T.C.

MARMARA UNIVERSITY

FACULTY of ENGINEERING

CSE4062 Introduction to Data Science and Analytics

Spring 2021

Group #1

Delivery #4: Predictive Analysis

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

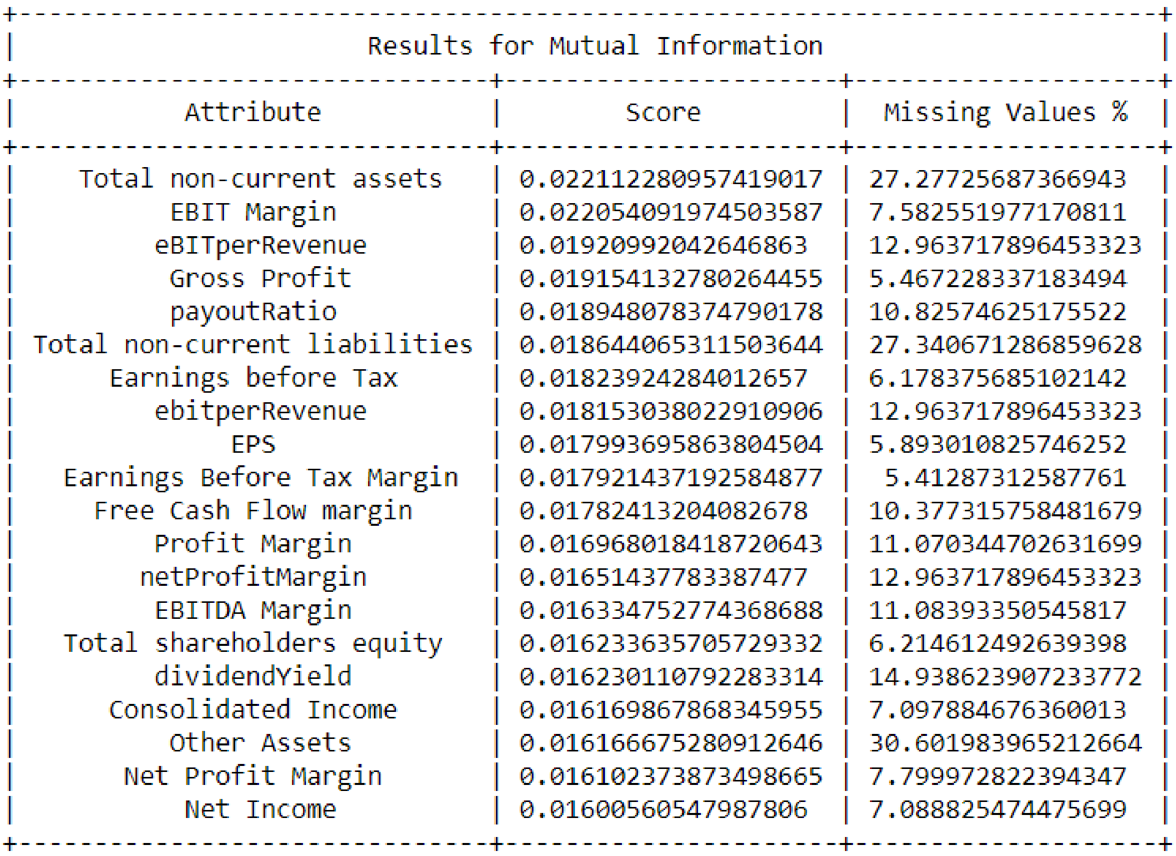
Lecturer

Doç. Dr. Murat Can Ganiz

1- Feature Selection Methods

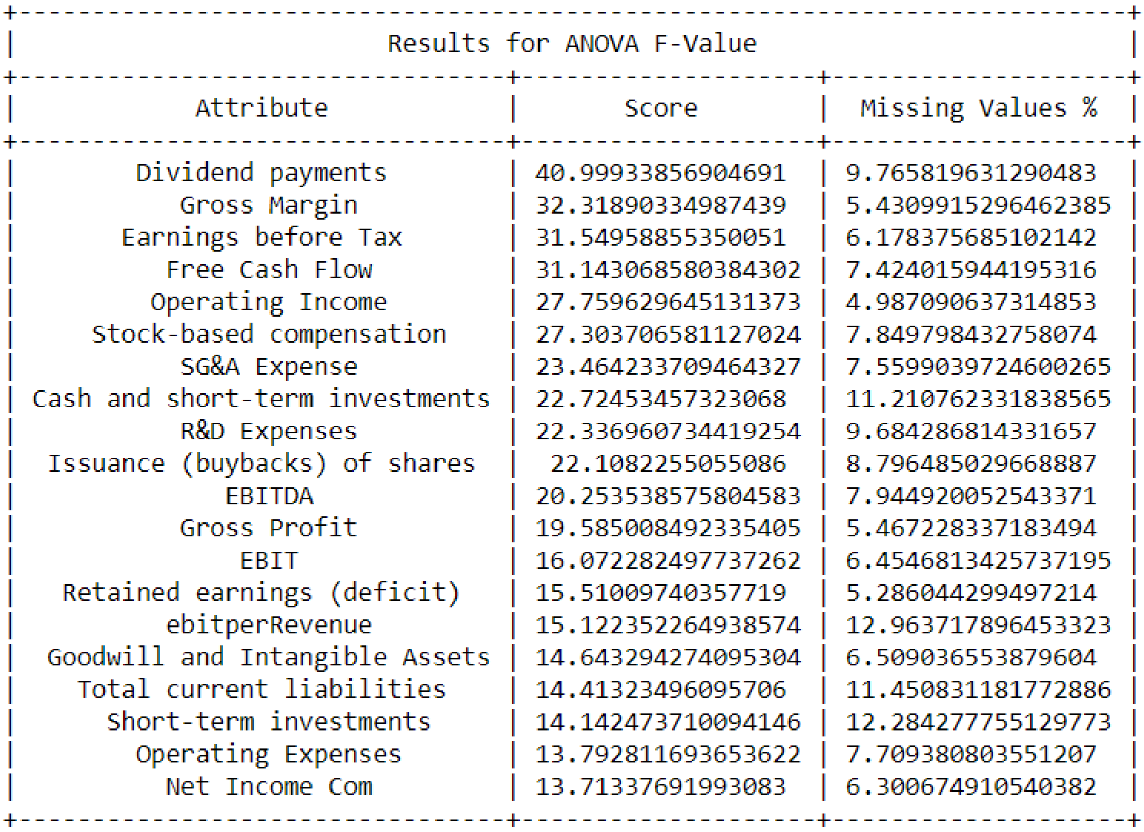
Due to data having large numbers of missing values we drop the attributes that has more than 33% missing values and we replace missing values with mean value of attributes.

Table 1. Mutual Information



```
+---------------------------------------------------------------------------+
|                      Results for Mutual Information                       |
+-------------------------------+----------------------+--------------------+
|           Attribute           |        Score         |  Missing Values %  |
+-------------------------------+----------------------+--------------------+
|   Total non-current assets    | 0.022112280957419017 | 27.27725687366943  |
|          EBIT Margin          | 0.022054091974503587 | 7.582551977170811  |
|        eBITperRevenue         | 0.01920992042646863  | 12.963717896453323 |
|         Gross Profit          | 0.019154132780264455 | 5.467228337183494  |
