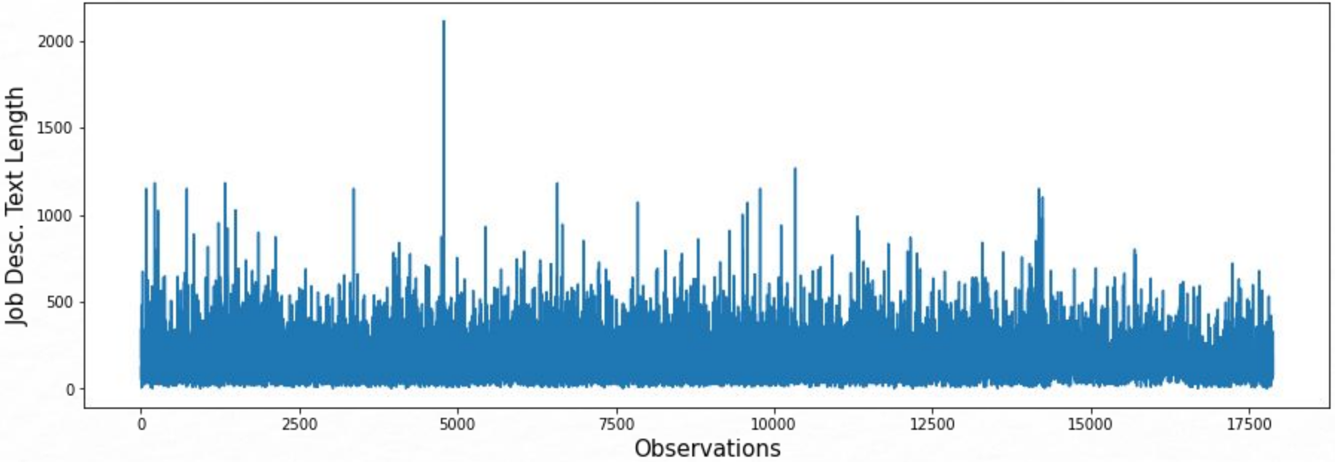
- The dataset is loaded directly from Kaggle to the python file, making it independent from the system. I used the open datasets library for this.
- Since I had so many NA values and the dataset mostly contained words and booleans and categories, the data cleanup part is mostly correcting the data type
- I also used the salary range to calculate the minimum and maximum salary.

