Final Project

# Project Title:

The purpose of our project is to build a model(s) that can predict the intention of an online shopper to make purchases at a webpage by analyzing their online behavior using different Machine Learning Methods.

# Data Source:

The dataset is from online site Kaggle:

[Online Shoppers Intention UCI Machine Learning | Kaggle](https://www.kaggle.com/datasets/henrysue/online-shoppers-intention)

(See detailed information in Appendix: Data Source segment)

# Dataset Summary:

The dataset contains feature vectors (behavior) of 12,330 sessions belonging to different visitors to a webpage. It has 17 quantitative and qualitative independent variables, in addition to a “target” column, which identifies if there was or was not a revenue from the shopper. Also, the dataset was formed so that each session would belong to a different user in a 1-year period to avoid any tendency to a specific campaign, special day, user profile, or period.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Abstract**: Of the 12,330 sessions in the dataset, 84.5% (10,422) were negative class samples that did not end with shopping, and the rest (1908) were positive class samples ending with shopping. So, it is an unbalanced dataset.   |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | **Data Set Characteristics:** | Multivariate | **Number of Instances:** | 12330 | **Area:** | Business | | **Attribute Characteristics:** | Integer, Real | **Number of Attributes:** | 18 | **Date Donated** | 2018-08-31 | | **Associated Tasks:** | Classification, Clustering | **Missing Values?** | N/A | **Number of Web Hits:** | 198385 | |  |

(Source of this table: Dataset Origin in Appendix)

(Refer Appendix: Data Summary - for list of Variables and Description.)

# **Data Statistics:**

The Dataset is clean, and it does not contain Missing Values. It has 12330 instances and 18 Variables. By analyzing the graphs, we can see none of the graphs are normally distributed. All the graphs are slightly rightly skewed. So, there is a possibility of some big outliers. We will certainly analyze outliers before building models.

(Refer Appendix: Data Statistics- for general statistics and graphs.)

# **Project Plan (Tools/ Methods):**

To achieve this goal of this project, the team will train different models using below Machine Learning methods and compare their performances.

* Binominal Logistic Regression (Lasso, Ridge)
* Support Vector Machine (Linear, RBF (Radial Basis Function), Sigmoid)

After training the models, we will evaluate the performance of each model's test data and then compare them by plotting Confusion Matrix and ROC. Also, the dataset has two limitations: Big Outliers and Unbalanced dataset. We will be using **SMOTE analysis** to make data more balanced, and we are training models using SVM (Support Vector Machine) method as it has better tolerance for outliers. All the project work will be done using both Programming languages: **Matlab and Python**. We will be using PyCharm to write all Python codes and Matlab Online to write Matlab codes.

**Project Description:**

This project aims to build an appropriate model(s) that could predict the intention of a visitor to make a purchase on a webpage. We think it is an interesting project topic and are excited to deliver it, as this project could be extremely helpful for most of the E-commerce Business. By comparing the model’s performance, the project will focus on delivering the most optimal model as per the business requirements.

The Project Summary (writeup):

This project is an overview of how Machine Learning methods can be applied to the retail industry, specifically to the task of Churn Prediction. The objective of this project is to predict the intention of an online shopper to make purchases on a website by analyzing their online behavior. To achieve this goal, we have trained different models using machine learning methods: Binominal logistics regression (Lasso and Ridge) and Support Vector Machine (using different kernel tricks: Linear, RBF, polynomial and Sigmoid) with and without cross-validation.

Because our dataset was imbalanced, we oversampled the data using SMOTE analysis before training our models. To recommend the best-performing model based on business requirements, we compared the test data result of all the models by evaluating the Confusion matrix and ROC curve. Additionally, we used different types of cross-validation techniques, such as GridSearchCV and RepeatedStratifiedKFold to improve and evaluate the performance of a model better.

We found that all models performed very well, SVM RBF Kernel being the best with 90% Model AUC. Since it is an unbalanced date we can't use accuracy, so it's better to use Model AUC. For Binominal Logistic Regression, both Lasso and Ridge performed similarly. Therefore, we concluded by recommending the SVM RBF model for this kind of dataset and business. Working on this project was a great learning experience for us, it made us more confident using Machine Learning methods in the real-world dataset. We faced many pitfalls while going through the process, the biggest among them was training our models without using Cross Validation, which gave us ambiguous results. Further, we corrected our mistake and re-did it to obtain the correct results. We could have also included other Machine Learning methods like- Naïve Bayesian and Random Forest in our project to explore more possibilities.

(Refer Appendix: Project Summary - for Model Performance Graphs)

Appendix:

# Data Source

* **Source of data:**

The dataset is taken from online site Kaggle:

[Online Shoppers Intention UCI Machine Learning | Kaggle](https://www.kaggle.com/datasets/henrysue/online-shoppers-intention)

* **Dataset Origin:**

<https://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset>

* **Creator:**

1. C. Okan Sakar  
   Department of Computer Engineering, Faculty of  
   Engineering and Natural Sciences, Bahcesehir University,  
   34349 Besiktas, Istanbul, Turkey
2. Yomi Kastro  
   Inveon Information Technologies Consultancy and Trade,  
   34335 Istanbul, Turkey

* **Citation:**

Sakar, C.O., Polat, S.O., Katircioglu, M. et al. Neural Comput & Applic (2018)

[Real-time prediction of online shoppers’ purchasing intention using multilayer perceptron and LSTM recurrent neural networks | SpringerLink](https://link.springer.com/article/10.1007/s00521-018-3523-0)

## **Sampling Frame: -**

* The dataset was collected by combining two modules. In Module 1 The data was collected to predict the purchasing intention of the visitor using aggregated pageview data kept track during the visit along with some session and user information. Module two used only sequential clickstream data. The modules are used together to determine the visitors who have purchasing intention but are likely to leave the site in the prediction horizon and take actions accordingly to improve the Web site abandonment and purchase conversion rates. (Refer: Source: Citation)
* We do not know the circumstances or methods for how these records were collected, or if that may play a role in our sample data’s relevancy to the population of interest. However, we have no reason to believe this sample data is misrepresentative of the population, so for this project's purposes, we assume the data available is relevant to the population of interest.

