ALY 6140 Module 4 — Capstone Project Draft Report

Student's name: Mohammad Hossein Movahedi

Assignment title: ALY 6140 Module 4 — Capstone Project Draft Report

Course number and title: ALY6140 71379 Analytics Systems Technology SEC 12 Fall 2022 CPS [TOR-B-HY]

Term: 202315_1 Fall 2022 CPS Quarter

Instructor's name: Jay Qi, Ph.D.

Nov 23, 2022,

Introduction

The goal of this project is to find an optimized algorithm that can distinguish fake jobs from legit jobs. In this project, the question is can we use attributes of the job posting to training a machine learning model to successfully categorize jobs into two categories of fake and real jobs?

Based on the goal and the question of the project concerning my knowledge of machine learning algorithms from previous works. The following algorithms are used for this project:

1. Naive buyers
2. SVM
3. GLM

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

In this part, I briefly describe the EDA part of the project at each level. Exploratory data analysis is the crucial process of doing preliminary analyses of data to find patterns, identify anomalies, test hypotheses, and double-check assumptions with the aid of summary statistics and graphical representations (Patil, 2018).

Analysis Description

Determining the number of samples (rows) and features (columns) in the dataset. The amount of data provides information about potential computing bottlenecks. A correlation matrix calculation, for instance, can take a long time when applied to huge datasets. Subsampling can be useful if the dataset is too large to work within a Jupyter notebook but still adequately depicts the data(Shopify, 2021).

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

Data columns (total 20 columns):

#	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

The gathering of data is a crucial step in exploratory data analysis. It speaks of the method used to locate and load data into our system. You can purchase trustworthy information from private companies or find it on a variety of public websites. Websites like Kaggle, Github, the Machine Learning Repository, etc. are trustworthy sources for data acquisition (Avijeet Biswal, 2021).

The dataset is loaded directly from Kaggle to the python file making it independent from the system. I used opendatasets library for this.

Data Cleanup

The act of purging your dataset of any extraneous variables and values as well as any errors is known as data cleansing. To clean data, the following actions can be taken: removing incorrect rows and columns, outliers, and missing values. reformatting and re-indexing the data (Avijeet Biswal, 2021).

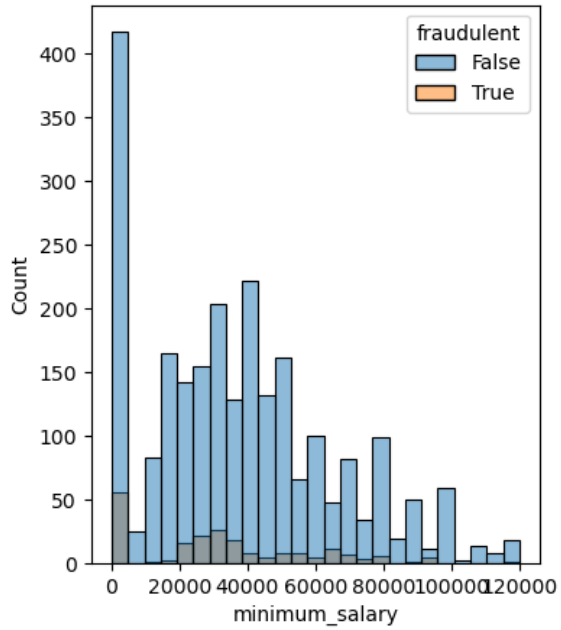
There are numerous ways for data to be duplicated or incorrectly categorized when merging multiple data sources. Correcting or erasing inaccurate, corrupted, improperly formatted, duplicate, or insufficient data is known as data cleaning. Due to the fact that procedures differ from dataset to dataset, there is no one definite way to specify the precise phases in the data cleaning process (Guide to Data Cleaning: Definition, Benefits, Components, and How to Clean Your Data, 2022).