|          payoutRatio          | 0.018948078374790178 | 10.82574625175522  |
| Total non-current liabilities | 0.018644065311503644 | 27.340671286859628 |
|      Earnings before Tax      | 0.01823924284012657  | 6.178375685102142  |
|        ebitperRevenue         | 0.018153038022910906 | 12.963717896453323 |
|              EPS              | 0.017993695863804504 | 5.893010825746252  |
|  Earnings Before Tax Margin   | 0.017921437192584877 |  5.41287312587761  |
|     Free Cash Flow margin     | 0.01782413204082678  | 10.377315758481679 |
|         Profit Margin         | 0.016968018418720643 | 11.070344702631699 |
|        netProfitMargin        | 0.01651437783387477  | 12.963717896453323 |
|         EBITDA Margin         | 0.016334752774368688 | 11.08393350545817  |
|   Total shareholders equity   | 0.016233635705729332 | 6.214612492639398  |
|         dividendYield         | 0.016230110792283314 | 14.938623907233772 |
|      Consolidated Income      | 0.016169867868345955 | 7.097884676360013  |
|         Other Assets          | 0.016166675280912646 | 30.601983965212664 |
|       Net Profit Margin       | 0.016102373873498665 | 7.799972822394347  |
|          Net Income           | 0.01600560547987806  | 7.088825474475699  |
+-------------------------------+----------------------+--------------------+
```

