T.C.

MARMARA UNIVERSITY

FACULTY of ENGINEERING

CSE4062 Introduction to Data Science and Analytics

Spring 2021

Group #1

Delivery #3: Exploring Your Data Part 2

Title of the Project

Machine Learning Approach to U.S. Stock Investments

Group Members

CSE 150118825 Ahmet Hakan Ekşi ahe9953@gmail.com

ENVE 150215036 Canberk Köroğlu canberkkoroglu@hotmail.com

ENVE 150215045 Erim Varış ermvrs@live.com

Lecturer

Doç. Dr. Murat Can Ganiz

1- Feature Selection & Charts

We have 225 features. One of them is for classification, one of them for generating class label, one of them is year, one of them is the Stock ID. There are 221 features left. One is nominal/categorical feature that specifies companies' sector. Other 220 features are financial indicator and all of them is numerical attribute.

Our data has considerable number of missing values. Therefore, in feature selection we first drop the attributes that has more than 33% missing values, 16 attributes. There are 206 attributes left. And then we replace missing values with mean value of attributes, considering each sector individually.

In final we choose our best 20 attributes by using ANOVA F-value. We also try with Mutual Information but with MI, our 20 attributes' missing values percentages before replacing as very close to our threshold 33%. We want to use more raw data; therefore, we choose ANOVA F-value for feature selection.