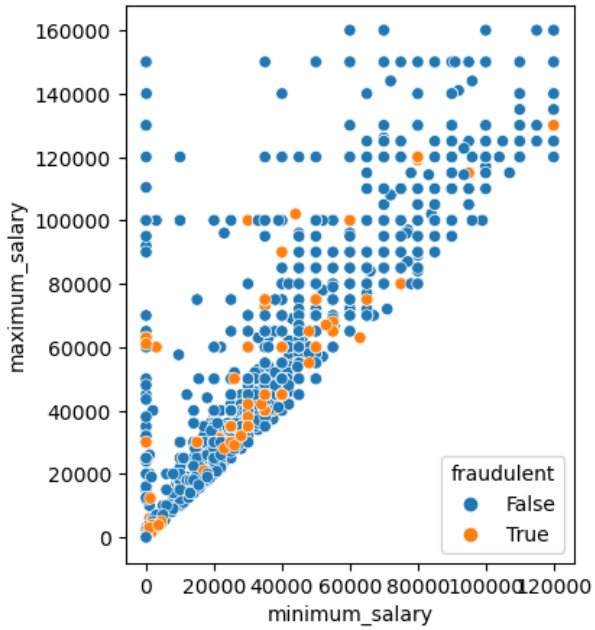
Since I had so many NA values and the dataset mostly contains words and booleans and categories the data cleanup part is mostly correcting the data type, I also used the range of salary to calculate the minimum and maximum salary.

Data Visualization

The depiction of data through the use of typical graphics, such as infographics, charts, and even animations, is known as data visualization. Concept creation, idea illustration, visual discovery, and data visualization are the four main categories used by Harvard Business Review to classify data visualization. After a fresh insight has been discovered, data visualization helps with the subsequent storytelling. When text mining unstructured data, an analyst may use a word cloud to identify important ideas, patterns, and undiscovered connections. As an alternative, they can show the connections between things in a knowledge graph using a graph structure (IBM Cloud Education, 2021).

