CAPSTONE PROJECT

Lottery Winning Numbers Prediction

Abstract

Problems are solved based on techniques and models of:

Natural Language Processing (NLP)
Classification models
Long Short-Term Memory (LSTM)

VO NHI

Date: 15/April/2023

Contents

Problem statement

There are various ways of earning money from the fix income or investments in long-term or short-term plan. Because of the prosperous development of stock market in stage of the economy going up, more people are involved in spending their budget on trading stocks. Investors try to come up with such powerful models to help them predict stock prices and market behaviour in different phases. Abundant of professional researches and public papers are mentioning different strategies and hypothetical theories to interpret and predict stock market.

However, when we come to talk about lottery. No one really sees it as an investment rather than just gambling and trying to get one luck one day after hundreds of trials and failures. Therefore, it is barely to find a good paper unveiling some statistical and insightful analysis behind those winning numbers. It is worthy for us to prove if those winning numbers are drawn randomly or they are really following a particular frequency and the next drawn number could be affected by the previous numbers.

Industry/ domain:

Lotto NZ delivers part of prizes to do charity and serve back to the community. That 25 cents in every dollar spent on Lotto goes to the community, after it pays for prizes, retail commissions, levies, taxes and overheads such as staff and advertising. Lotto playing environment makes itself so easy to access to any players regardless of their age, gender, education background, etc. Due to the cheap initial cost of paying the lottery ticket, Lotto service provider creates rules and wide range of numbers for choosing. That attracts a large number of participants every year.

The prospect of TOTO numbers analysis is under the sub-domain of analysing stock market or financial market. It just takes the branch of statistics with the absence of other financial factors.

Stakeholders

Through the analysis report, the main stakeholders here regarded as Singapore Toto buyers could have a better view on the chance of becoming winner with different prizes. This business report provides different aspects of statistical analysis and model prediction accuracies. Stakeholders can apply those prediction models to narrow their predicting each winning number from a range of 1-49 down to 1-10 groups.

Business Question:

Instead of predicting the exact number from 1-49. Lottery numbers prediction analysis is just to target on predicting the possible groups of each winning number could fall into. 5 numbers are put in one group. There are 49 numbers, therefore they are divided into 10 groups.

Predict the correct group of each winning number can give lottery players a better guide about their number selection.

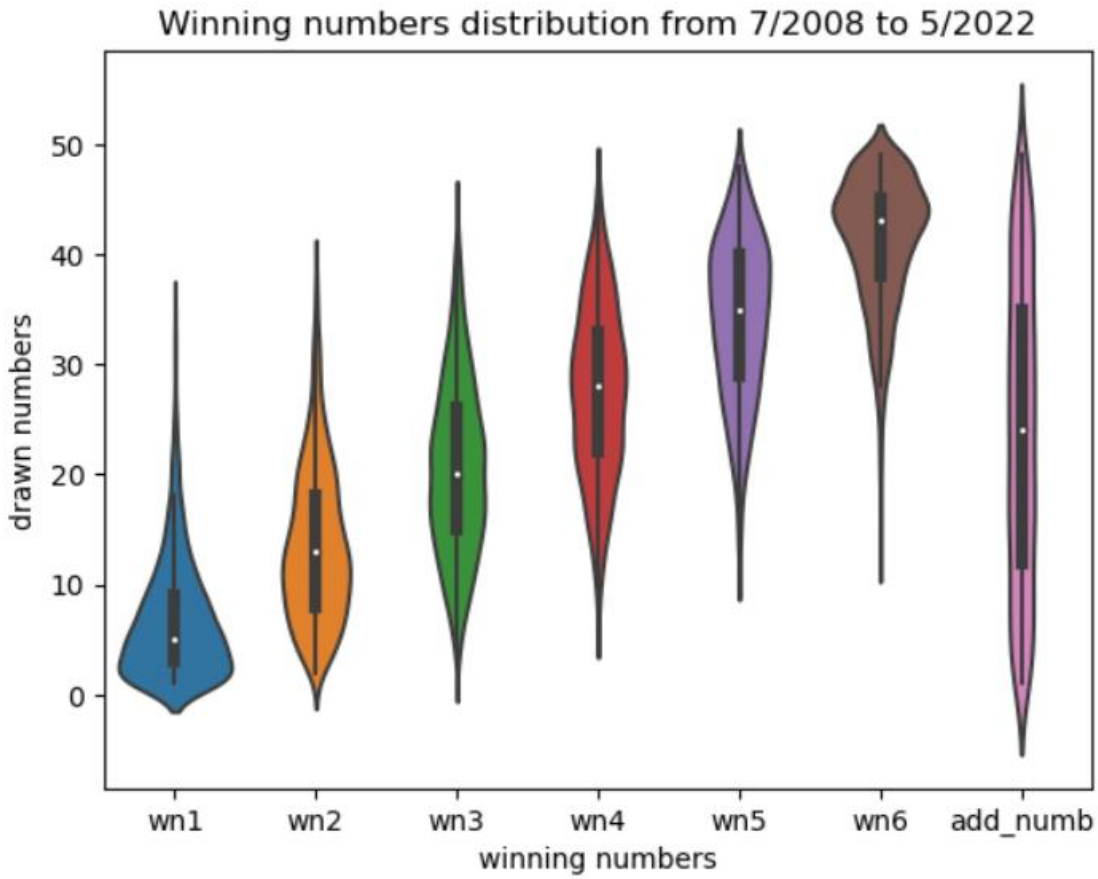
Data

The dataset of Singapore lottery numbers is sourced from Kaggle website. Data is collected from 2008 to 2022 with 1425 observations.