# 2) Dataset Summary

**List of Variables:**

|  |  |  |
| --- | --- | --- |
| **Variable name** | **Level of Measurement**  **Numerical = N**  **Categorical = C** | **Description** |
| Administrative | N | The number of pages of this type that the user visited. |
| Administrative\_Duration | N | This is the amount of time spent in this category of pages. |
| Informational | N | This is the number of pages of this type (informational) that the user visited. |
| Informational\_Duration | N | This is the amount of time spent in this category of pages. |
| ProductRelated | N | This is the number of pages of this type (product related) that the user visited. |
| ProductRelated\_Duration | N | This is the amount of time spent in this category of pages. |
| BounceRates | N | The percentage of visitors who enter the website through that page and exit without triggering any additional tasks. |
| ExitRates | N | The percentage of pageviews on the website that end at that specific page. |
| PageValues | N | The average value of the page averaged over the value of the target page and/or the completion of an eCommerce transaction. |
| SpecialDay | N | This value represents the closeness of the browsing date to special days or holidays (e.g., Mother's Day or Valentine's Day) in which the transaction is more likely to be finalized. |
| Month | C | Contains the month the pageview occurred, in string form |
| OperatingSystems | C | An integer value representing the operating system that the user was on when viewing the page. |
| Browser | C | An integer value representing the browser that the user was using to view the page. |
| Region | C | An integer value representing which region the user is in. |
| TrafficType | C | describes how traffic arrived on the website (Direct, Organic, Referral, Social, Email, Display and Paid) |
| VisitorType | C | A string representing whether a visitor is New Visitor, Returning Visitor, or Other. |
| Weekend | C | A Boolean representing whether the session is on a weekend. |
| Revenue | C | A Boolean representing whether the user completed the purchase. |

# Dataset Statistics:

1. **Snapshot of Dataset:**

Graphical user interface, application, table, Excel

Description automatically generated

1. **Number of Rows and Columns:**

df = pd.read\_csv("online\_shoppers\_intention.csv")  
pd.set\_option("display.max\_columns", None)  
print(df.shape)

**(12330, 18)**

1. **Statistics:**

print(df.describe())

**Administrative Administrative\_Duration Informational**

**count 12330.000000 12330.000000 12330.000000**

**mean 2.315166 80.818611 0.503569**

**std 3.321784 176.779107 1.270156**

**min 0.000000 0.000000 0.000000**

**25% 0.000000 0.000000 0.000000**

**50% 1.000000 7.500000 0.000000**

**75% 4.000000 93.256250 0.000000**

**max 27.000000 3398.750000 24.000000**

**Informational\_Duration ProductRelated ProductRelated\_Duration**

**count 12330.000000 12330.000000 12330.000000**

**mean 34.472398 31.731468 1194.746220**

**std 140.749294 44.475503 1913.669288**

**min 0.000000 0.000000 0.000000**

**25% 0.000000 7.000000 184.137500**

**50% 0.000000 18.000000 598.936905**

**75% 0.000000 38.000000 1464.157214**

**max 2549.375000 705.000000 63973.522230**

**BounceRates ExitRates PageValues SpecialDay**

**count 12330.000000 12330.000000 12330.000000 12330.000000**

**mean 0.022191 0.043073 5.889258 0.061427**

**std 0.048488 0.048597 18.568437 0.198917**

**min 0.000000 0.000000 0.000000 0.000000**

**25% 0.000000 0.014286 0.000000 0.000000**

**50% 0.003112 0.025156 0.000000 0.000000**

**75% 0.016813 0.050000 0.000000 0.000000**

**max 0.200000 0.200000 361.763742 1.000000**

**OperatingSystems Browser Region TrafficType**

**count 12330.000000 12330.000000 12330.000000 12330.000000**

**mean 2.124006 2.357097 3.147364 4.069586**

**std 0.911325 1.717277 2.401591 4.025169**

**min 1.000000 1.000000 1.000000 1.000000**

**25% 2.000000 2.000000 1.000000 2.000000**

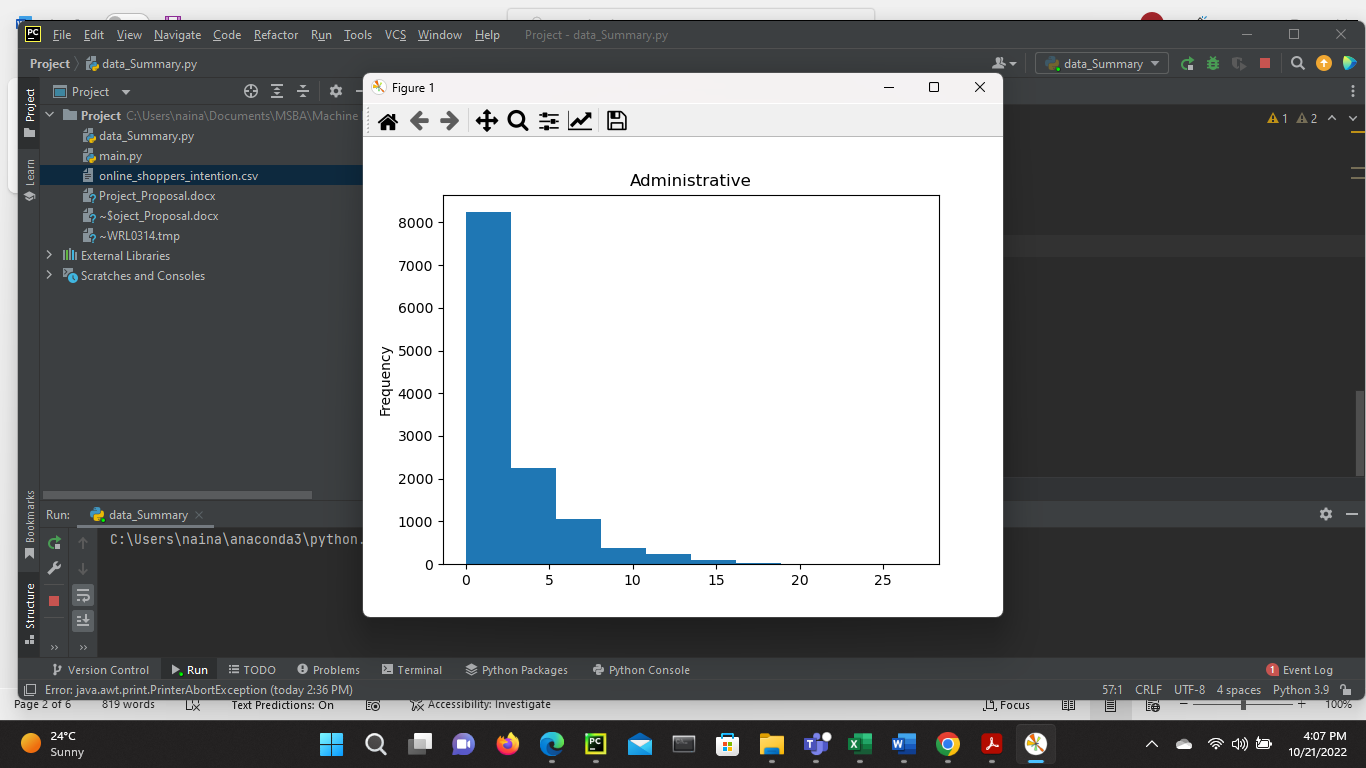
**50% 2.000000 2.000000 3.000000 2.000000**