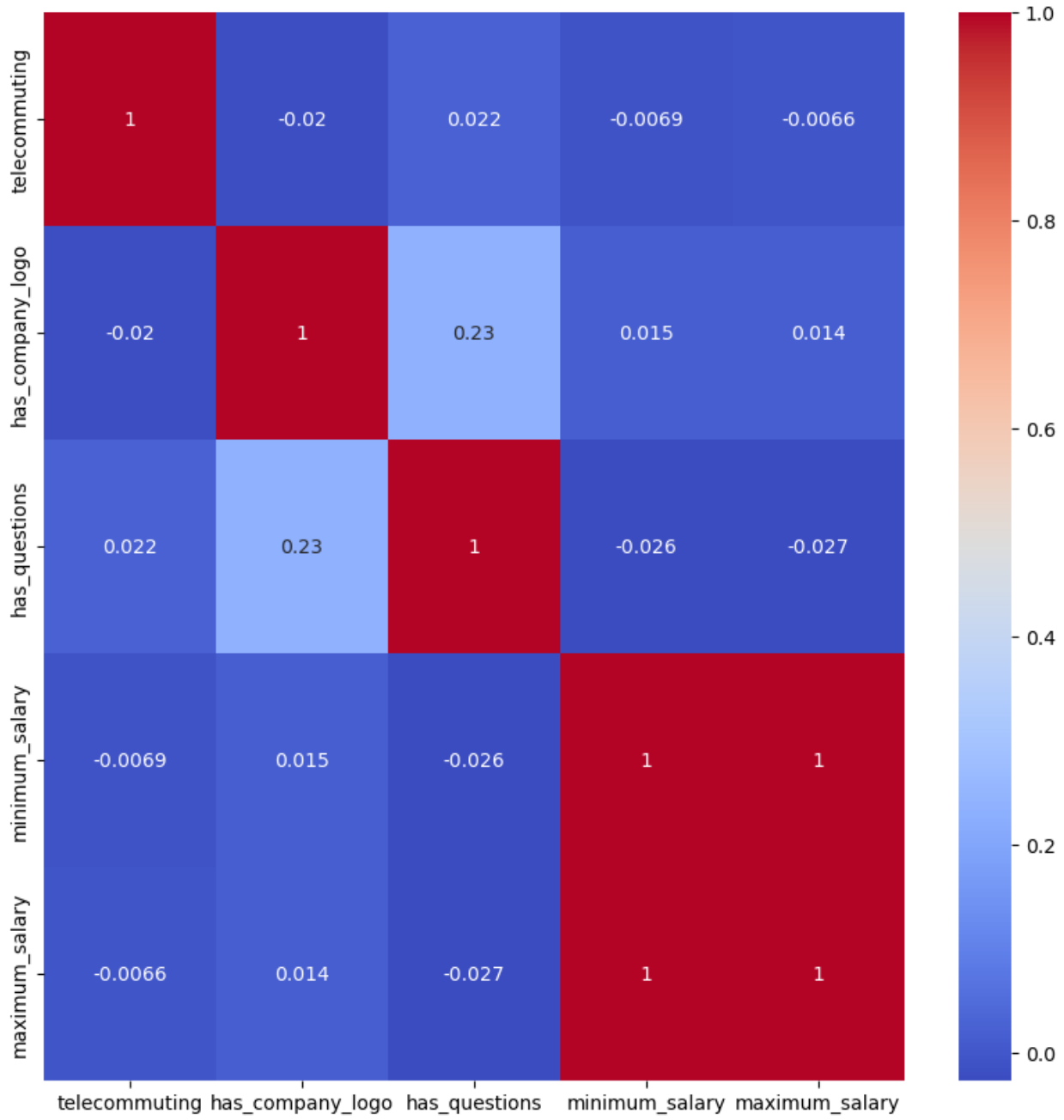
The chart below shows the minimum and maximum salary in scatter plot and also in histogram comparing them to each other





As can be seen there is no visible trend in the fraudulent job in terms of minimum and maximum salary.

The chart below shows the correlation between the 10 main variables of the dataset.



As can be seen except for minimum and maximum salary all other variables are almost independent of each other .this assure us that we can use naive buyers as it doesn't check for singularity.