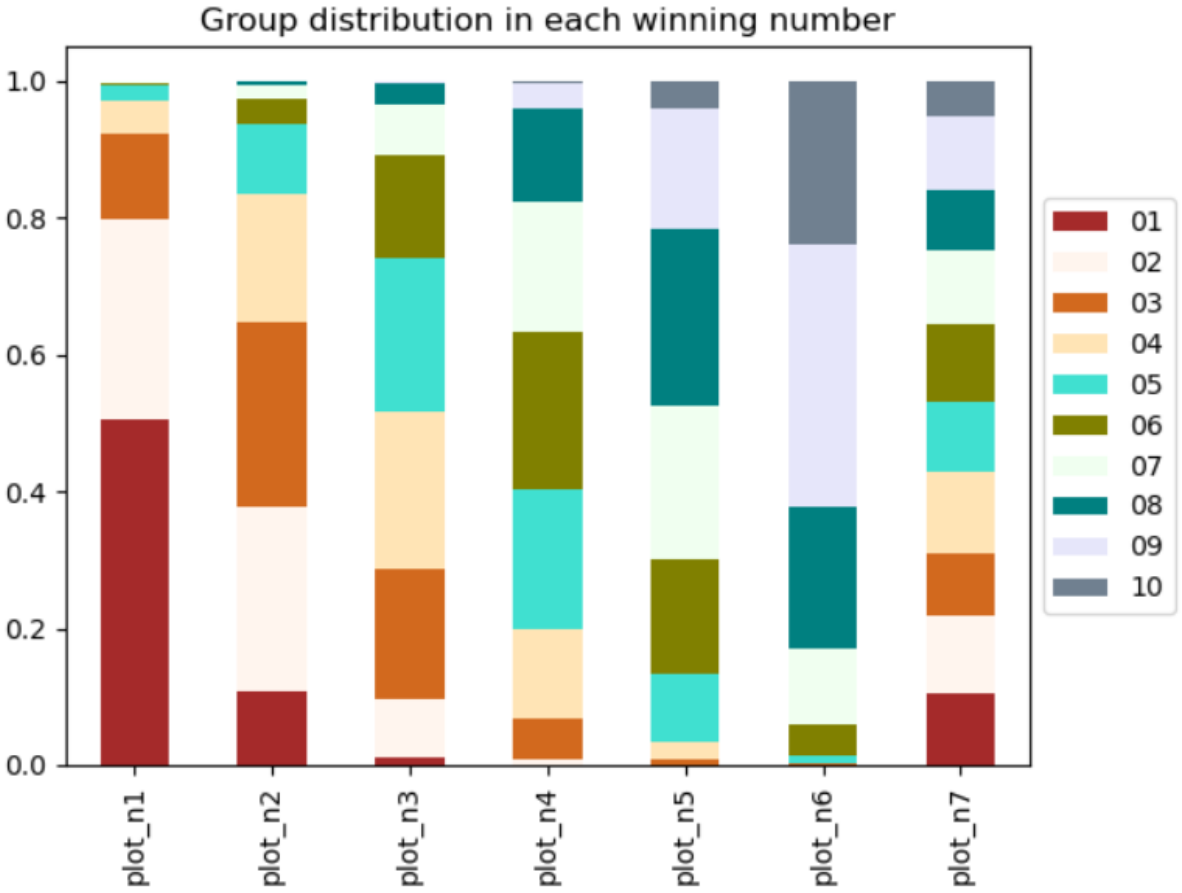
7 features are selected from the csv file corresponding to series of historical winning numbers of 7 numbers.

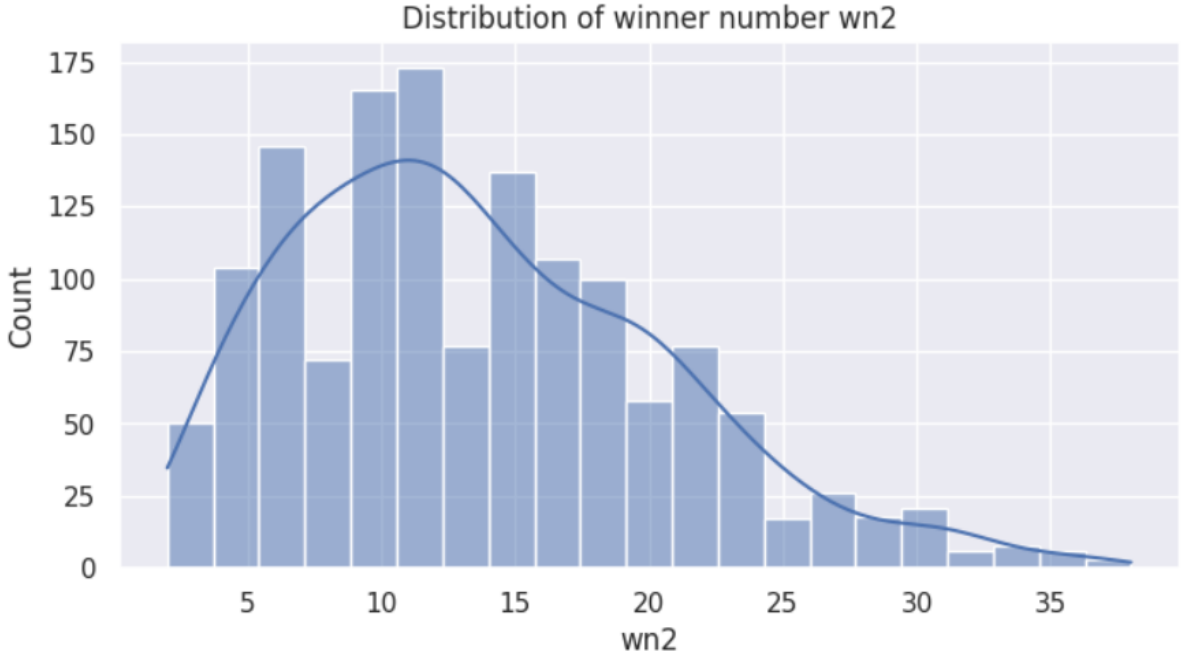
Data Analysis

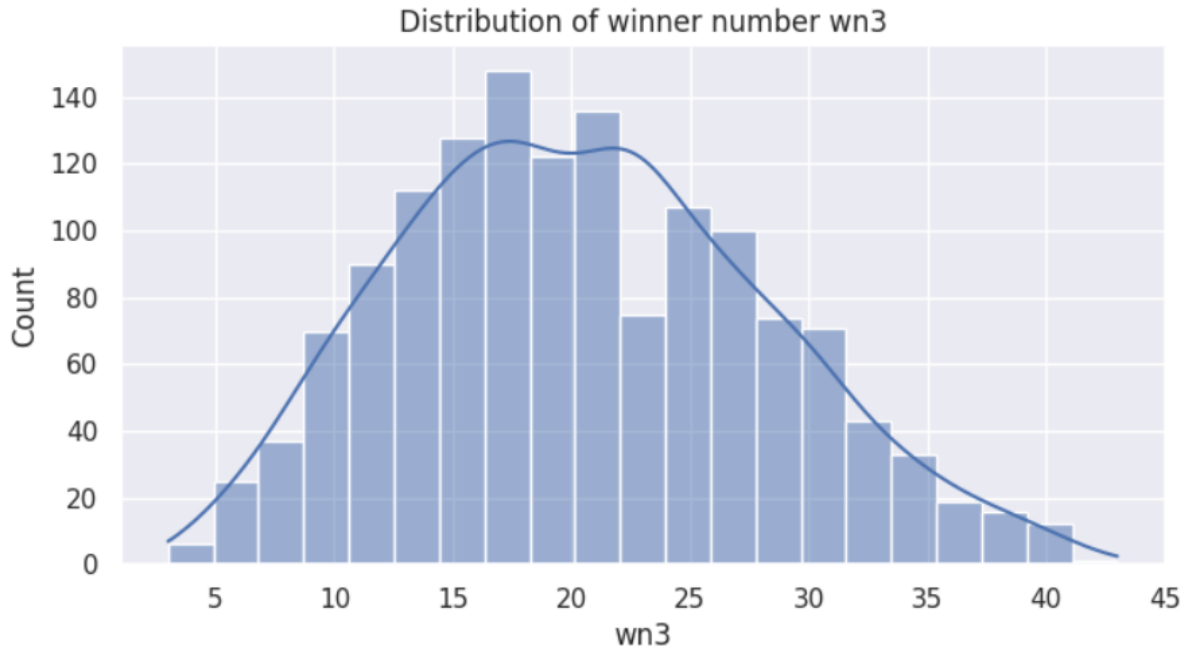
Each winning number is randomly drawn from 1 to 49. However, each of these numbers has significant different mean values as its moving upwards for the latter numbers compared to the former drawn numbers. The additional number seemingly has the most even distribution of numbers 1 to 49.