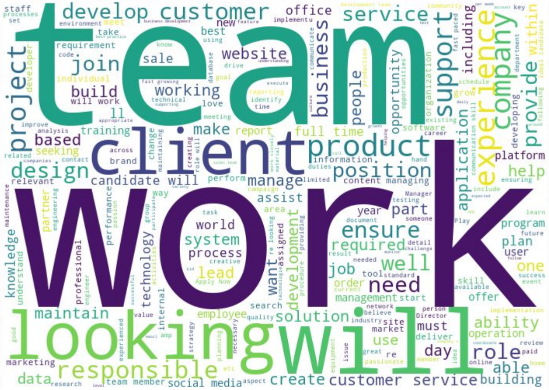
Data Visualization

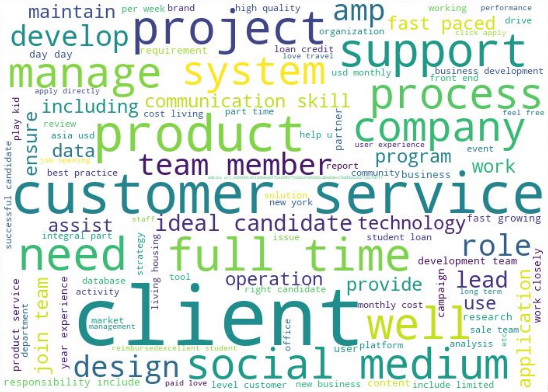




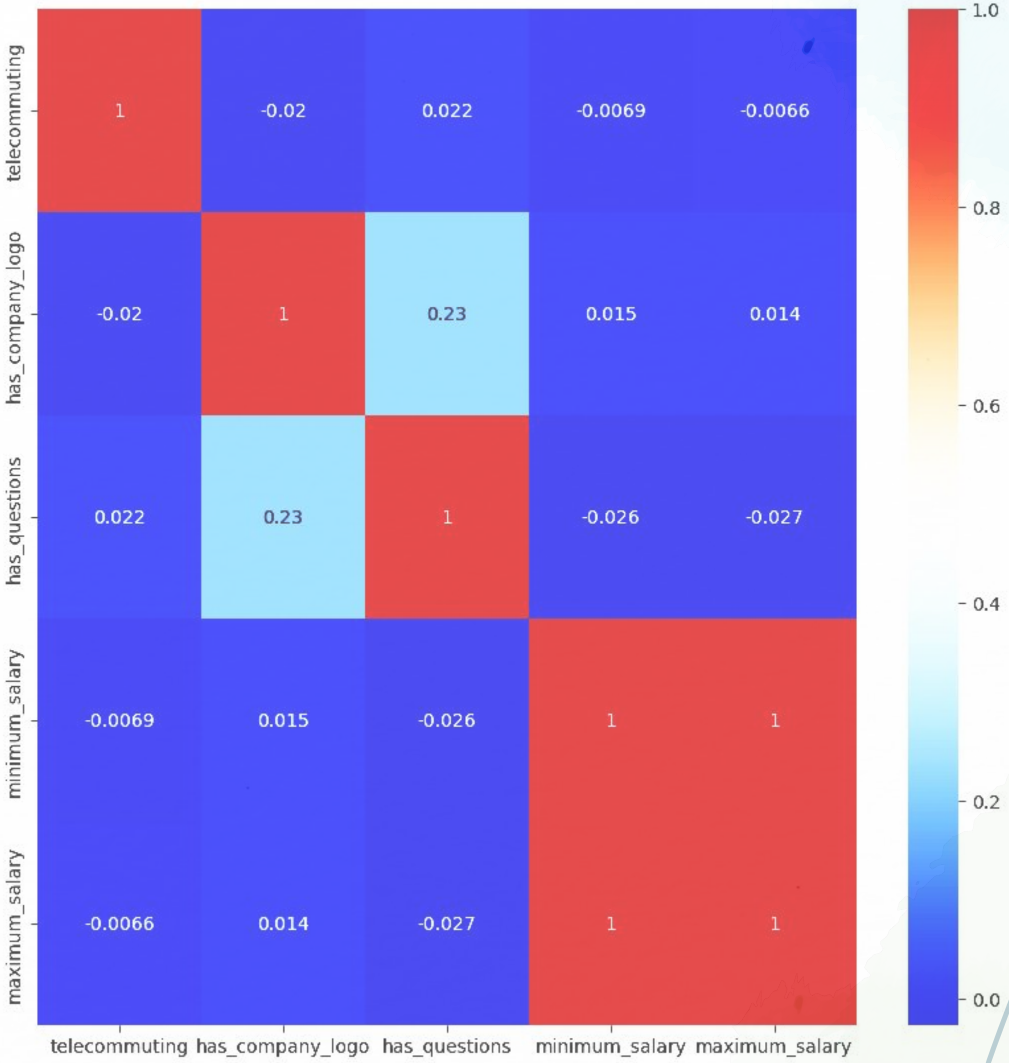
Data Visualization





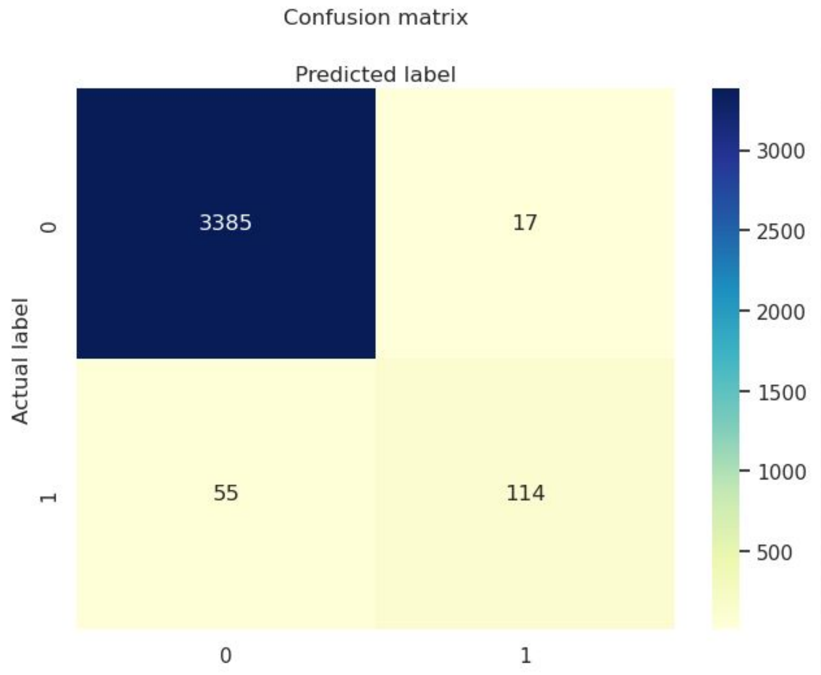


Data Visualization



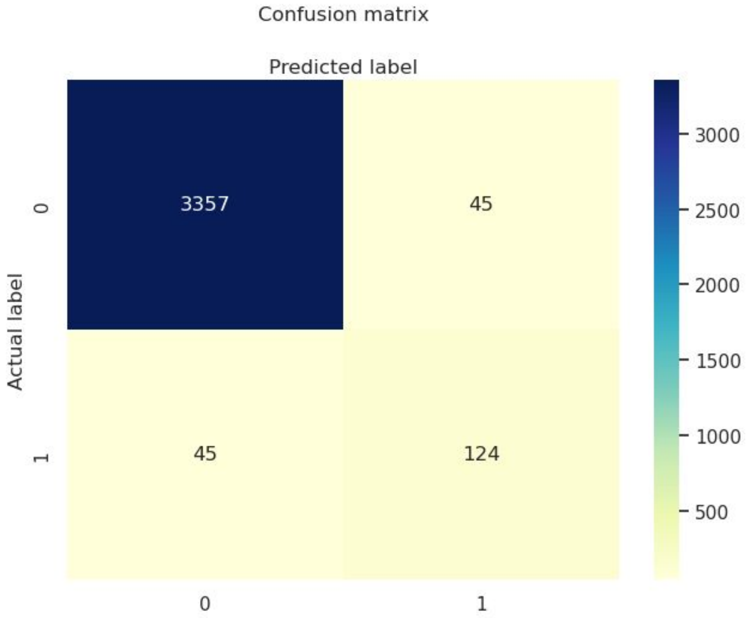
Predictive Models (Generalized linear model)

- As can be seen, it has done a great job distinguishing classes from each other.