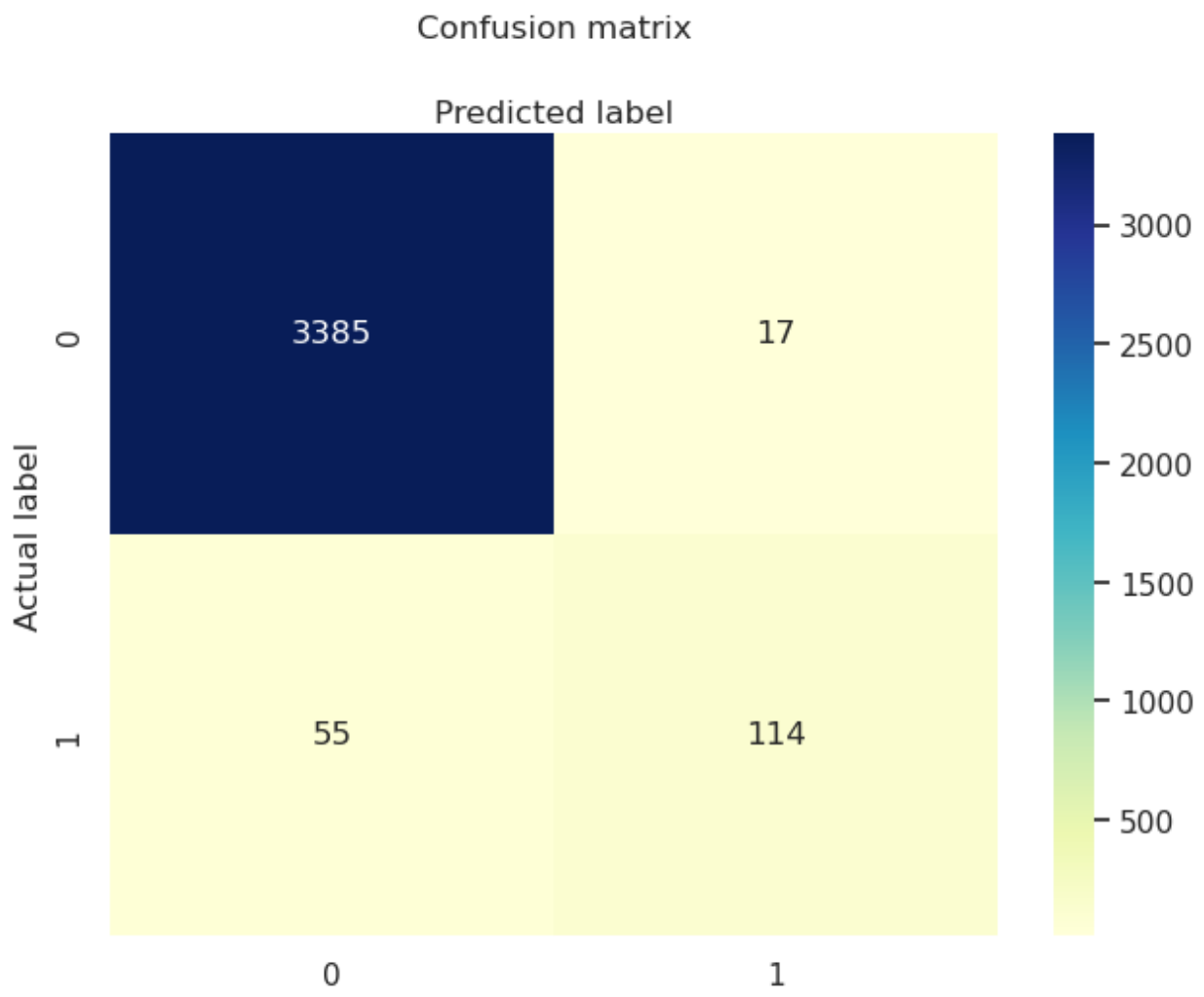
Predictive Models

As stated in the introduction section of the project Naive buyers, SVM, and GLM models are used in this project. The final result and findings of each model are shown in each subsection below.

Generalized linear model

GLM models change the response variable to allow least squares fitting. The variance's connection to the mean is described by the link function. Additionally, GLM models may be used to fit data when the variance is proportional to a known variance function. For some GLM models and considerably less so for others, residual graphs are beneficial. The variance is used to normalize Pearson residuals, which are anticipated to remain constant over the entire prediction range. Wald test of coefficients is not favored over nested model tests for a coefficient's significance(Generalized Linear Models in R, 2021).

