Springboard - DSC Capstone Project II

Job Change of Data Scientist

Final Report

Akbanu Tleubayeva

Data Science Career Track

December 2021

Table of Contents

1. Introduction

A company which is active in Big Data and Data Science wants to hire data scientists among people who successfully pass some courses which are conducted by the company. Many people sign up for their training. Company wants to know which of these candidates really wants to work for the company after training or looking for a new employment because it helps to reduce the cost and time as well as the quality of training or planning the courses and categorization of candidates. Information related to demographics, education, experience are in hands from candidates signup and enrollment.

This dataset is designed to understand the factors that lead a person to leave their current job for HR research too. By model(s) that uses the current credentials,demographics,experience data you will predict the probability of a candidate to look for a new job or will work for the company, as well as interpreting affected factors on employee decision.

1.1 Objective

The objectives of this project are to:

• Explore a dataset compiled to understand which factors lead a person to leave their current job as a data scientist.
• Identify the key features that lead enrollees to look for new employment.
• Develop machine learning models that predict the probability of enrollees looking for new jobs
• Identify the final model that captures the most target enrollees within the top 20% and top 50% of the test dataset in descending order by their prediction scores

This report is divided into the following sections:
• Section 2: Dataset
• Section 3: Package Introduction

• Section 4: Data Wrangling
• Section 5: Exploratory Data Analysis
• Section 6: Machine Learning
• Section 7: Final Model Selection and Hyperparameters Tuning
• Section 8: Conclusion
• The programming codes used for this report can be found in this Github Repository.
https://github.com/akisd2020/Capstone2_Job_Change_Of_Data_Scientist

1.2 Significance

By thoroughly exploring the dataset, we will identify the important features that affect the enrollee's decision of career change. We will also develop machine learning models that can be used by the recruiting team of a company to filter out the potential candidates from the user database and approach them with better efficiency and accuracy.

2. Dataset

2.1 Data Description

The datasets are sourced from the website kaggle, a subsidiary of Google LLC, is an online community of data scientists and machine learning practitioners that allows users to find and publish datasets, explore and build models in a web-based data-science environment.

The dataset aug_train.csv used in this project was collected in August, 2020. This data set contains 1**9,158 observations,** 14 columns. i.e 13 features + 1 label.

10 categorical variables: city, gender, relevent_experience, enrolled_university, education_level, major_discipline, experience, company_size,last_new_job,company_type

3 numerical variables: enrollee_id, city_development_index,training_hours

1 label variable - target

Each row contains credentials/demographics/experience data for each unique enrollee. A description of each of the columns is provided in Table 1.1

	enrollee_id	city	city_development_index	gender	relevent_experience	enrolled_university	education_level	major_discipline	experience	company_size	company_type	last_ne
0	8949	city_103	0.920	Male	Has relevent experience	no_enrollment	Graduate	STEM	>20	NaN	NaN	
1	29725	city_40	0.776	Male	No relevent experience	no_enrollment	Graduate	STEM	15	50-99	Pvt Ltd	
2	11561	city_21	0.624	NaN	No relevent experience	Full time course	Graduate	STEM	5	NaN	NaN	
3	33241	city_115	0.789	NaN	No relevent experience	NaN	Graduate	Business Degree	<1	NaN	Pvt Ltd	
4	666	city_162	0.767	Male	Has relevent experience	no_enrollment	Masters	STEM	>20	50-99	Funded Startup	

2.2 Dataset Characteristic

Dealing with missing values.

This data contains numerical and categorical columns. i.e in some columns the order of the categories is significant and in others there is no meaning to the order or the quantity of the value.

The numerical columns in our data are : relevent_experience, enrolled_university, education_level, experience, company_size,last_new_job,

The categorical columns are: city, gender, major_discipline, company_type.

Table 2.1

	null_numm	null_percent
gender	4508	0.240000
enrolled_university	386	0.020000
education_level	460	0.020000
major_discipline	2813	0.150000
experience	65	0.000000
company_size	5938	0.310000
company_type	6140	0.320000
last_new_job	423	0.020000

1. The company type has the largest missing value, with a missing percentage of 32%.
2. The company size has the largest missing value, with a missing ratio of 31%
3. The same conclusions for the test set - most of the gender columns are missing, as well as few values in the enrolled_university, education_level, experience and last_new_job columns.

3. Package Introduction

In this study we used JupyterLab (2.1.5) to run all the code. Numpy (1.18.5), Pandas (1.0.5), Matplotlib (3.2.2), Seaborn (0.11.0) were installed as the basic package. Scikit-learn (0.23.1) was installed as the machine learning library. Imblearn (0.7.0) was installed for data oversampling. Scikit Plot (0.3.7) was installed for plotting Cumulative Lift.