**75% 3.000000 2.000000 4.000000 4.000000**

**max 8.000000 13.000000 9.000000 20.000000**

1. **Frequency Distribution:**

**Numerical Variables:**

Graphical user interface, application, Word

Description automatically generated

Graphical user interface, application

Description automatically generatedGraphical user interface, application, Word

Description automatically generatedGraphical user interface, application, Word

Description automatically generatedGraphical user interface, application

Description automatically generated

Graphical user interface, application, Word

Description automatically generatedChart, histogram

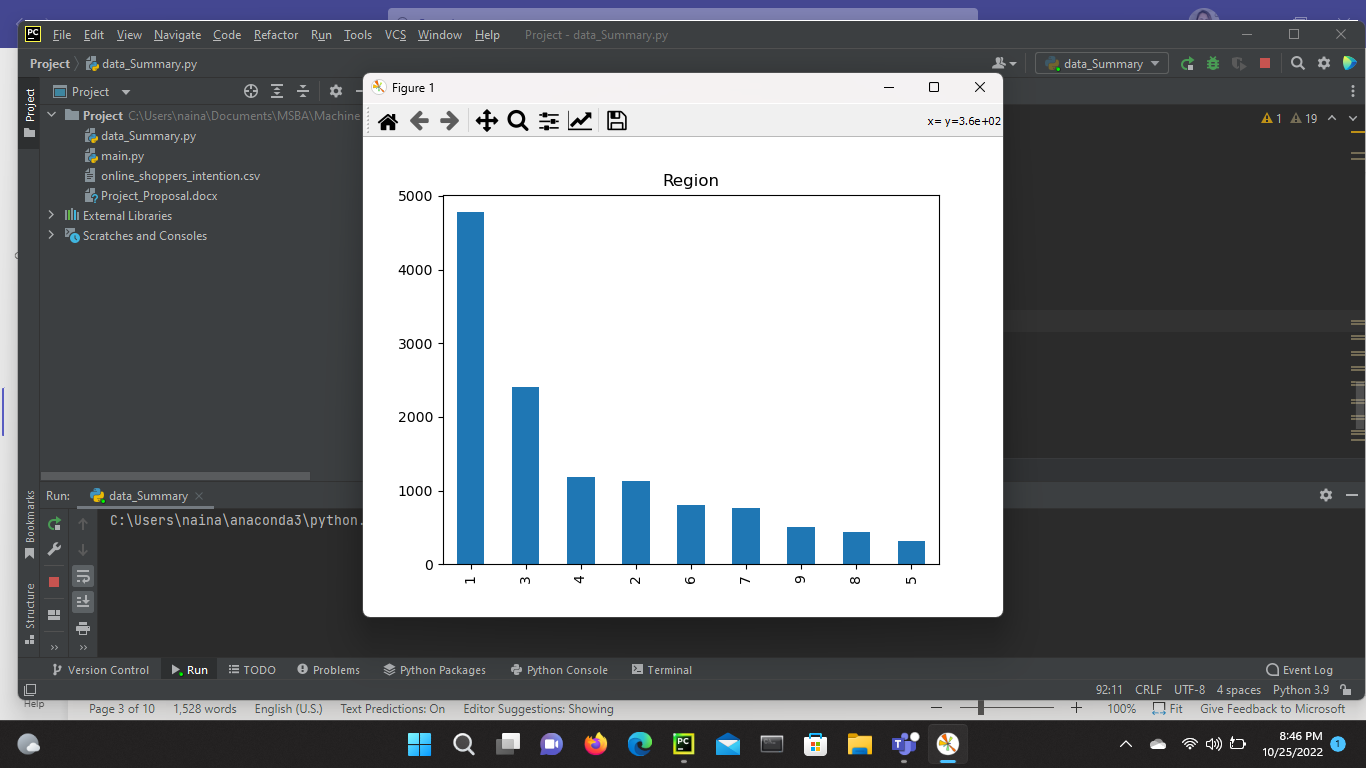
Description automatically generated

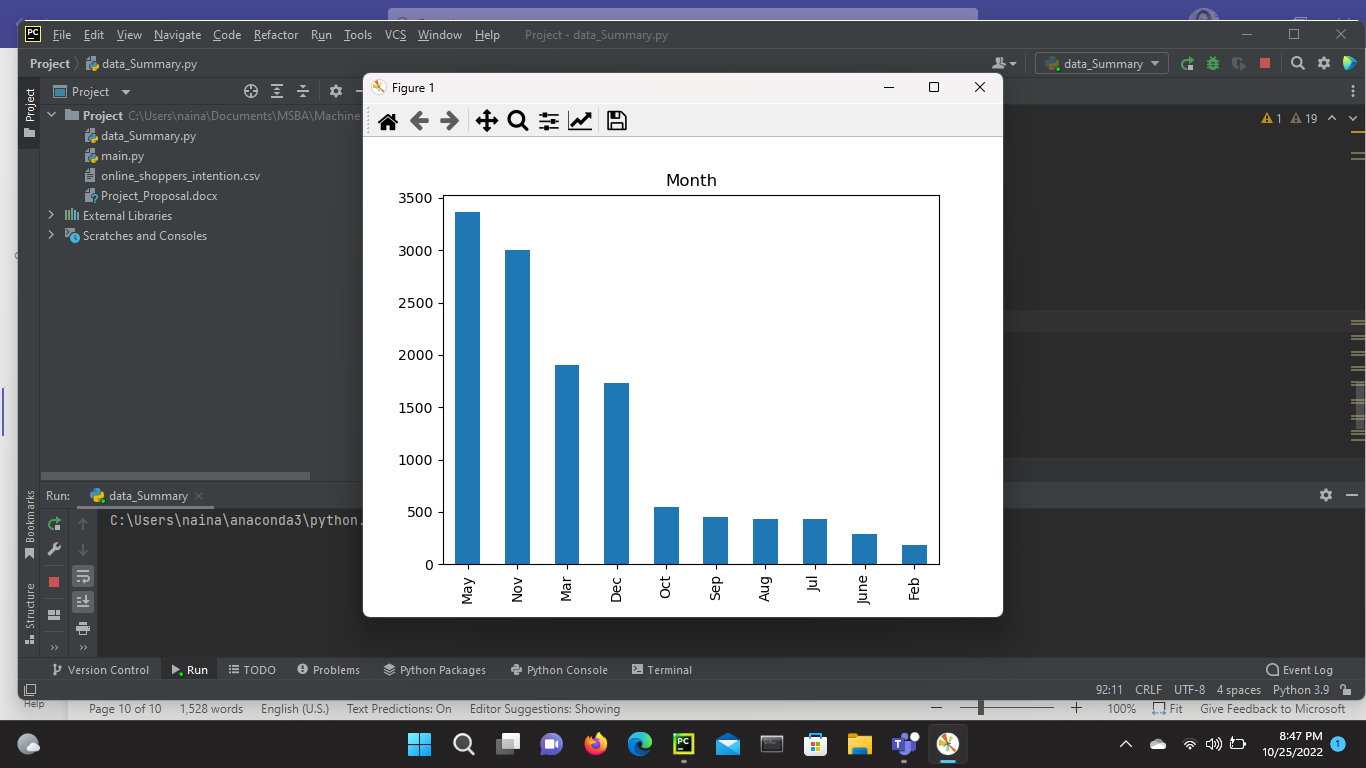
Graphical user interface, application

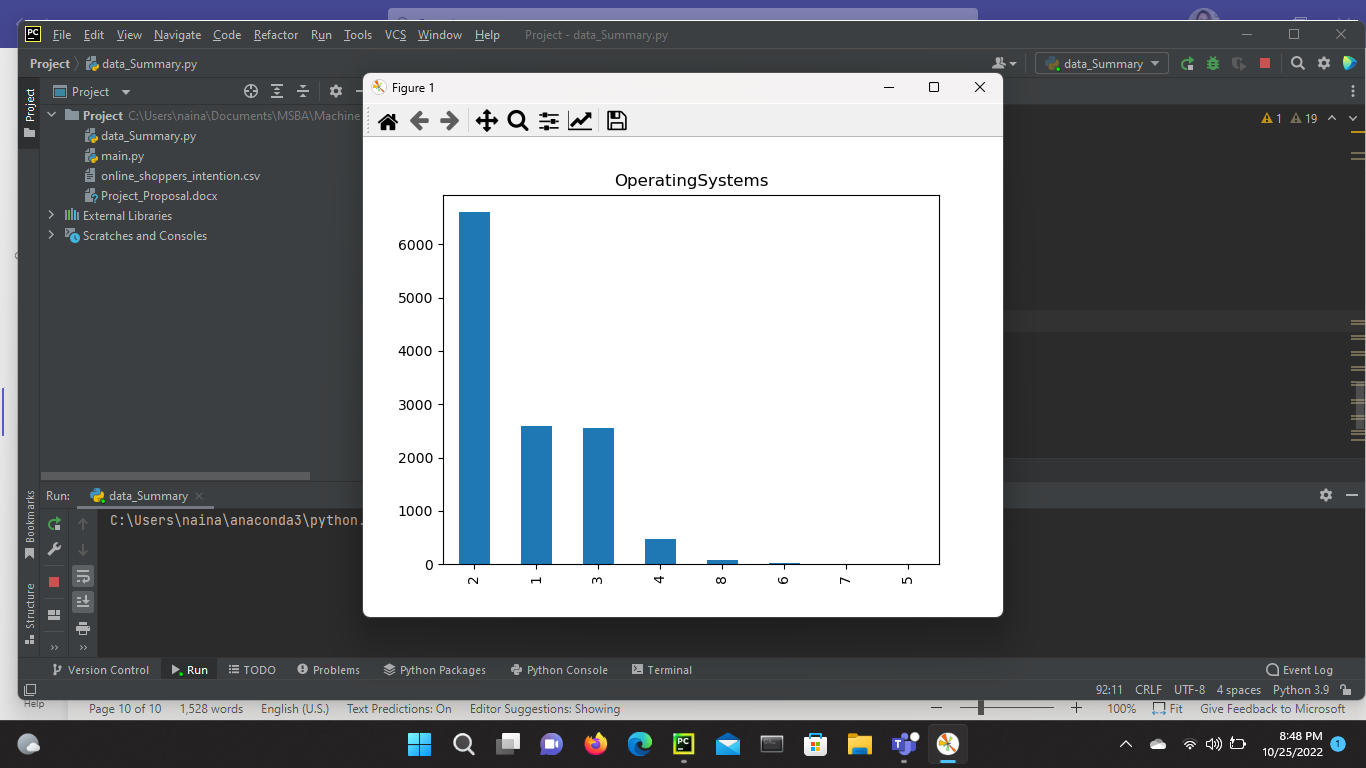
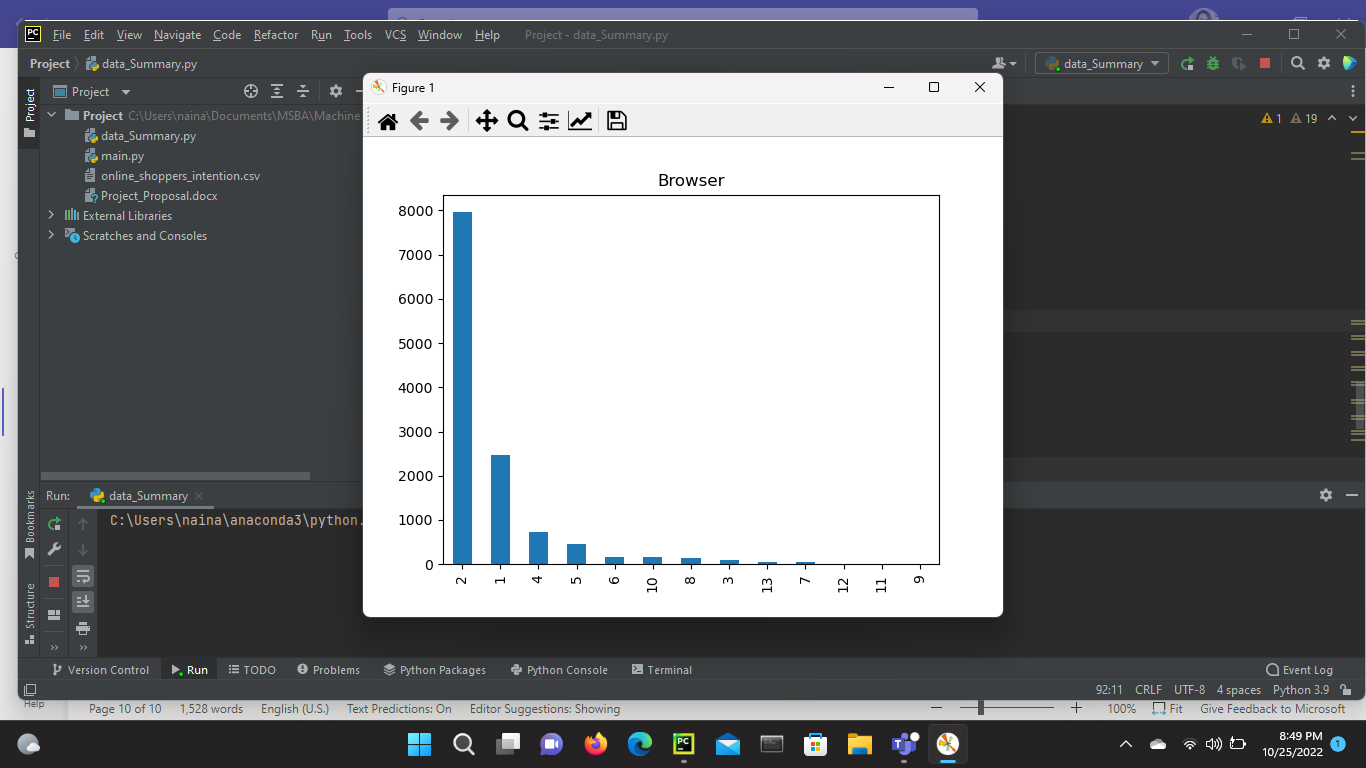
Description automatically generatedGraphical user interface, application, Word