The plot below shows the confusion matrix for GLM

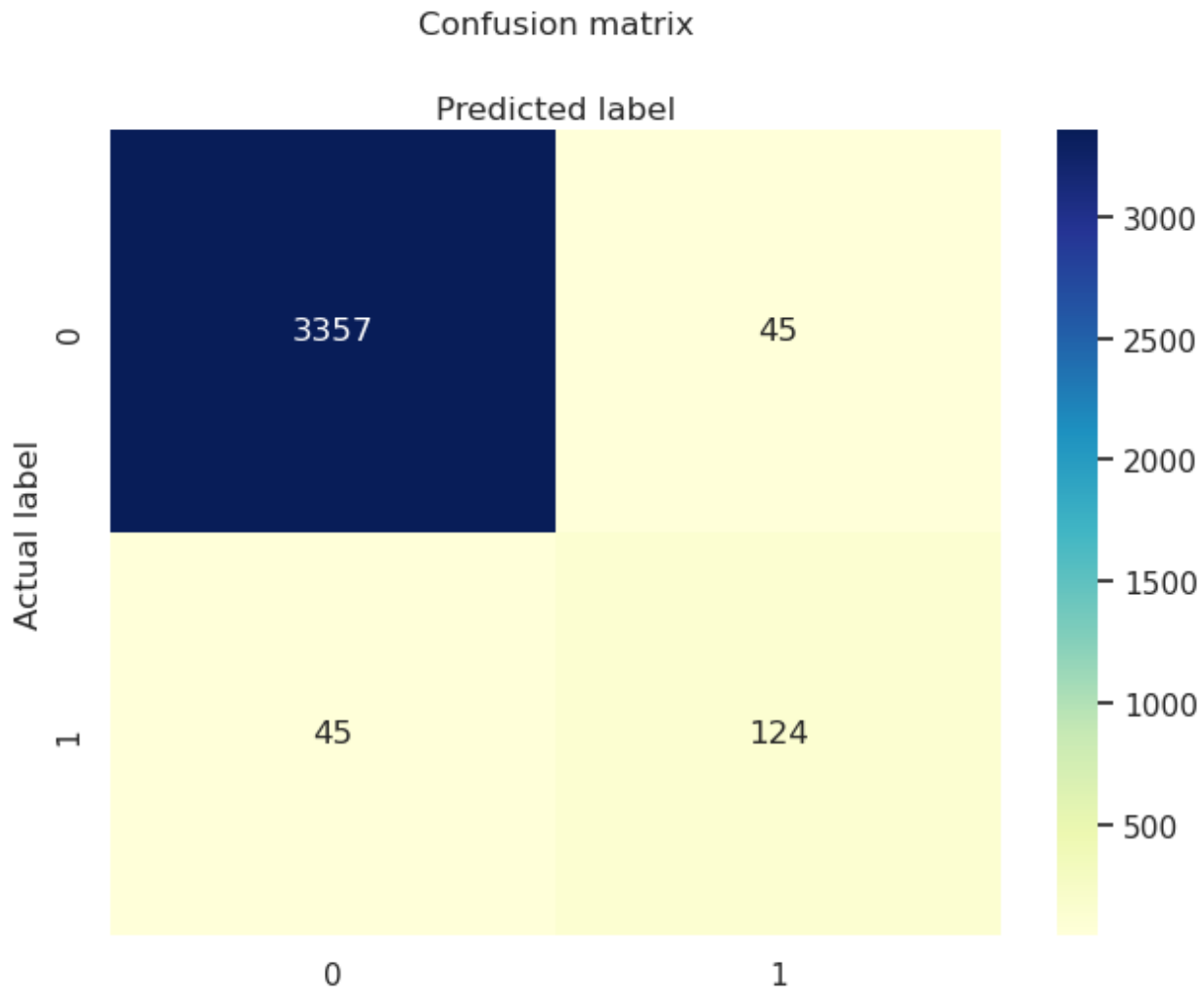


As it can be seen it has done a great job distinguishing classes from each other.

Support vector machines

Finding a hyperplane in N-dimensional space (N is the number of features) that categorizes the data points clearly is the goal of the support vector machine algorithm. Different classes can be given to the data points that fall on each side of the hyperplane. The goal of SVM is to increase the distance between the data points and the hyperplane. To balance margin maximization and loss, we additionally include a regularization parameter in the cost function. If the projected value and the actual value have the same sign, the loss function is 0 (Gandhi, 2018).