Table 2. ANOVA F-Value



```
+---------------------------------------------------------------------------+
|                         Results for ANOVA F-Value                         |
+---------------------------------+--------------------+--------------------+
|            Attribute            |       Score        |  Missing Values %  |
+---------------------------------+--------------------+--------------------+
|        Dividend payments        | 40.99933856904691  | 9.765819631290483  |
|          Gross Margin           | 32.31890334987439  | 5.4309915296462385 |
|       Earnings before Tax       | 31.54958855350051  | 6.178375685102142  |
|         Free Cash Flow          | 31.143068580384302 | 7.424015944195316  |
|        Operating Income         | 27.759629645131373 | 4.987090637314853  |
|    Stock-based compensation     | 27.303706581127024 | 7.849798432758074  |
|          SG&A Expense           | 23.464233709464327 | 7.5599039724600265 |
| Cash and short-term investments | 22.72453457323068  | 11.210762331838565 |
|          R&D Expenses           | 22.336960734419254 | 9.684286814331657  |
|  Issuance (buybacks) of shares  |  22.1082255055086  | 8.796485029668887  |
|             EBITDA              | 20.253538575804583 | 7.944920052543371  |
|          Gross Profit           | 19.585008492335405 | 5.467228337183494  |
|              EBIT               | 16.072282497737262 | 6.4546813425737195 |
|   Retained earnings (deficit)   | 15.51009740357719  | 5.286044299497214  |
|         ebitperRevenue          | 15.122352264938574 | 12.963717896453323 |
| Goodwill and Intangible Assets  | 14.643294274095304 | 6.509036553879604  |
|    Total current liabilities    | 14.41323496095706  | 11.450831181772886 |
|     Short-term investments      | 14.142473710094146 | 12.284277755129773 |
|       Operating Expenses        | 13.792811693653622 | 7.709380803551207  |
|         Net Income Com          | 13.71337691993083  | 6.300674910540382  |
+---------------------------------+--------------------+--------------------+
```

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

2- Classification Experiments

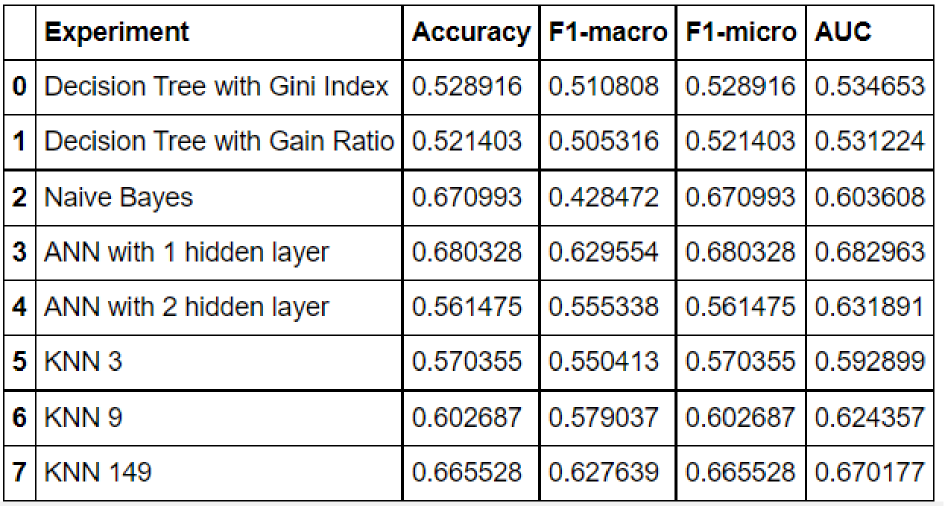
We did not use Cross Validation or any other similar methods alike. We split our train and test set according to years. Last year's csv, 2018, is our test set.