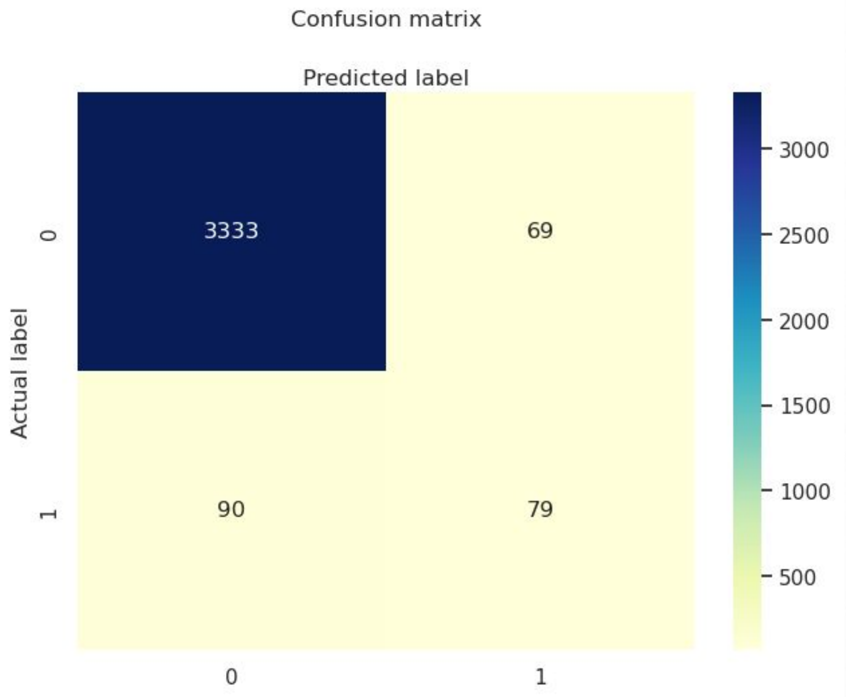
Predictive Models (Support vector machines)

- SVM aims to increase the distance between the data points and the hyperplane. To balance margin maximization and loss, we include a regularization parameter in the cost function. If the projected and actual values have the same sign, the loss function is 0