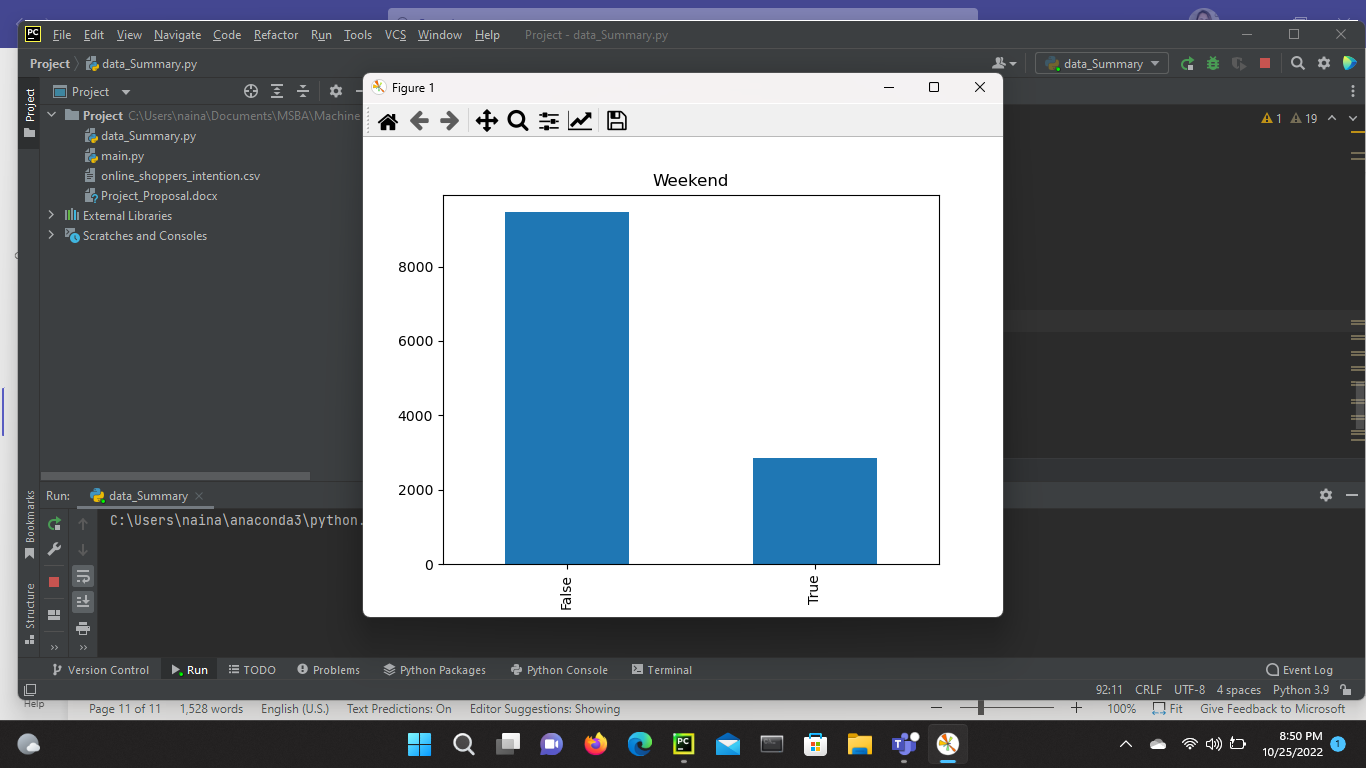
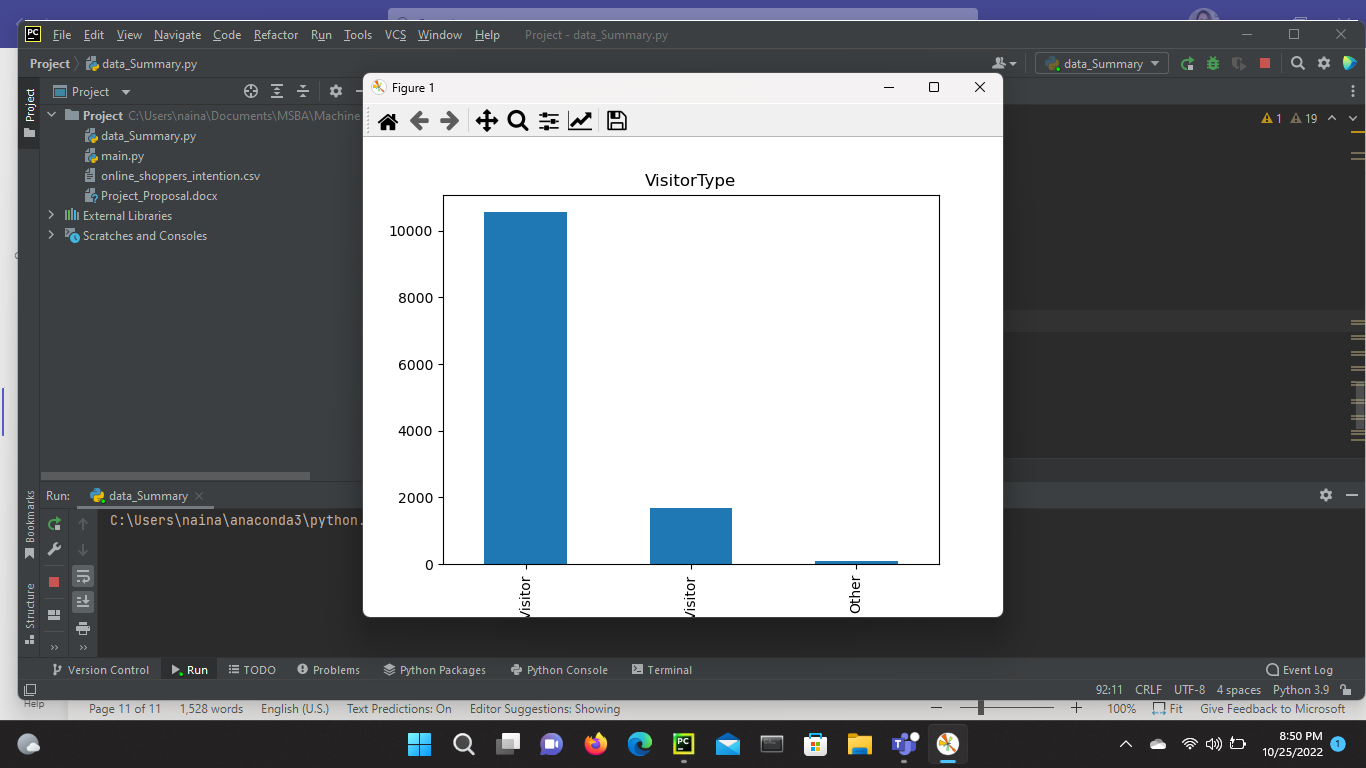
Description automatically generated