4. Data Wrangling

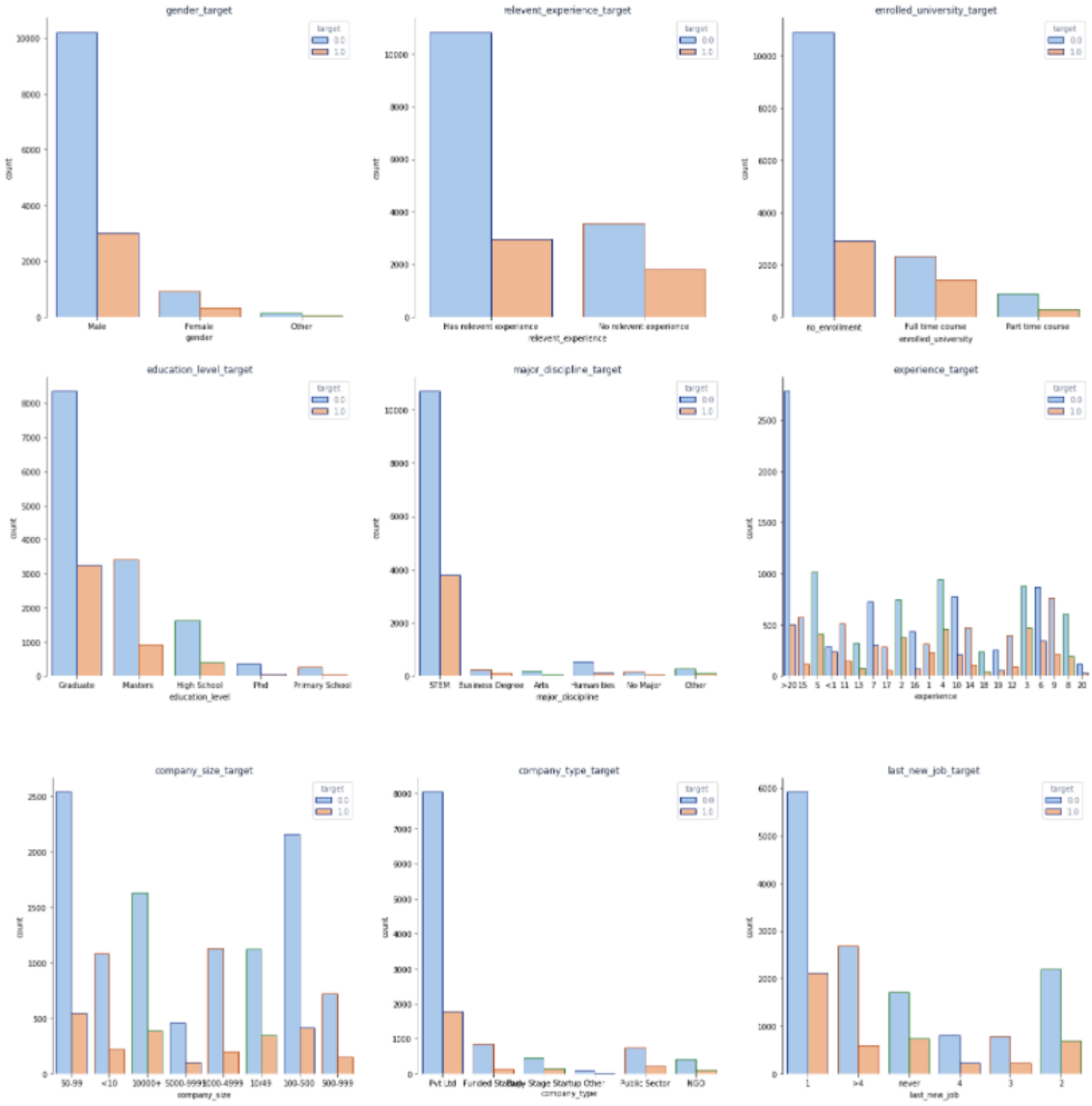
4.1 Dataset Information

First, I visualized the relationships between each variable to the other variables in the data. Then I used methods like head(), describe(), info() ect. to display the first five rows of each data set, display basic statistical details and general information about the data.

Afterwards I visualized the missing data in order to decide which columns to drop and how to deal with the missing values.