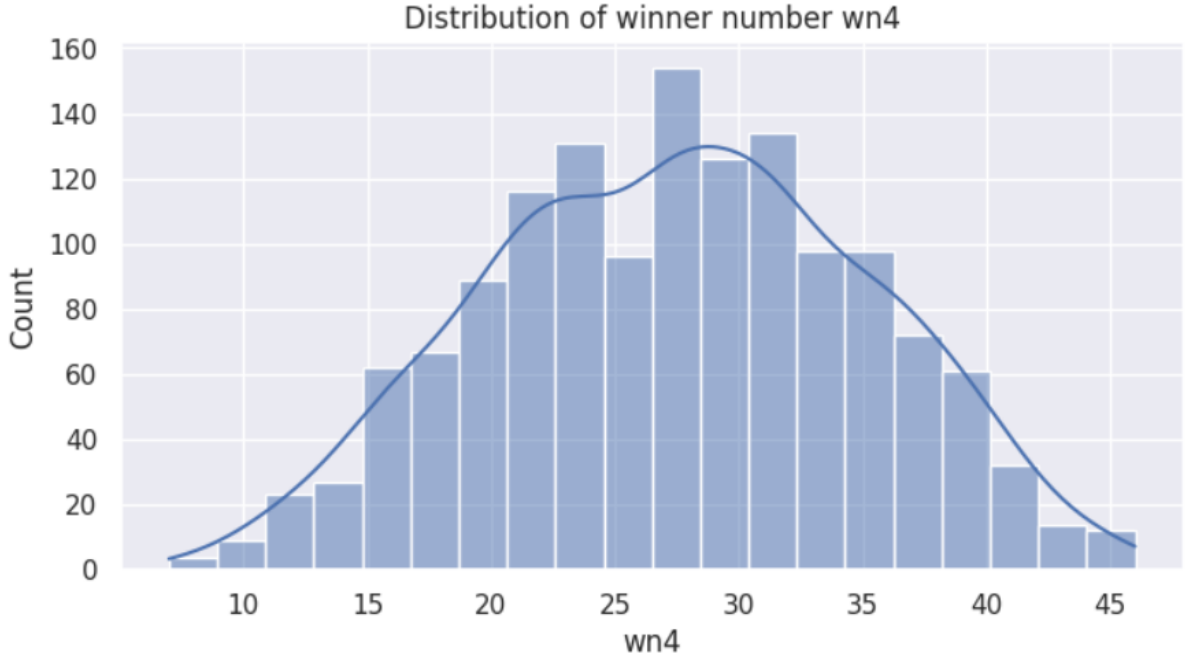


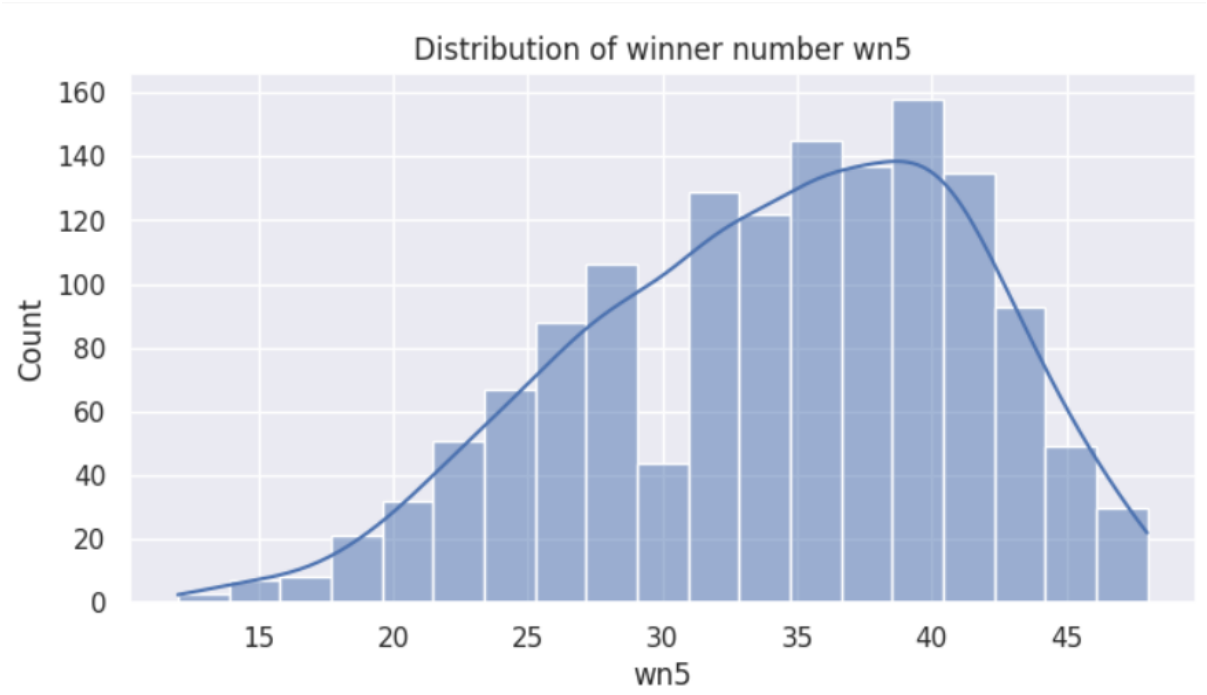
Each winning number has its own favour groups with high frequency of falling into. The most favour groups of this winning number will become the less favour groups of the other winning numbers. Some groups exist in one number but disappear in other numbers.