The reasoning behind this is we thought ourselves at the end of 2018 and wanted to profit next year by using the algorithm. Thus, we use features selected in ANOVA F-Value.

Methods used for classification in this experiment;

- Decision Tree with Gini Index which is calculated by subtracting the sum of squared probabilities of each class from one.
- Decision Tree with Gain Ratio which determines the information gain of all the attributes, and then computes the average information gain.
- Naïve Bayes which are based on applying Bayes' theorem with strong independence assumptions between the features.
- Artificial Neural Networks which are designed to simulate the way the human brain analyzes and processes information.
- K Nearest Neighbor assumes that similar things exist in close proximity.

Table 3. Table for Evaluation for Classification Experiments



	Experiment	Accuracy	F1-macro	F1-micro	AUC
0	Decision Tree with Gini Index	0.528916	0.510808	0.528916	0.534653
1	Decision Tree with Gain Ratio	0.521403	0.505316	0.521403	0.531224
2	Naive Bayes	0.670993	0.428472	0.670993	0.603608
3	ANN with 1 hidden layer	0.680328	0.629554	0.680328	0.682963
4	ANN with 2 hidden layer	0.561475	0.555338	0.561475	0.631891
5	KNN 3	0.570355	0.550413	0.570355	0.592899
6	KNN 9	0.602687	0.579037	0.602687	0.624357
7	KNN 149	0.665528	0.627639	0.665528	0.670177

The performance evaluation table shows the most accurate method is ANN with 1 hidden layer. Considering Area Under Curve (AUC) which measures performance across all possible classification thresholds it suggests ANN with 1 hidden layer overperforms when compared with Naïve-Bayes which is also less accurate. Also, F1-micro (micro-averages) suggests ANN with 1 hidden layer performs better. Furthermore, F1-macro (macro-averages) indicate ANN with 1 hidden layer performs better.

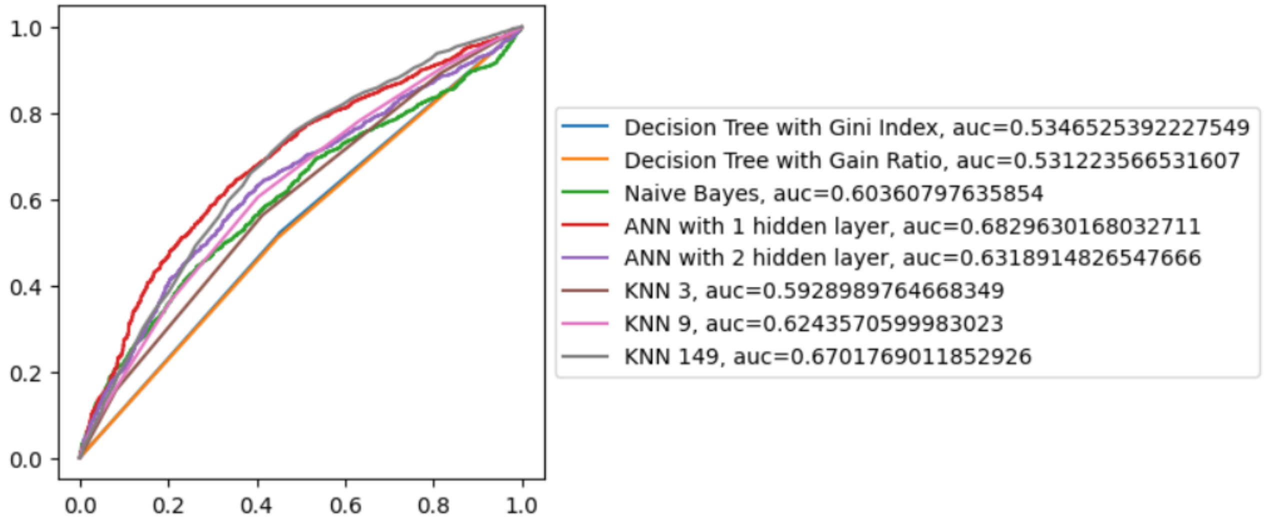


Fig.1 ROC Curve

The ROC curve shows the trade-off between sensitivity and specificity. Classifiers that give curves closer to the top-left corner indicate a better performance. The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test. This also shows ANN with 1 hidden layer is the best performing method followed by Naïve-Bayes.