4.2 Data Processing

Table 4.1



1. According to the first picture, we can see that the proportion of men who are not looking for a job change far exceeds the proportion of men who are looking for a job change.

2. In the second picture, we can see that most candidates with relevant experience do not look for job changes in a large proportion.

3. In the types of registered courses, most people are not registered for courses and are not willing to look for job changes.

4. Most of these groups have a high degree of education.

5. The candidate's major is basically STEM. This shows that many people are not changing industries.

6. In the group that does not change their jobs, many people have more than 20 years of work experience. This can actually explain in disguise that the longer you work, the more you hope you can stabilize. In contrast, those with less work experience will have a significantly higher rate of changing jobs.

7. Among the groups that do not plan to change jobs, the number of their employer companies is basically between 50-500.

8. Among the groups that do not change their jobs, the type of employer is basically pv ltd

9. The proportion of unchanged jobs exceeds the proportion of changed jobs, and employees who have just joined the company for about a year are less willing to change jobs.

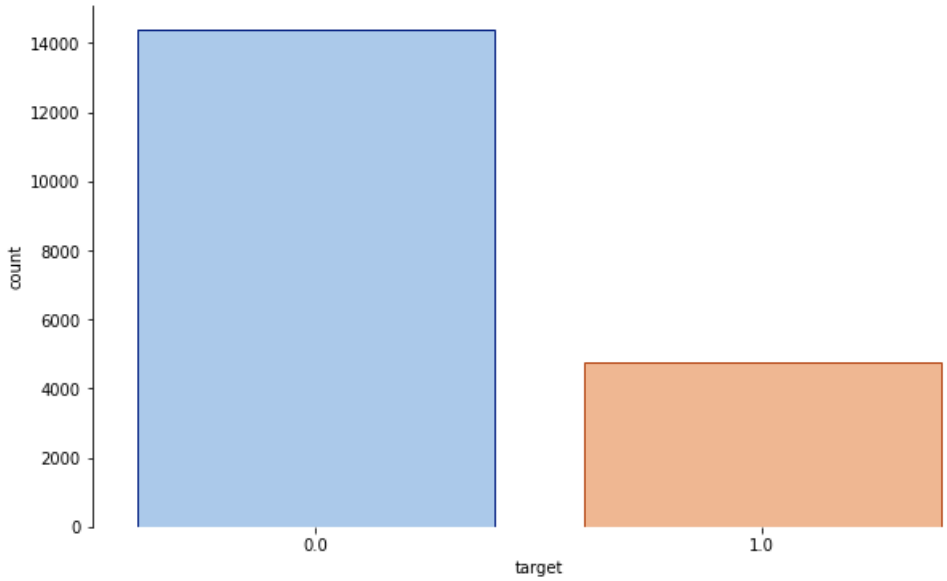
5. Exploratory Data Analysis

5.1 Summary Statistic

The statistics including mean, standard deviation, minimum values, maximum values and percentile for each variable were summarized.

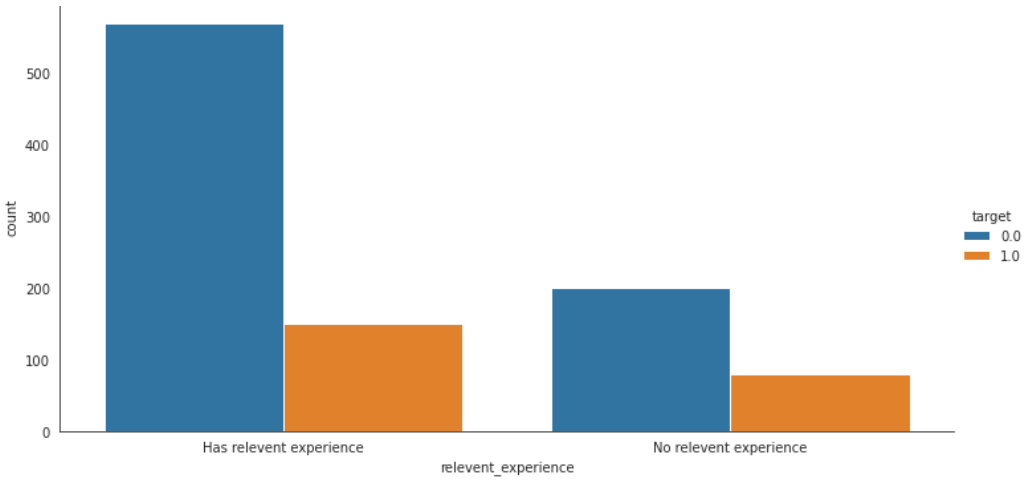
5.2 Overall Distribution

The overall distribution of numerical variables was visualized. Unique values and their counts breakdown by non-Target and Target enrollee for each categorical variable were visualized.