**Categorical Variables:**

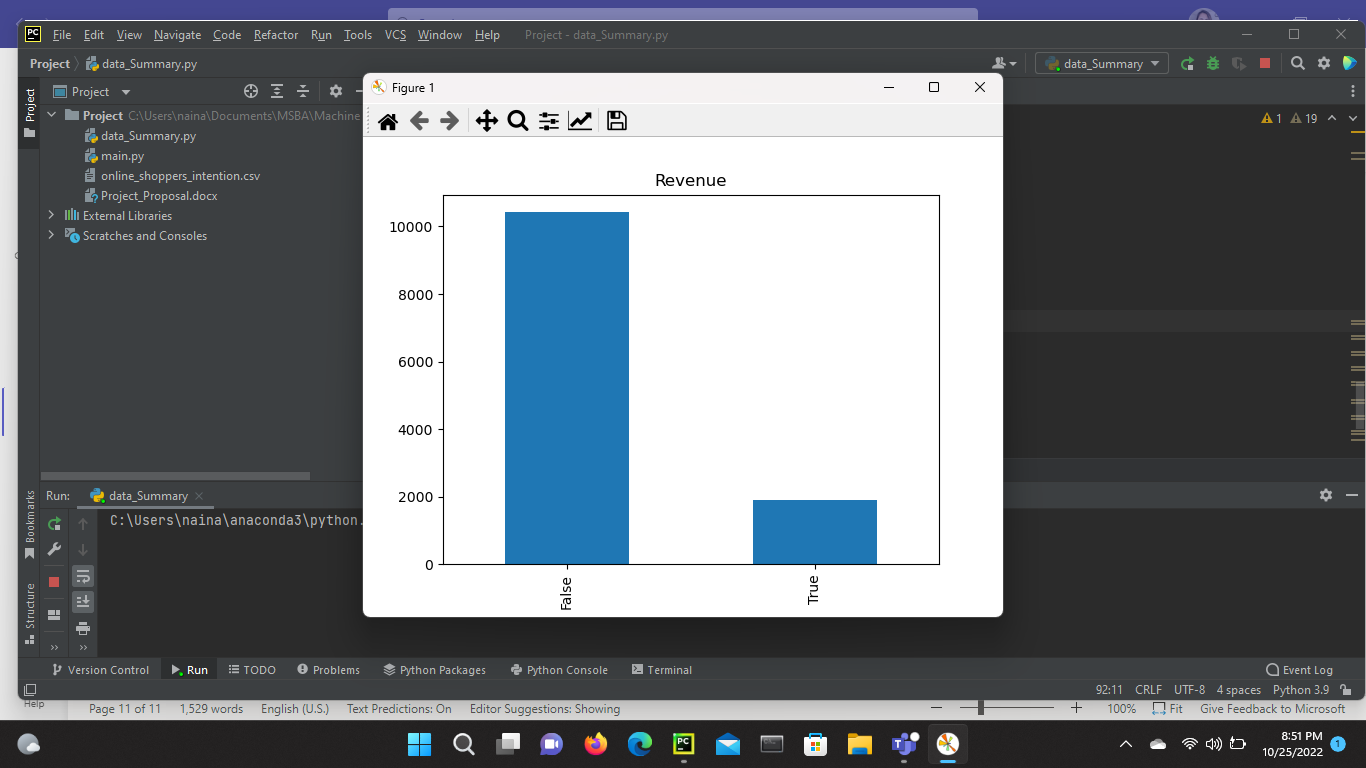




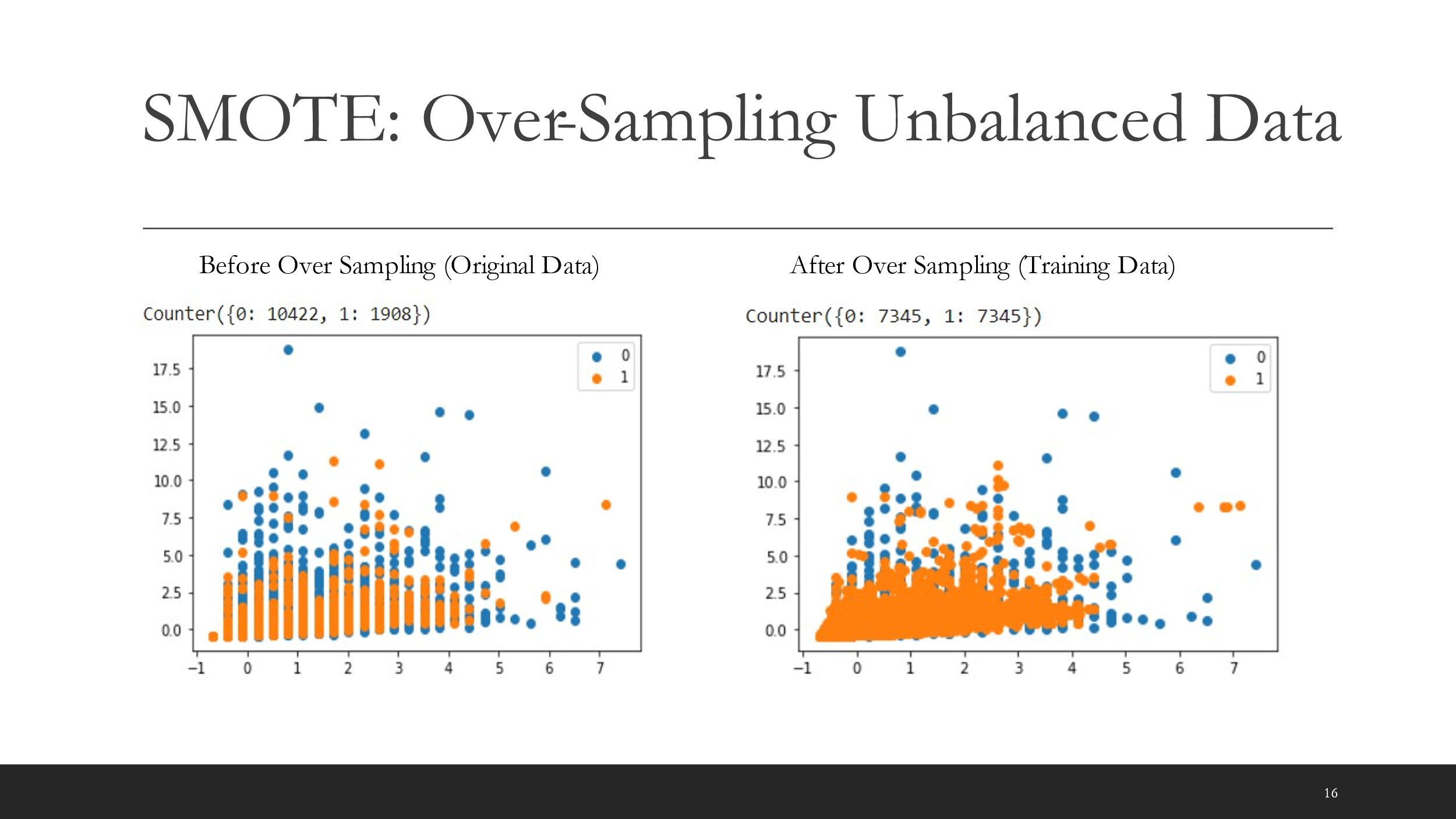


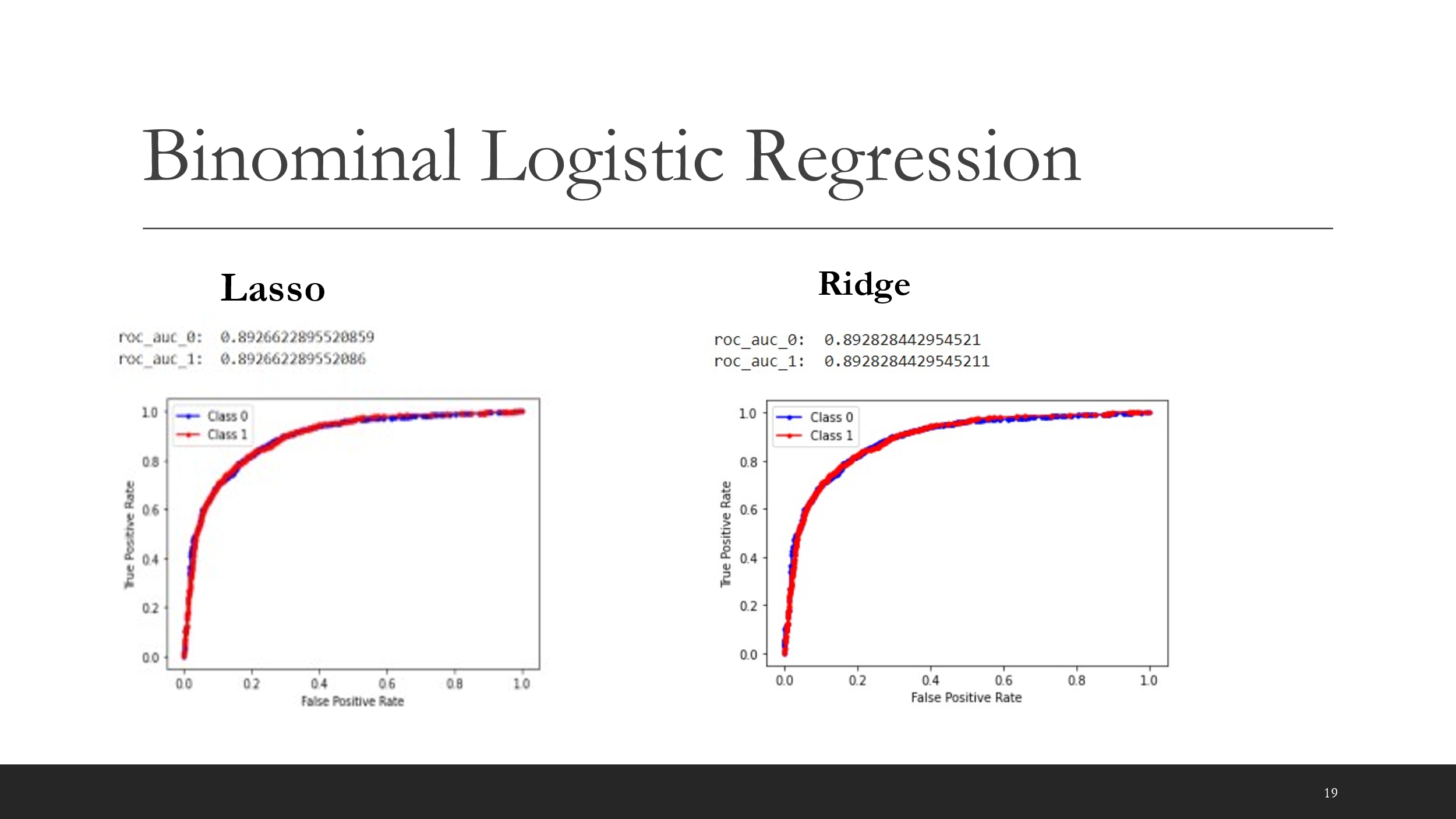