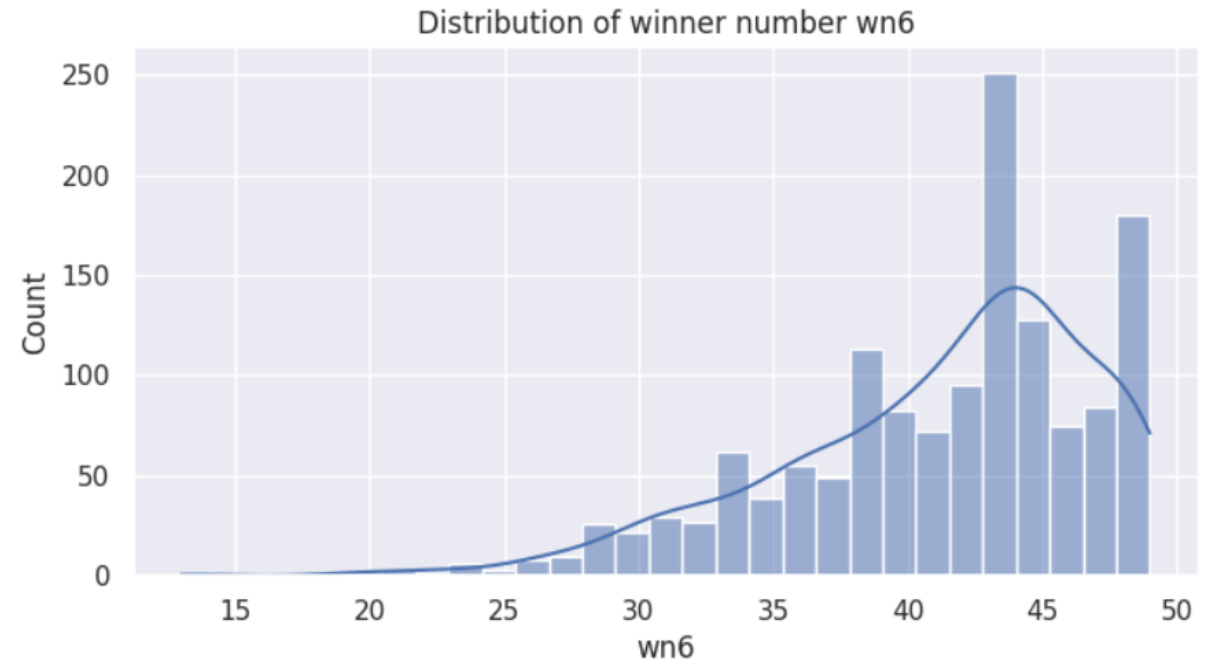

Group distribution in each winning number
1.0
0.8
0.6
0.4
0.2
0.0
plot_n1
plot_n2
plot_n3
plot_n4
plot_n5
plot_n6
plot_n7
01
02
03
04
05
06
07
08
09
10

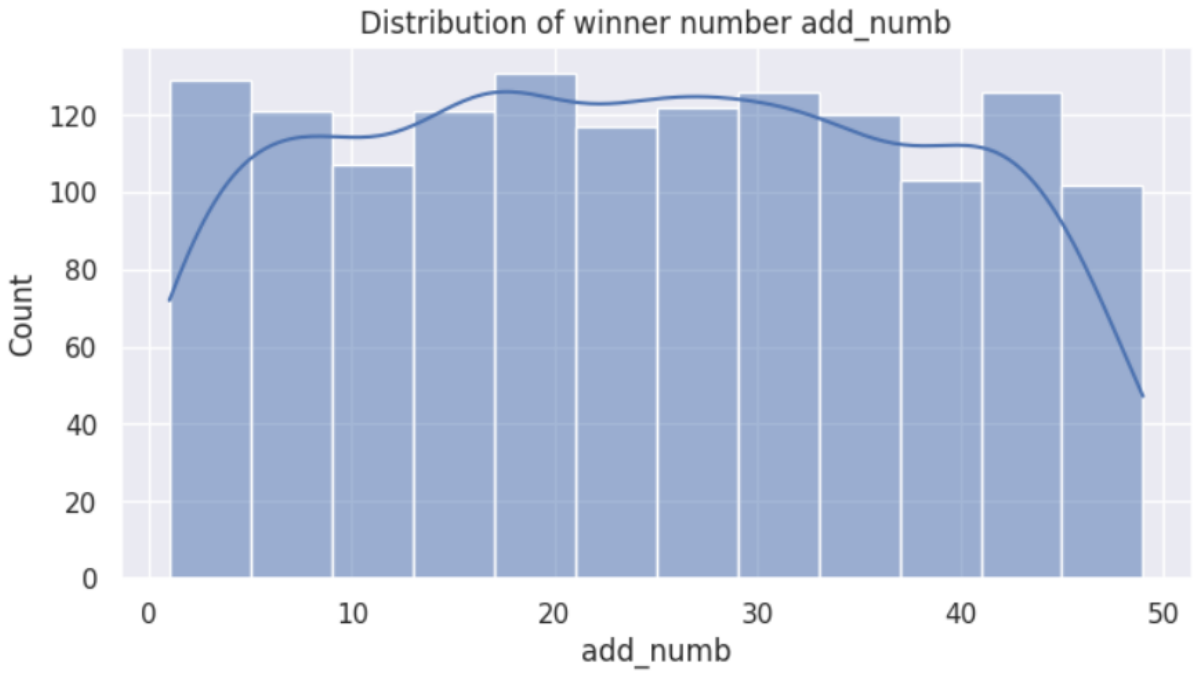

Distribution of winner number wn2
Count
175
150
125
100
75
50
25
0
5
10
15
20
25
30
35
wn2


Distribution of winner number wn3
Count
0
20
40
60
80
100
120
140
5
10
15
20
25
30
35
40
45
wn3


Distribution of winner number wn4
Count
0
20
40
60
80
100
120
140
160
10
15
20
25
30
35
40
45
wn4


Distribution of winner number wn5
Count
wn5
0
20
40
60
80
100
120
140
160
15
20
25
30
35
40
45


Distribution of winner number wn6
Count
wn6
0
50
100
150
200
250
15
20
25
30
35
40
45
50