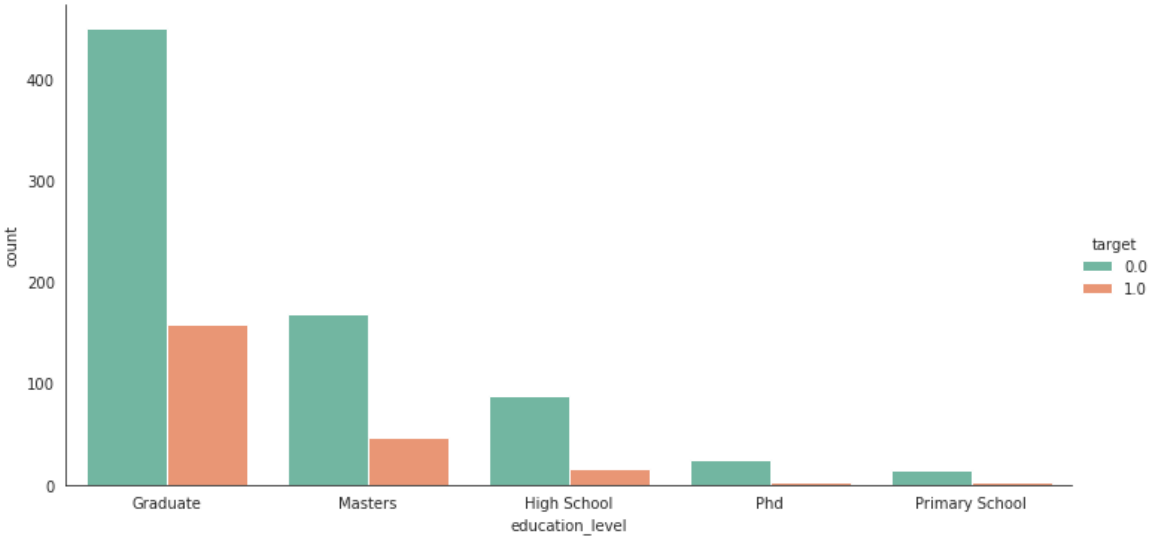
We can see that the number of people who do not plan to change jobs is the largest, and the data is indeed unbalanced.

5.3 Experience vs. Target



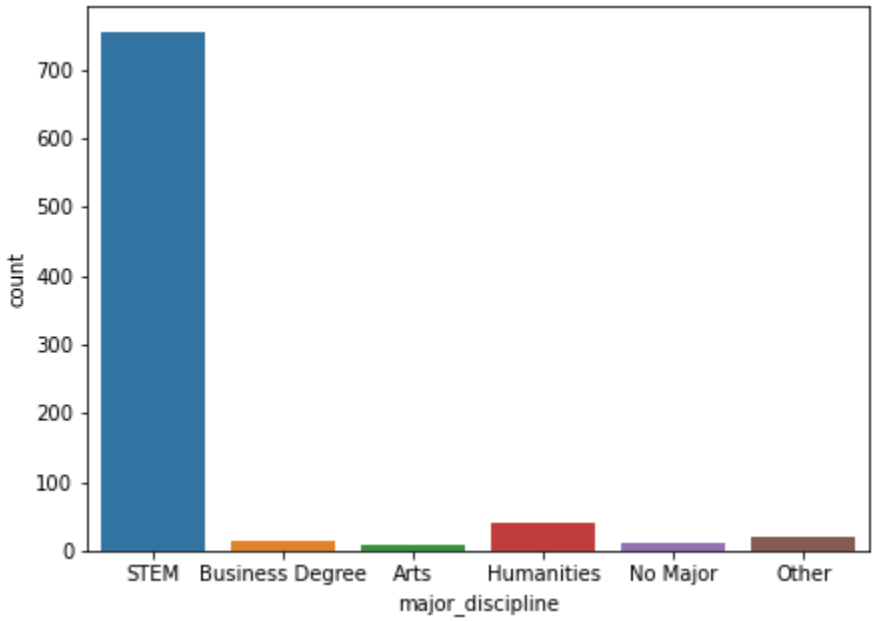
According to the analysis, those who are having less experience are more vulnerable to switch their job, rather than those who are having more experience.