The plot below shows the final results of the SVM method



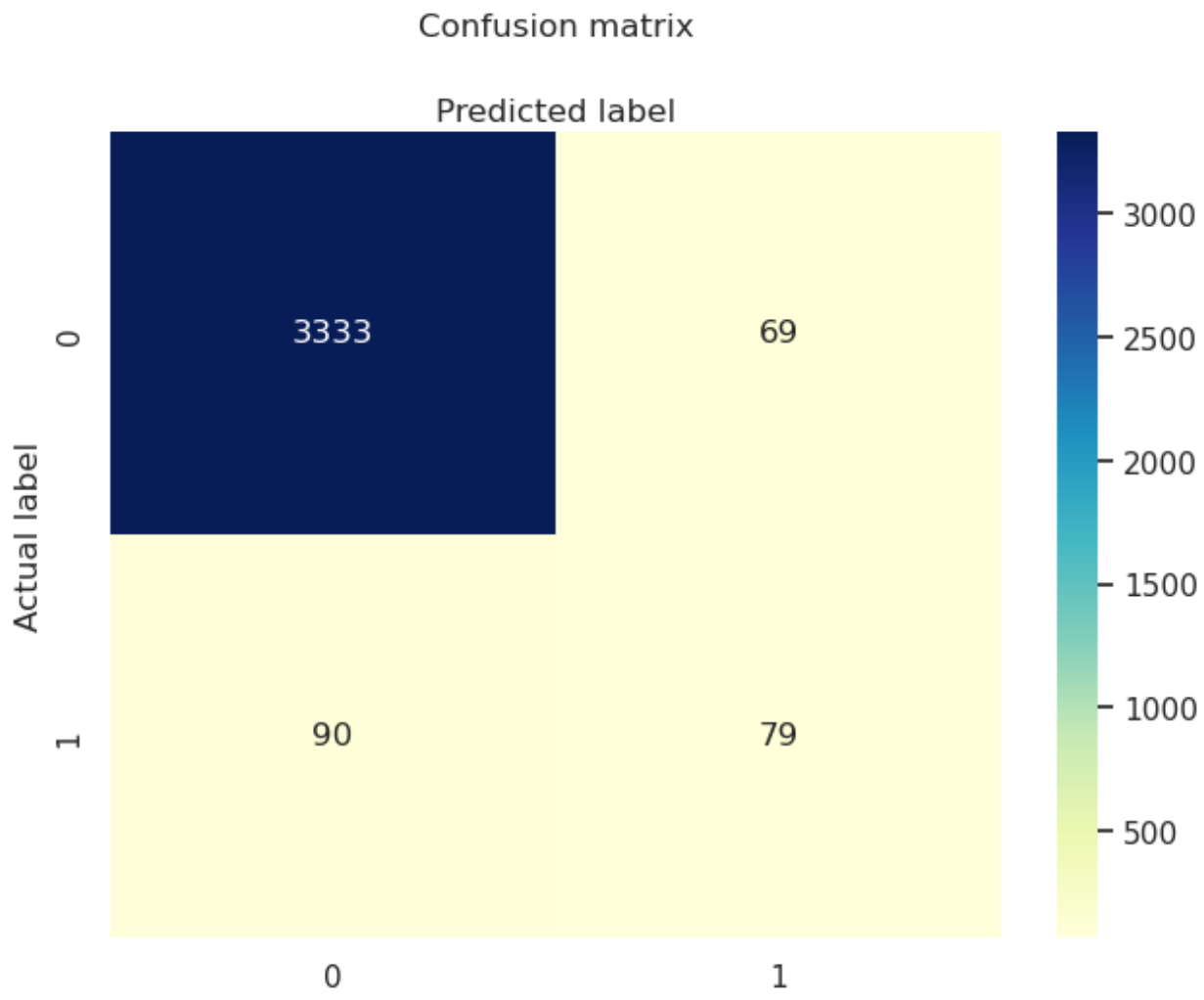
As it can be seen it also did a great job

Naive Bayes classifier

A group of classification algorithms built on the 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.

The plot below shows the confusion matrix for the naive buyer model



The results aren’t as impressive as GLM but the algorithm is considerably faster than the other two algorithms.