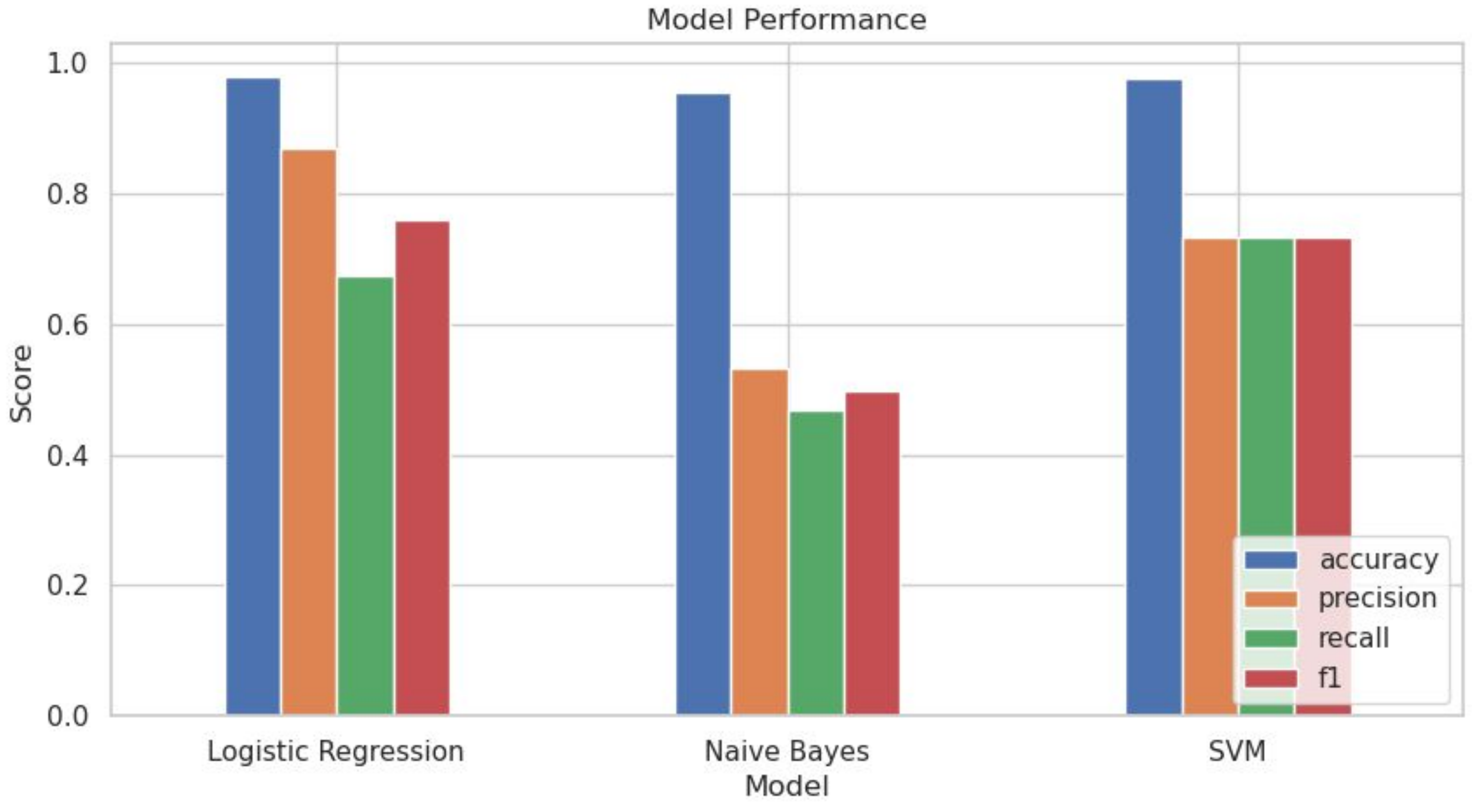


Predictive Models (Naive Bayes classifier)

- Bayes' Theorem is known as naive Bayes classifiers. It is a family of algorithms rather than a single method, and each character is individually important and relatively valuable. the inputs' probabilities for each potential value of the class variable y and choose the result with the highest likelihood.



Interpretive & Conclusions



References

Avijeet Biswal. (2021, November 9). *What is Exploratory Data Analysis? Steps and Market Analysis*. Simplilearn.com; Simplilearn. https://www.simplilearn.com/tutorials/data-analytics-tutorial/exploratory-data-analysis

Generalized Linear Models in R. (2021). Wisc.edu. https://sscc.wisc.edu/sscc/pubs/glm-r/

IBM Cloud Education. (2021, February 10). *What is Data Visualization?* Ibm.com. https://www.ibm.com/cloud/learn/data-visualization

Naive Bayes Classifiers - GeeksforGeeks. (2017, March 3). GeeksforGeeks. https://www.geeksforgeeks.org/naive-bayes-classifiers/

Patil, P. (2018, March 23). *What is Exploratory Data Analysis? - Towards Data Science*. Medium; Towards Data Science. https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15

Shopify. (2021, April 28). *A Five-Step Guide for Conducting Exploratory Data Analysis*. Shopify. https://shopify.engineering/conducting-exploratory-data-analysis