5.4 Level of Education vs. Target



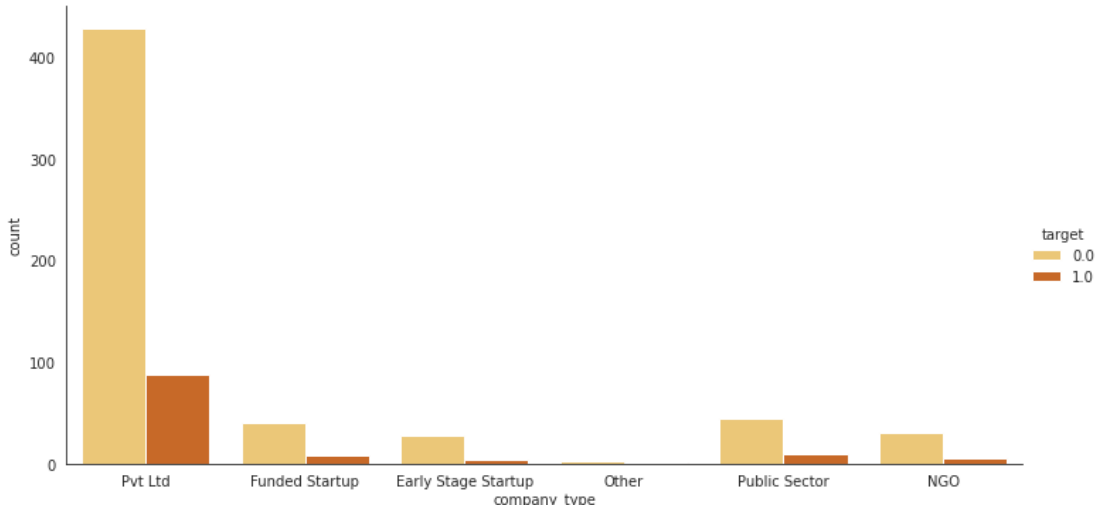
Most of the candidates are having graduation as their highest qualification followed by masters, and very few candidates are having Phd as their qualification.

5.5 Major vs. Target



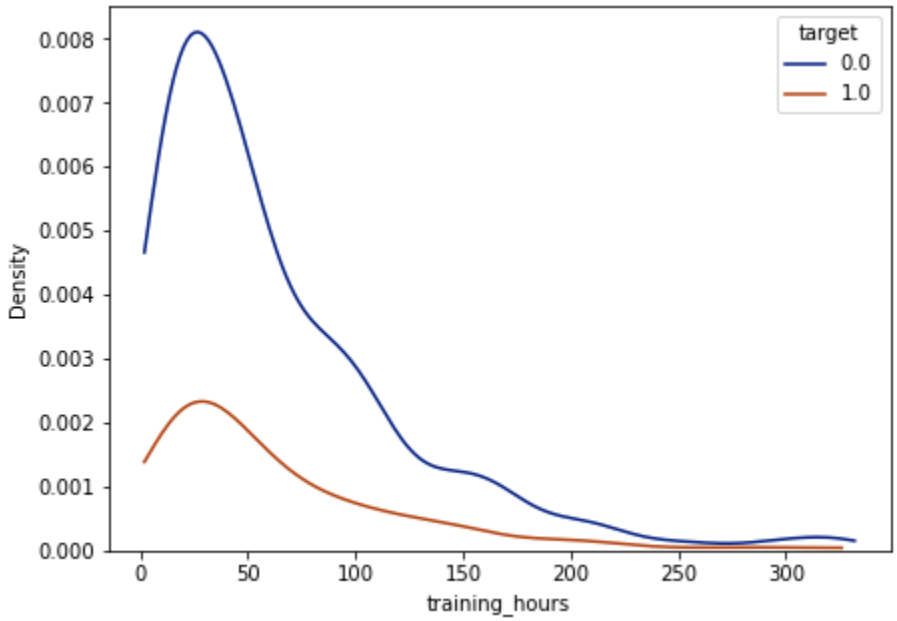
Majority of the candidates are from STEM.

5.6 Companies vs. Target



People who are working in private companies have less possibility of changing their job.