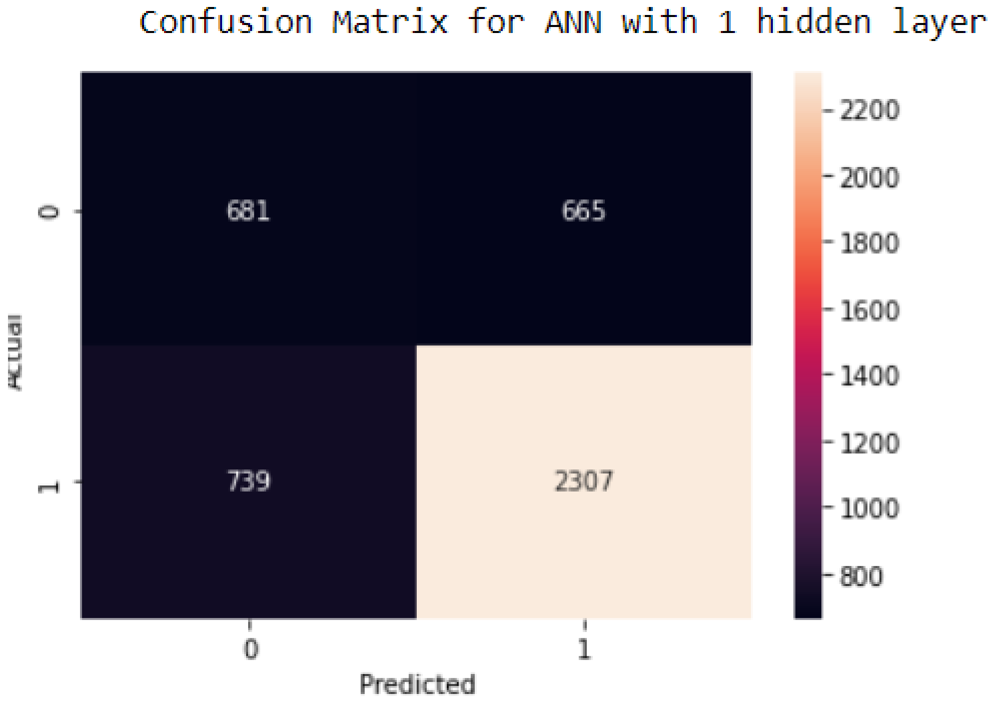


Fig.2 Confusion Matrix for ANN with 1 hidden layer

In Figure.2 it is observed that ANN with 1 hidden layer method predicted true positive cases 2307 times out of 4392 cases and true negative cases 681 times.

This suggests 68% of the algorithm is classifying correctly. Even though theoretically any algorithm can reach 100% accuracy, our achievement of 68% is good enough. Case being that if this algorithm reached maximum accuracy Group-1 would probably be very rich next year because of this achievement.

5- Statistical significance analysis between your best performing model and its closest competitor

Best Model: ANN with 1 hidden layer, Closest Competitor: Naïve-Bayes

Accuracy:

The P-value is = 0.002
The t-statistics is = 5.914
Since p<0.05, We can reject the null-hypothesis in terms of accuracy that both models perform equally well on this dataset. We may conclude that the two algorithms are significantly different.

F1-Macro:

The P-value is = 0.005
The t-statistics is = 4.883
Since p<0.05, We can reject the null-hypothesis in terms of f1_macro that both models perform equally well on this dataset. We may conclude that the two algorithms are significantly different.

F1-Micro:

The P-value is = 0.059
The t-statistics is = 2.428
Since p>0.05, we cannot reject the null hypothesis in terms of f1_micro may conclude that the performance of the two algorithms is not significantly different.

AUC:

The P-value is = 0.398
The t-statistics is = -0.923
Since p>0.05, we cannot reject the null hypothesis in terms of AUC and may conclude that the performance of the two algorithms is not significantly different.