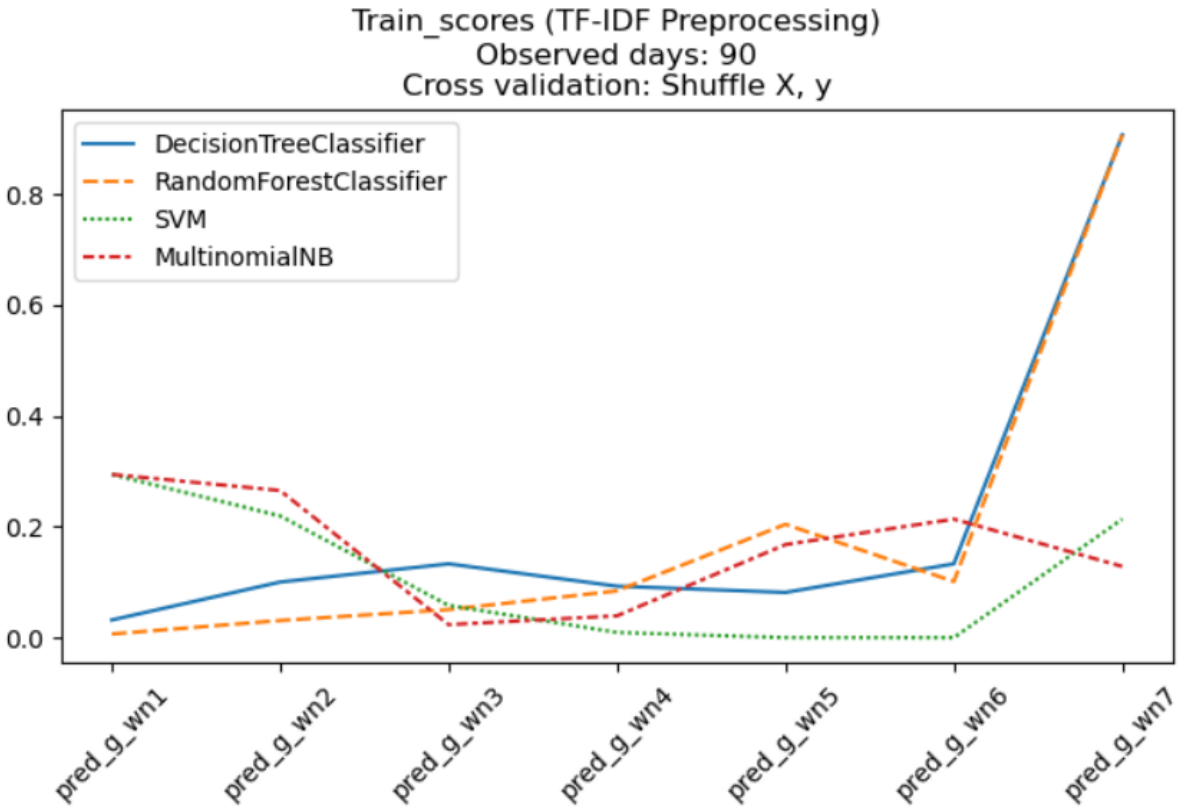


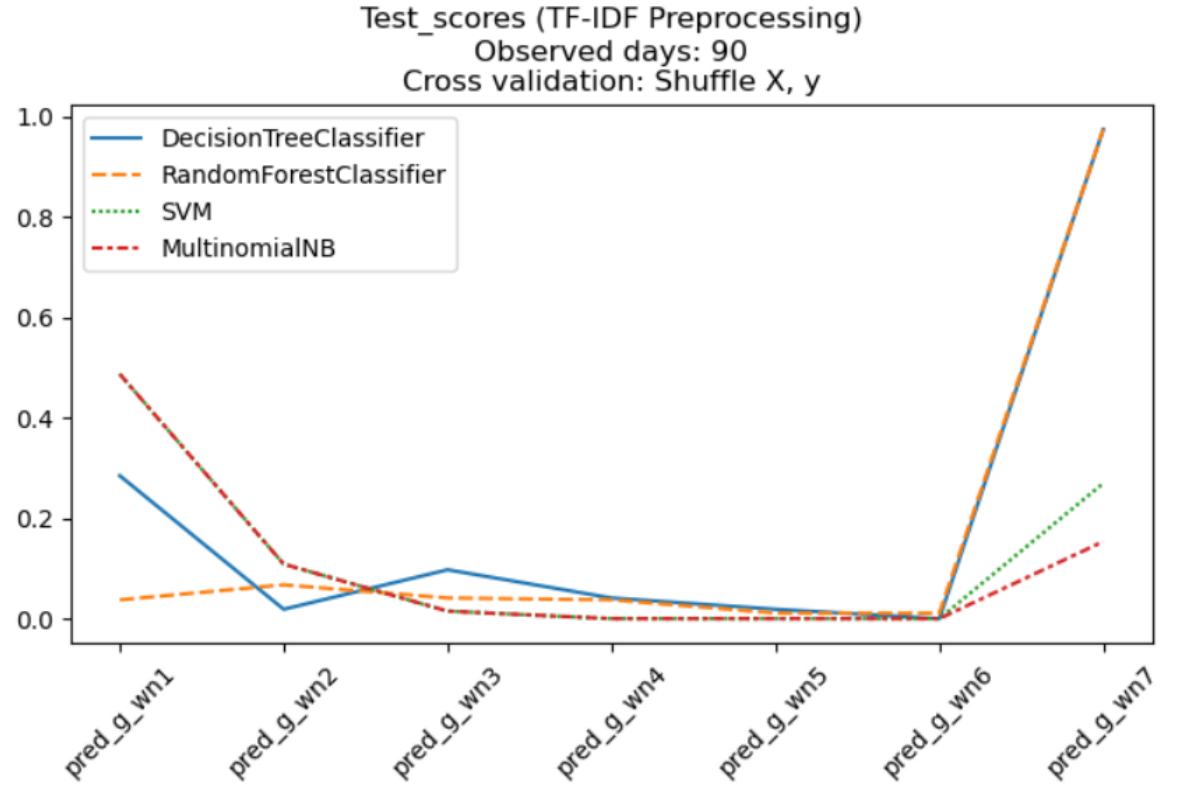
Modelling

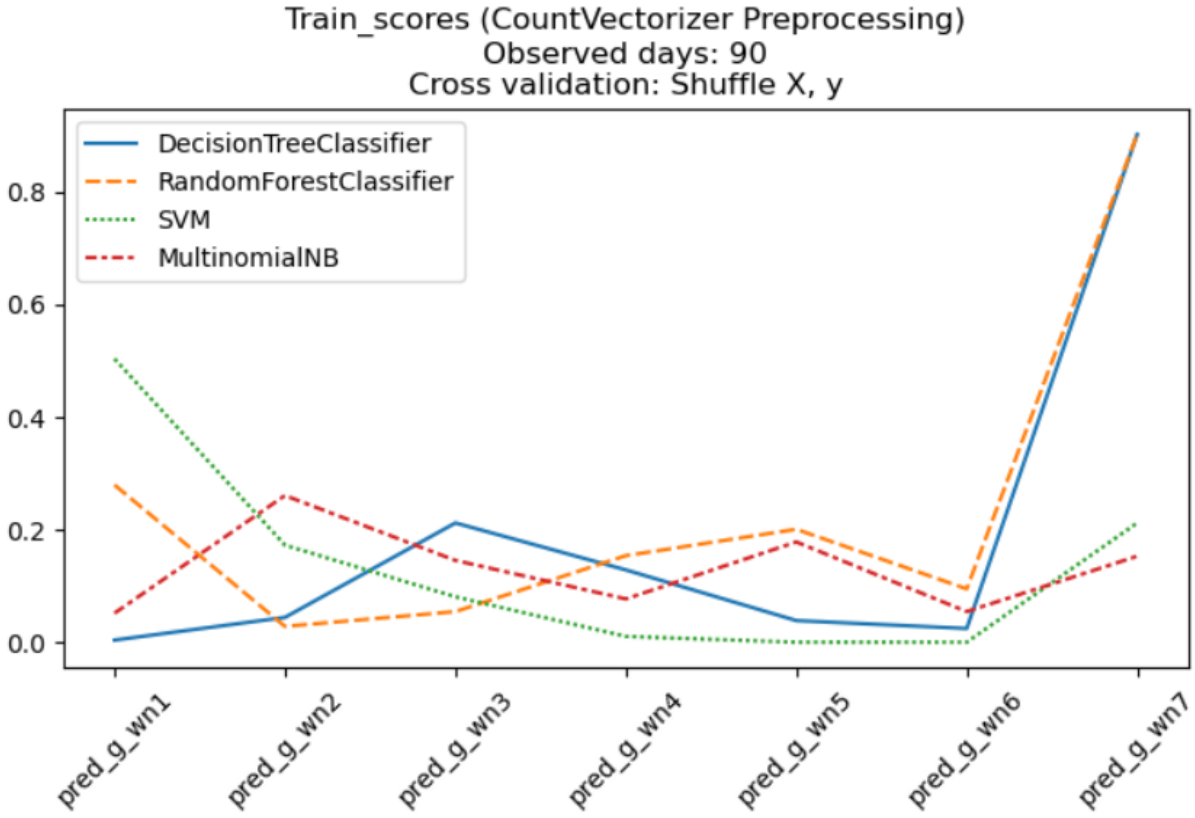
There are 2 different sets of post_processed data. One is processed by TF_IDF model and the other is processed by CountVectorizer model. Each set of data is tested with differetnt scenarios.