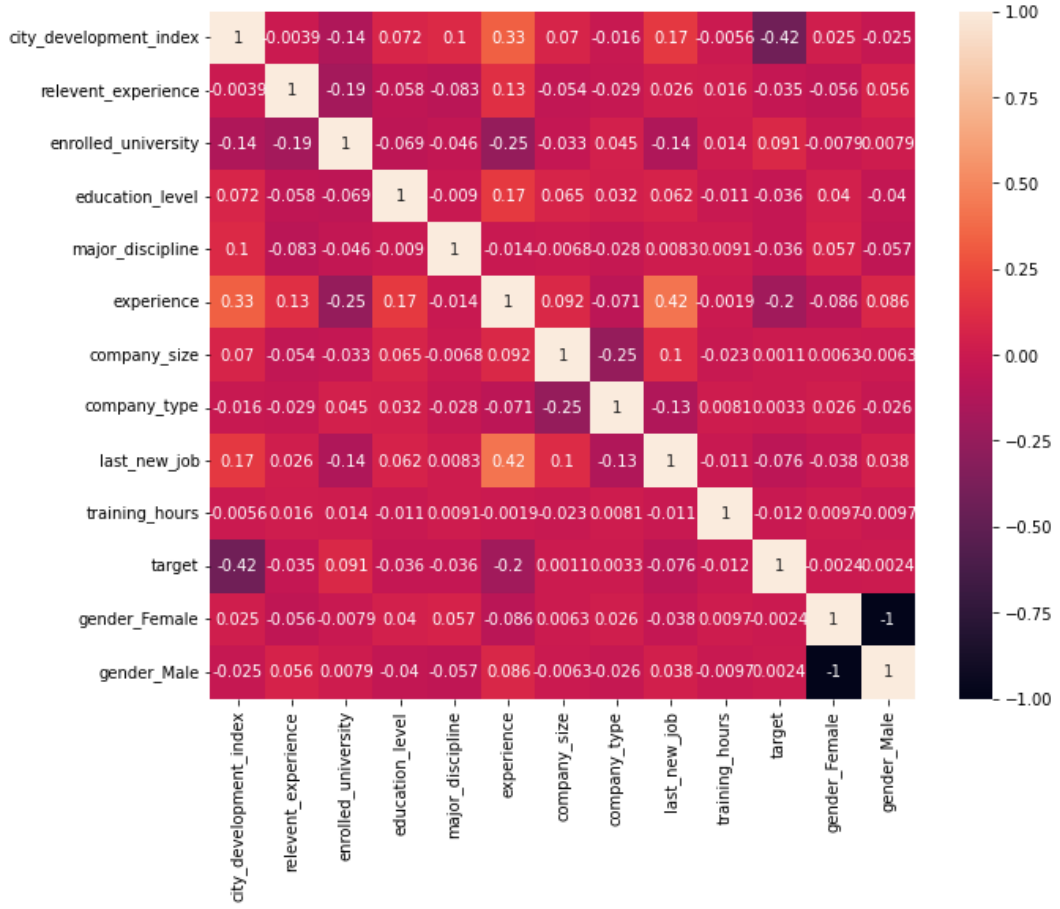
5.7 Training Hours vs. Target



From this graph, we can observe that employees with less training hours had a decreased likelihood of changing employment.

5.8 Variable Correlation Coefficient

After plotting heatmap Figure 5.1, we can see no significant correlation coefficient was found among variables.



6. Machine Learning

6.1 Data Preprocessing and Feature Selection

To Avoid From Dummy Variable Trap we will use `pd.get_dummies`

What is a Dummy Variable Trap?

The Dummy variable trap is a scenario where there are attributes which are highly correlated (Multicollinear) and one variable predicts the value of others. When we use one hot encoding for handling the categorical data, then one dummy variable (attribute) can be predicted with the help of other dummy variables. Hence, one dummy variable is highly correlated with other dummy variables. Using all dummy variables in models leads to a dummy variable trap. So, the models should be designed excluding one dummy variable.

For Example, Let's consider the case of gender having two values male (0 or 1) and female (1 or 0). Including both the dummy variables can cause redundancy because if a person is not male in such a case that person is a female, hence, we don't need to use both the variables in models. This will protect us from dummy variable traps.

6.2 Model Selection

In this section we studied the performance of 10 classification models: Logistic Regression, Gaussian Naïve Bayes, k-Nearest Neighbors, Support Vector Machine, Decision Tree, Random Forest, Gradient Boosting, XGBoost, LightGBM, CatBoost. The pros and cons of each models are summarized in Table 6.2

Table 6.2 Overview of classifier (need to summarize this table)