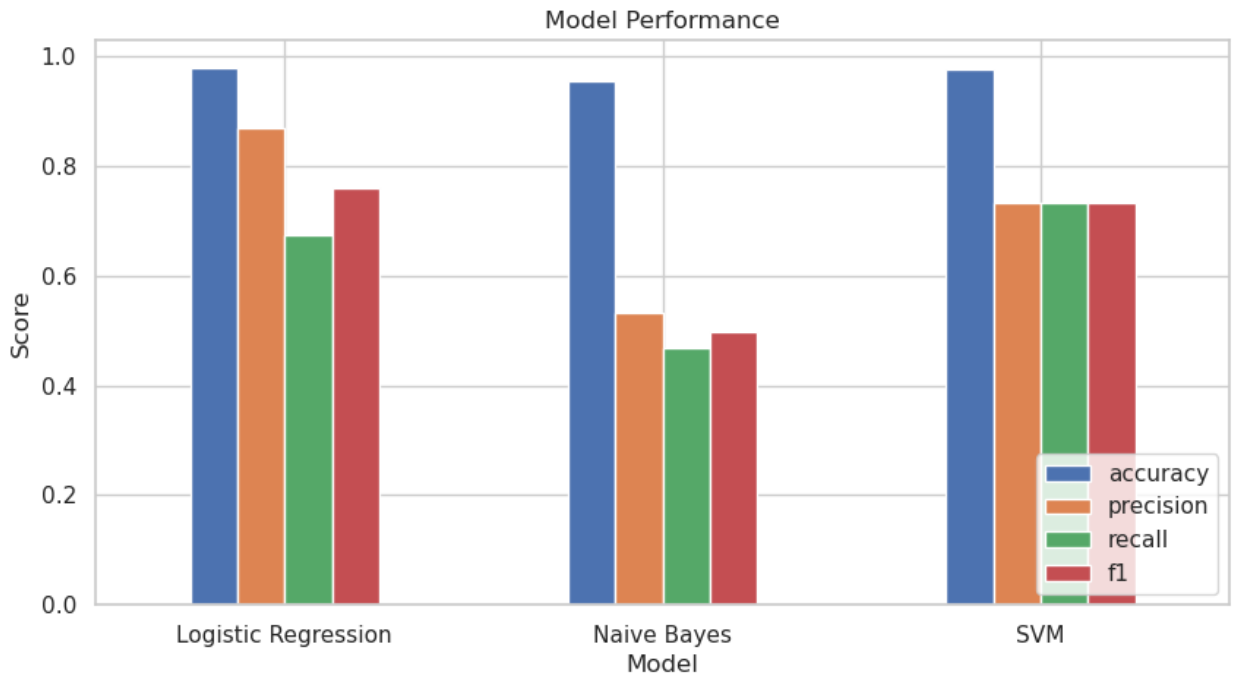
In many real-world contexts, including the infamous document categorization and spam filtering, naive Bayes classifiers have performed admirably. For them to estimate the required parameters, only a modest amount of training data is needed. Each model can be separately estimated as a one-dimensional distribution due to the separation of the class conditional feature distributions (Naive Bayes Classifiers - GeeksforGeeks, 2017).

Interpretive & Conclusions

The objective of this project is to identify an improved algorithm that can discriminate between legitimate jobs and bogus ones. For this project, the following algorithms are employed:. The vital process of performing the first analyses of data in order to uncover patterns and anomalies is known as exploratory data analysis.

The project uses the Naive Buyers, SVM, and GLM models as described in the introductory section. Each subsection displays the conclusion and results of each model. The response variable is modified in GLM models to enable least squares fitting. Additionally, they employ residual graphs to normalize Pearson residuals, which should hold steady throughout the full prediction range.

The plot below shows the accuracy, precision, recall , and F1 of three models in comparison



As it can be seen the overall performance of SVM is better than the other two, however, the Logistic regression has higher precision than the other two its recall score is second to SVM.

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