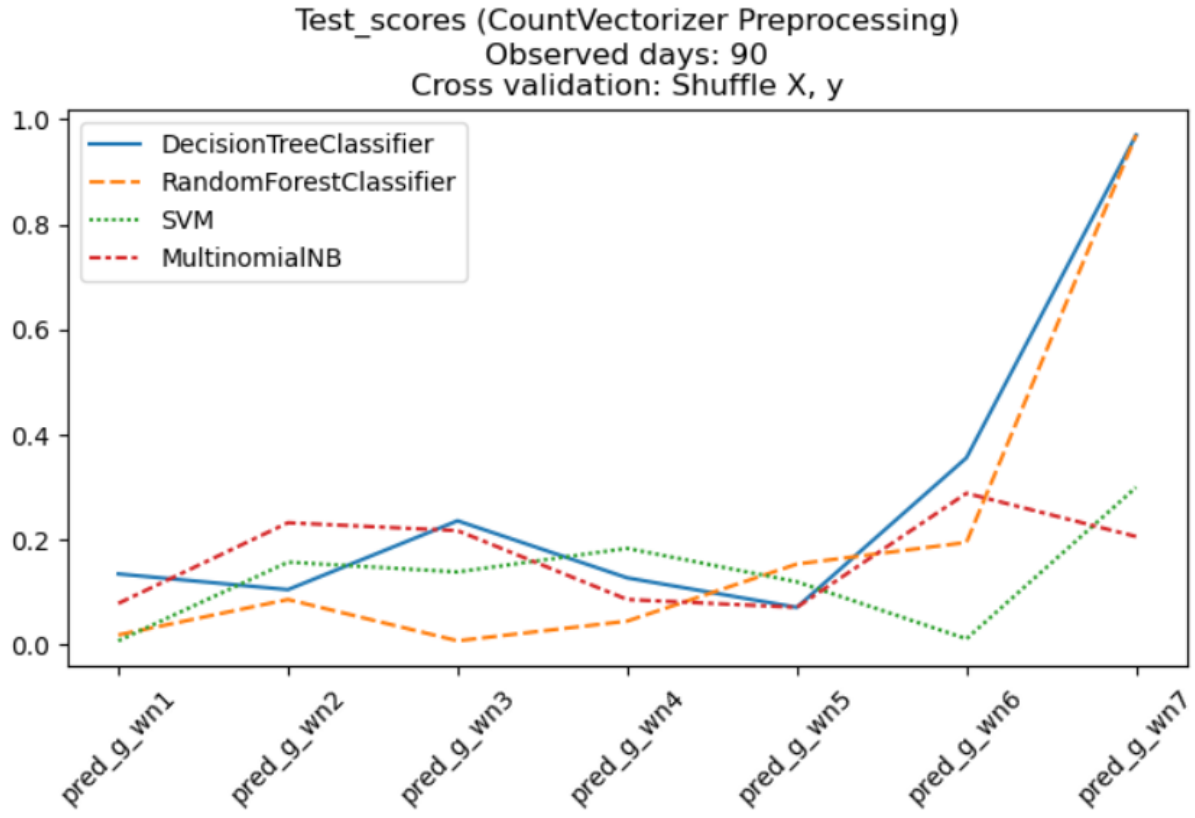
One scenario is tested with 3 different number of observed days of historical data. There are 3 options 90, 120 and 180 days.

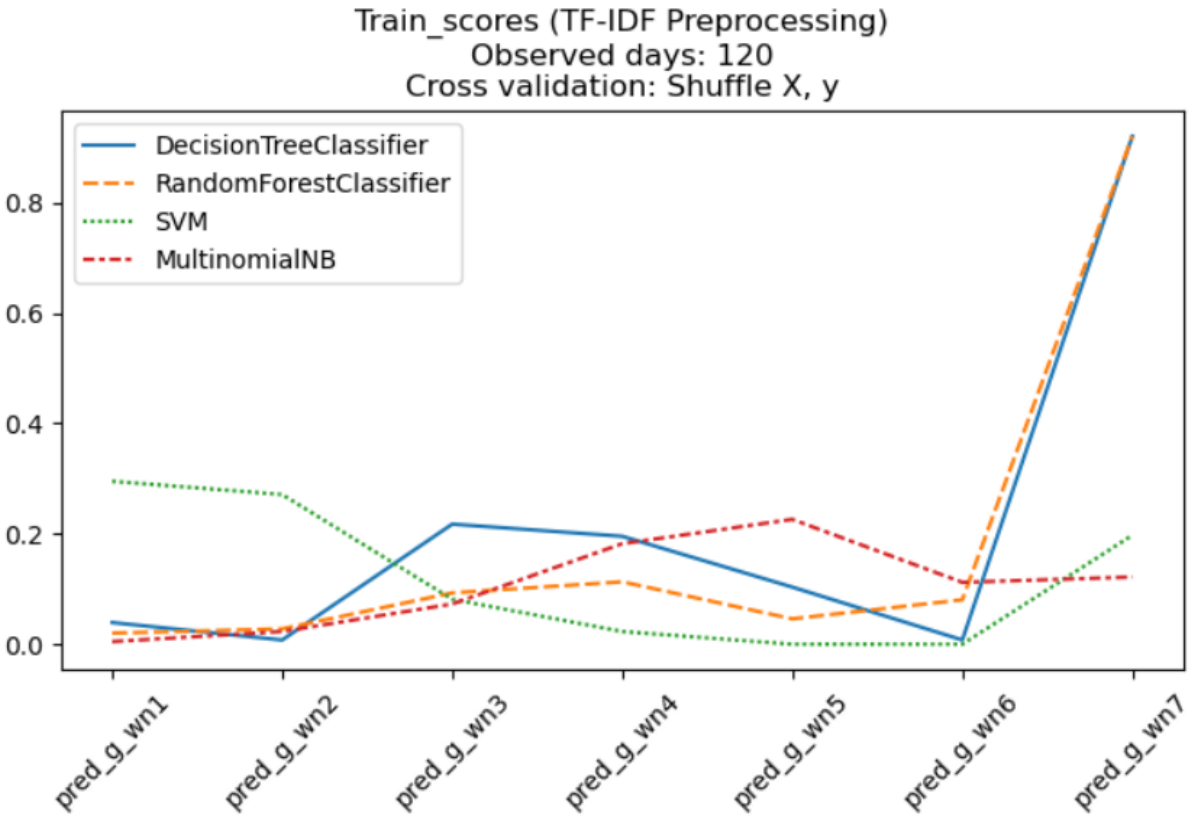
Second scenario is tested with cross validation which sets shuffling Xs but not shuffle ys. Doing so we can check if the predicted values of y is randomly determined by the frequency factor occurred some point in the past not necessary the most updated frequency.