No.	Binary Classifier	Advantages	Disadvantages
1	Logistic Regression	• East to interpret • Small number of hyperparameters • Overfitting can be addressed though regularization	• May overfit when provided with large numbers of features • Can only learn linear hypothesis functions • Input data might need scaling • May not handle irrelevant features well
2	Gaussian Naïve Bayes	• No training involved • Very little parameter tuning is required • Features do not need scaling	• Assumes that the features are independent, which is rarely true
3	k-Nearest Neighbors	• No training involved, easy to implement • Only one hyperparameter	• Need to find optimal number of K • Slow to predict • Outlier sensitivity
4	Support Vector Machine	• Accuracy • Works well on smaller cleaner datasets • Is effective when umber of dimensions is greater than the number of samples.	• Isn't suited to larger datasets as the training time can be high • Less effective on noisier datasets with overlapping classes
5	Decision Tree	• Easy to interpret • Work with numerical and categorical features. • Requires little data preprocessing • Performs well on large datasets • Doesn't require normalization	• Overfitting • Unable to predict continuous values • Doesn't work well with lots of features and complex large dataset
6	Random Forest	• Excellent predictive power • Requires little data preprocessing • Doesn't require normalization • Suitable for large dataset • Plenty of optimization options	• Overfitting risk • Parameter complexity • Limited with regression

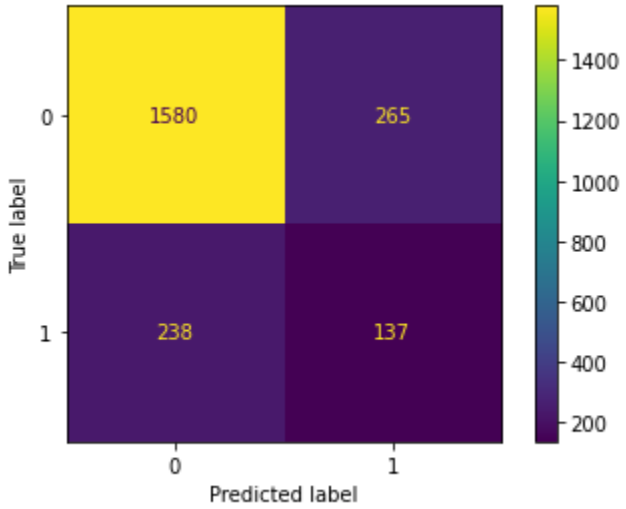
No.	Binary Classifier	Advantages	Disadvantages
7	Gradient Boosting	• Excellent predictive accuracy • Can optimize on different loss functions • Provides several hyperparameter tuning options that make the function fit very flexible.	• Overfitting risk • Computationally expensive • Parameter complexity
8	XGBoost	• Regularization to prevent overfitting • Computational efficiency and often better model performance • Can handle missing value	• Difficult to interpret and visualize • Parameter complexity • Time consuming when dataset is large
9	LightGBM	• High speed, high accuracy • Can handle missing value and categorical value • Low memory usage • Compatibility with large dataset	• Parameter complexity • Overfitting risk
10	CatBoost	• Can handle missing value and categorical value • Work well with both small and large dataset • Can monitor loss function	• Prevent overfitting • Normally doesn't need to tune hyperparameters to gain better result

6.3 Baseline Model Evaluation

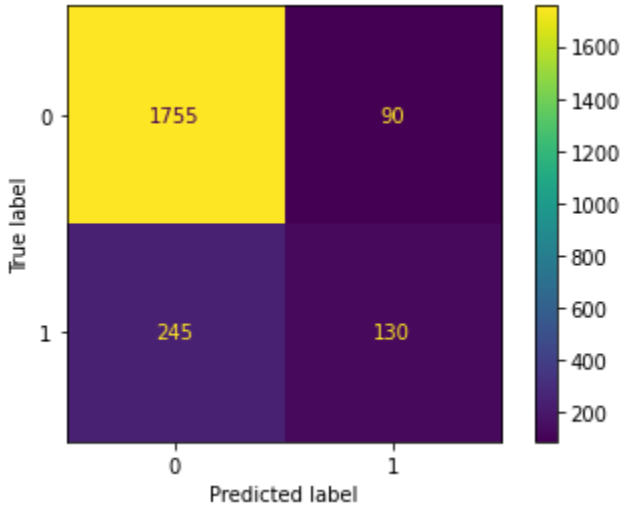
Step1: Firstly, we trained all the models with the baseline implementation, meaning all the hyperparameters of the models were left as the default value in the scikit-learn APIs.