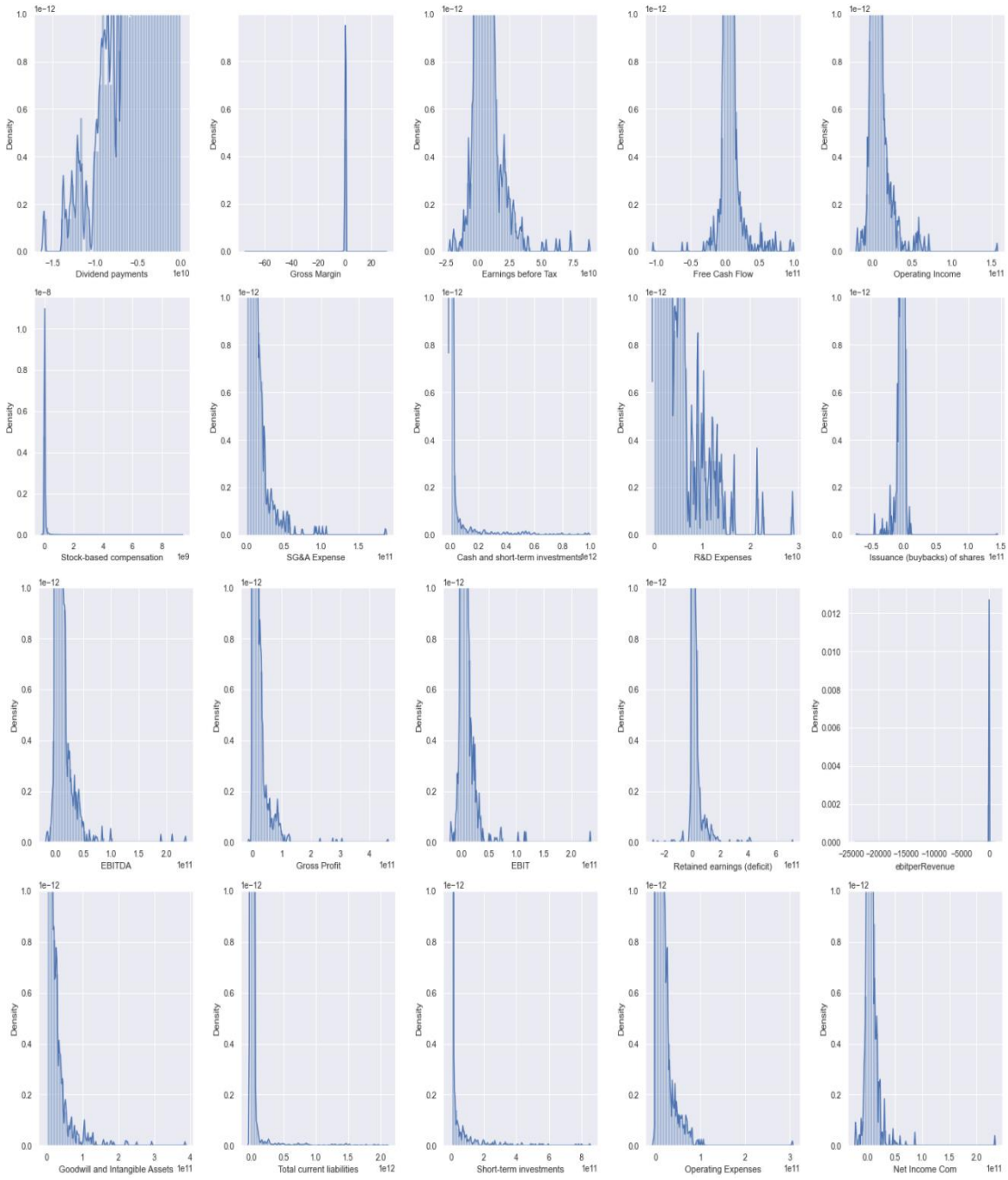

Fig. 1. Distribution charts

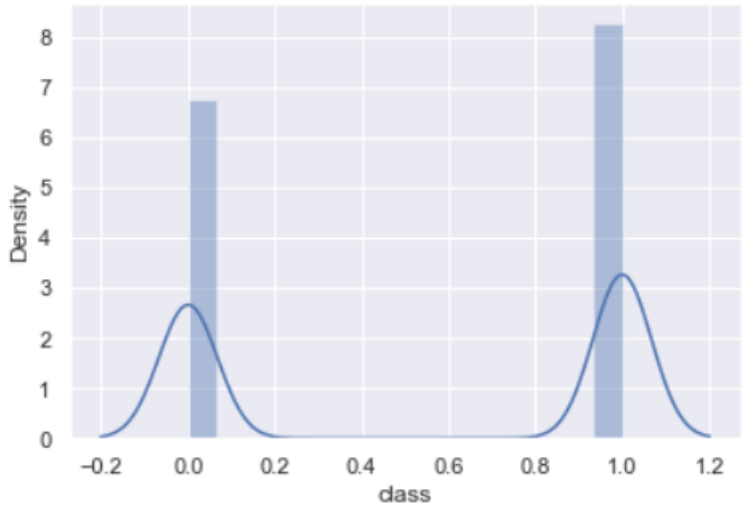

Fig. 2. Class Distribution chart

This is our class labels density chart. As you can see from the chart our stocks prices most likely to increase next year.

2- Scatter Plot Matrix

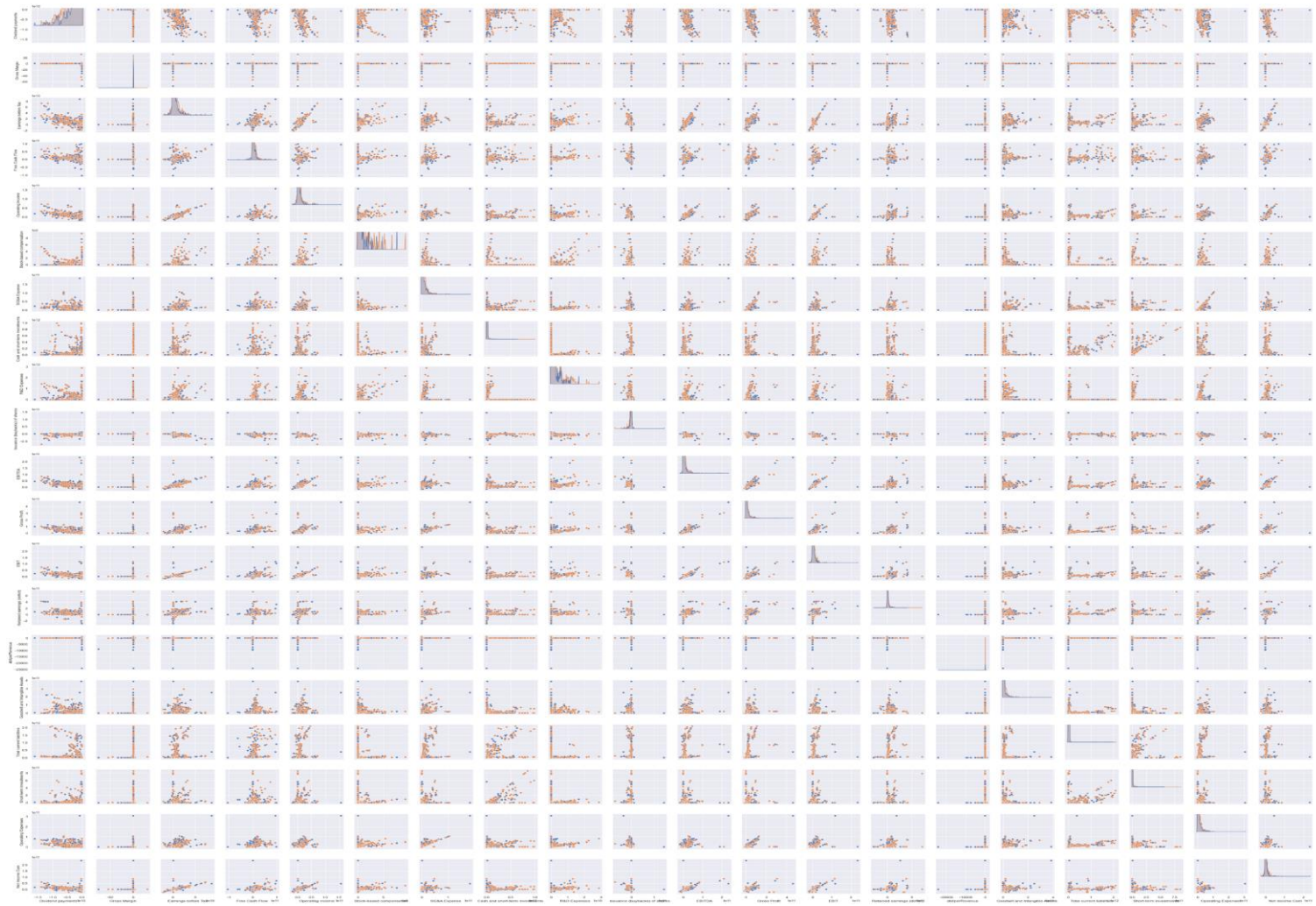

Fig. 3. Scatter Plot Matrix

In Figure 3 relationships between financial indicators are observed. X-axis shows a specific indicator, and y-axis shows another indicator, and the plot shows their relationships. Interactions between these indicators are mostly positively correlated for e.g. (3,6) or negatively correlated for e.g. (17,17) due to slopes being either negative or positive.