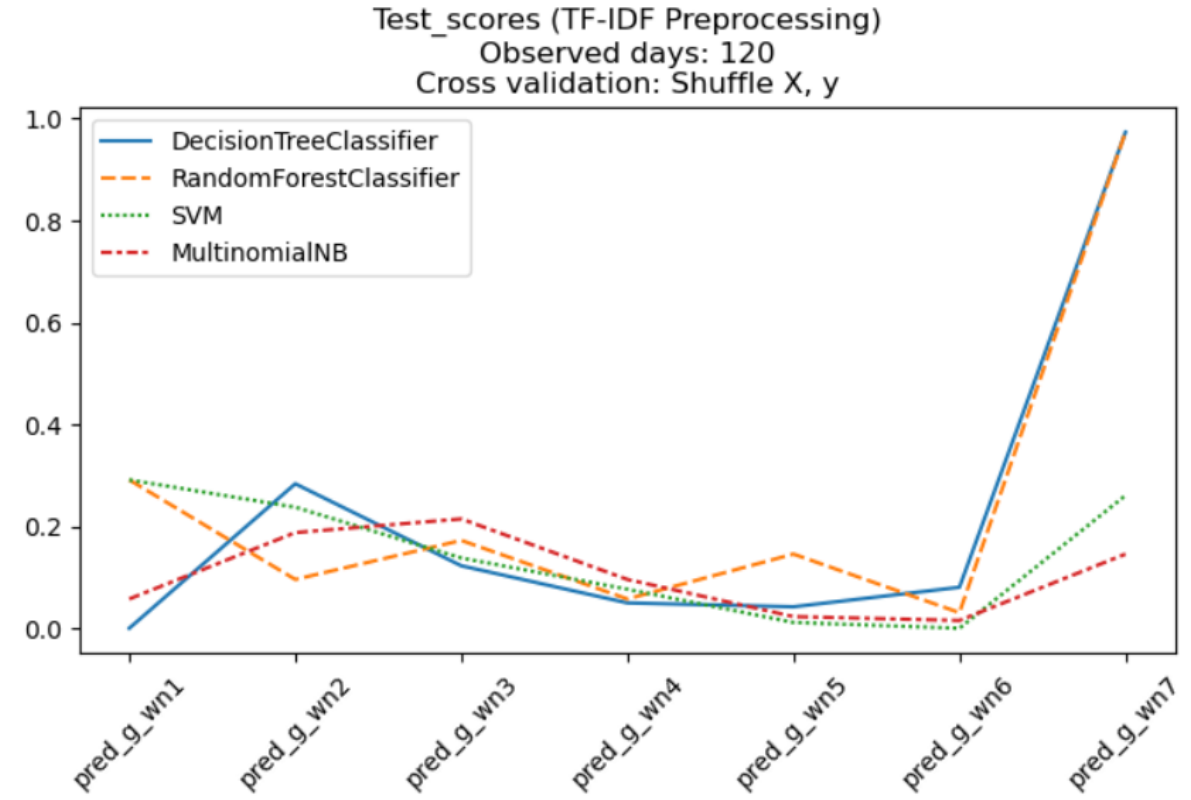

Train_scores (TF-IDF Preprocessing)
Observed days: 90
Cross validation: Shuffle X, y
DecisionTreeClassifier
RandomForestClassifier
SVM
MultinomialNB
0.0
0.2
0.4
0.6
0.8
pred_g_wn1
pred_g_wn2
pred_g_wn3
pred_g_wn4
pred_g_wn5
pred_g_wn6
pred_g_wn7

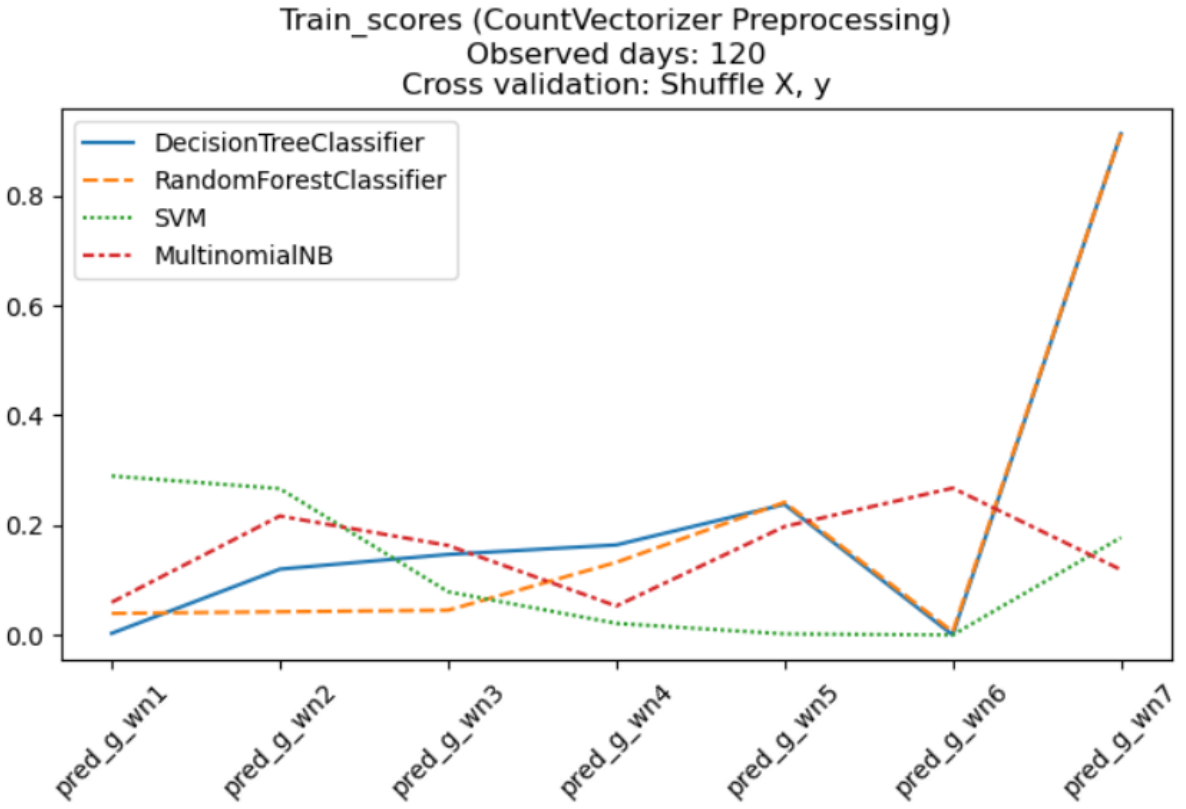

Test_scores (TF-IDF Preprocessing)
Observed days: 90
Cross validation: Shuffle X, y
DecisionTreeClassifier
RandomForestClassifier
SVM
MultinomialNB
0.0
0.2
0.4
0.6
0.8
1.0
pred_g_wn1
pred_g_wn2
pred_g_wn3
pred_g_wn4
pred_g_wn5
pred_g_wn6
pred_g_wn7