**Target:**



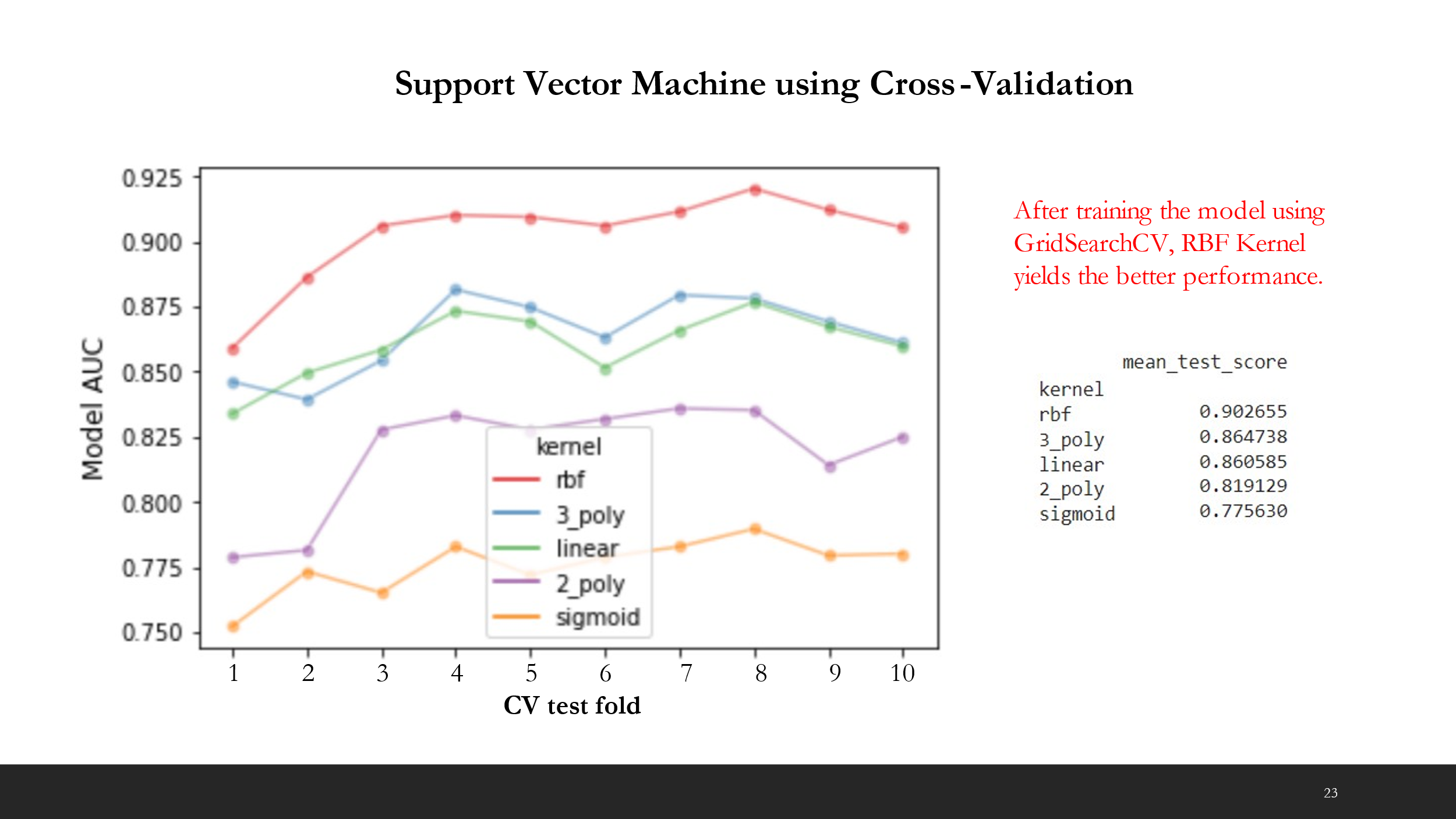
# Project Summary

**SMOTE Analysis**



**Results from Binominal Logistic Regression**

**Results from Support Vector Machine methods using GridSearchCV:**



**Pitfall:**

**Ambiguous Results before doing Cross Validation**