6.4 Model Comparison

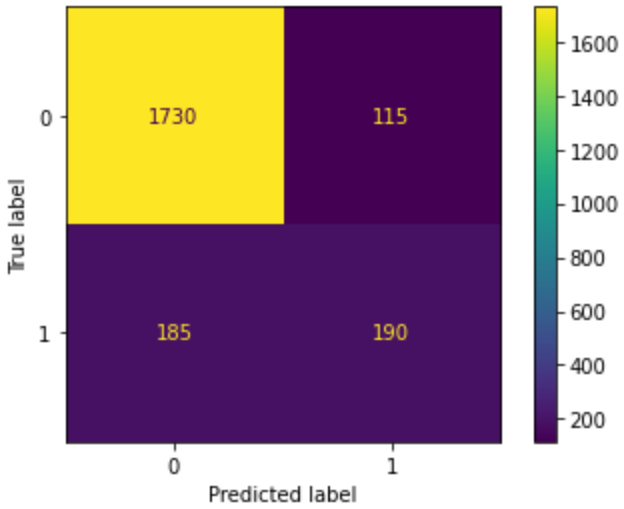
1、DecisionTreeClassifier



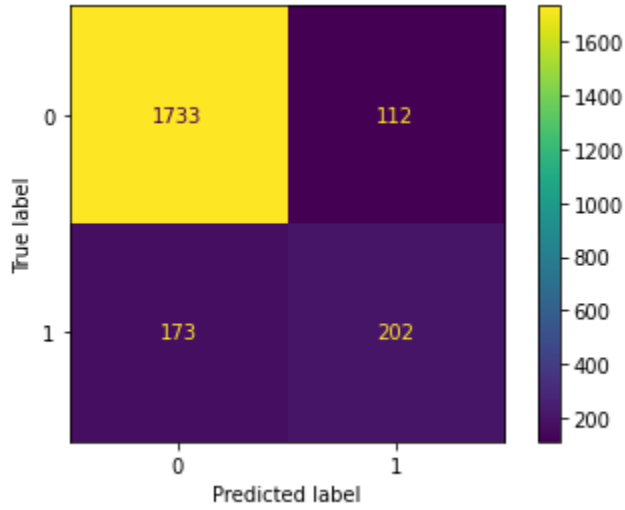
2.RandomForestClassifier



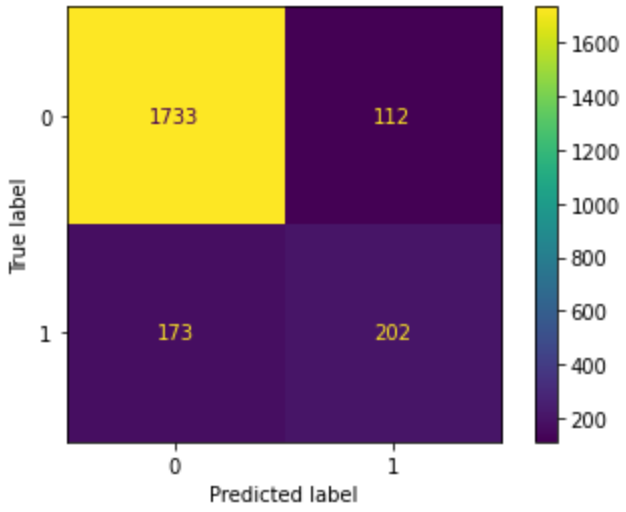
3.GradientBoostingClassifier



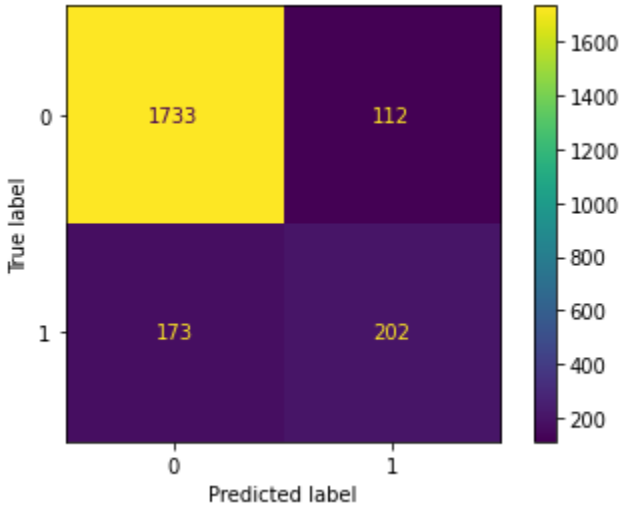
4.AdaBoostClassifier



5.MLClassifier



6.XGBClassifier



7. Final Model Selection and Hyperparameters Tuning

	Models	Accuracy Score
3	SVC	0.835
4	RandomForestClassifier	0.830
5	KNeighborsClassifier	0.805
0	LogisticRegression	0.800
2	BernoulliNB	0.775
1	GaussianNB	0.280