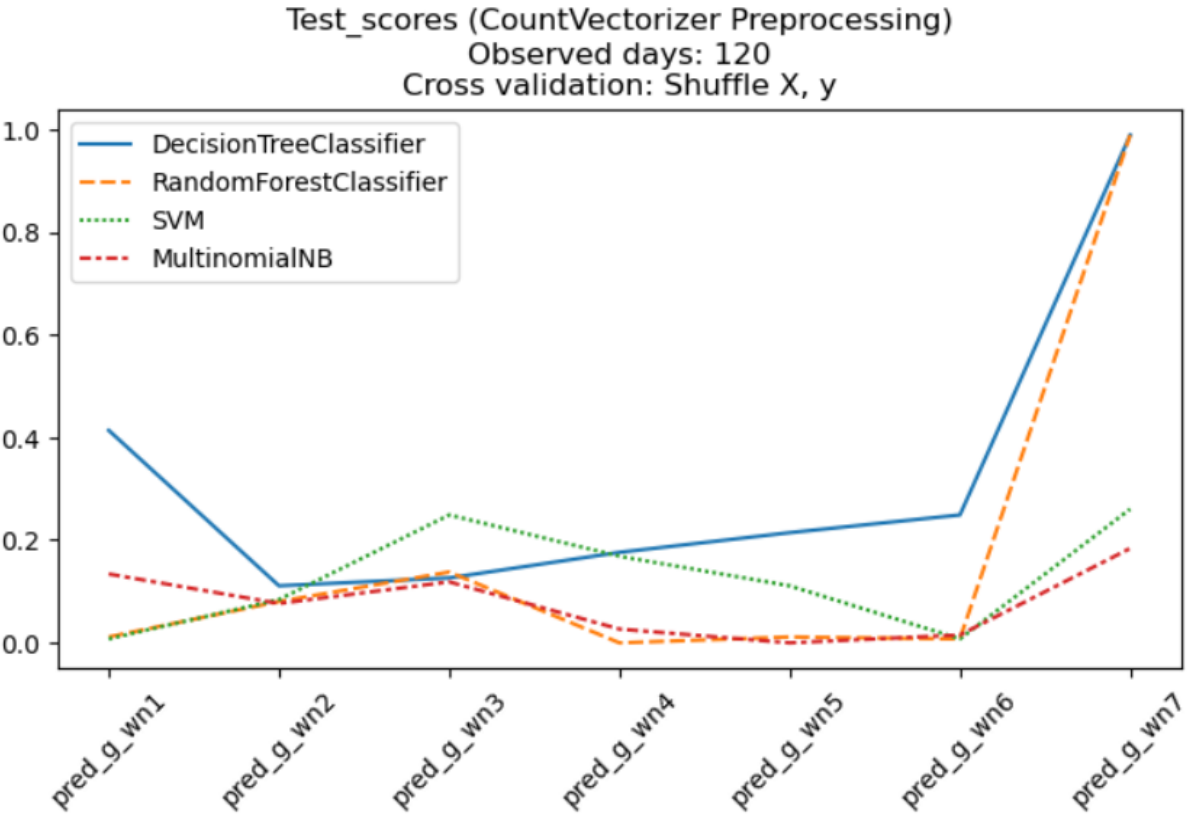

Train_scores (CountVectorizer Preprocessing)
Observed days: 90
Cross validation: Shuffle X, y
DecisionTreeClassifier
RandomForestClassifier
SVM
MultinomialNB
0.0
0.2
0.4
0.6
0.8
pred_g_wn1
pred_g_wn2
pred_g_wn3
pred_g_wn4
pred_g_wn5
pred_g_wn6
pred_g_wn7

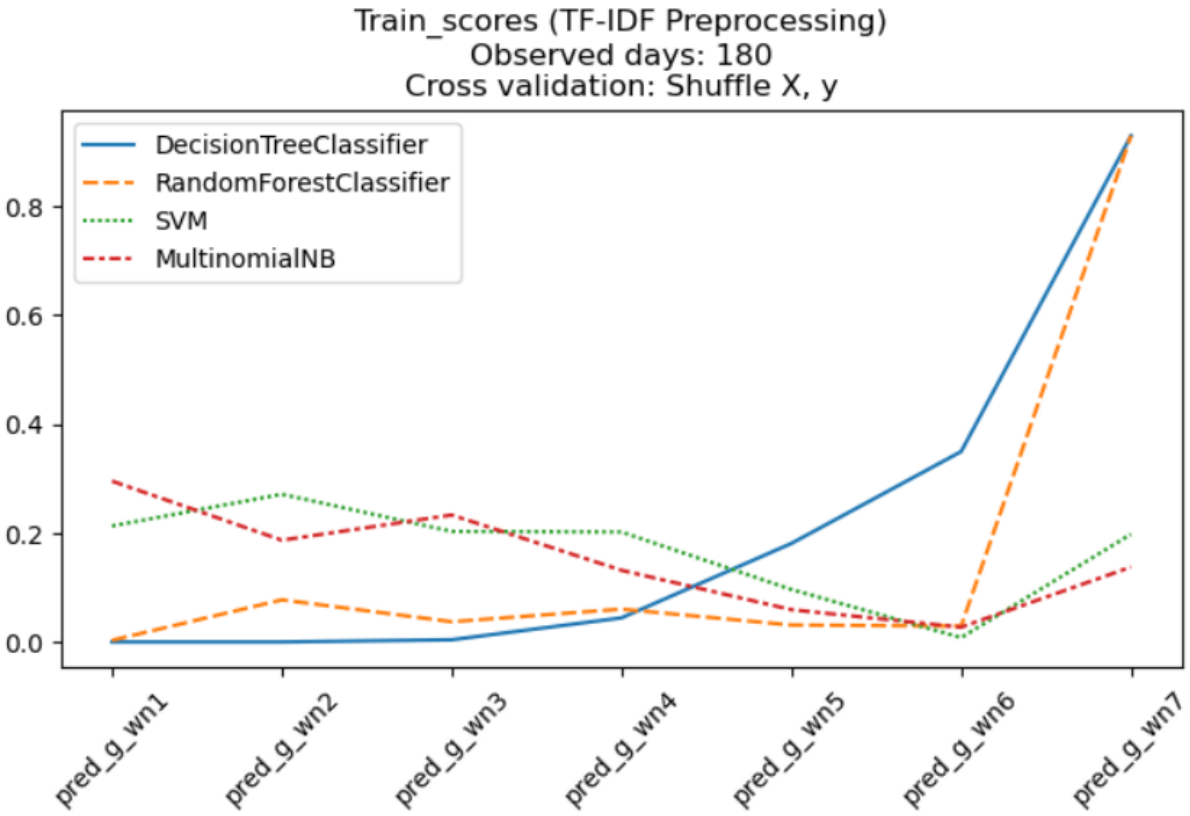

Test_scores (CountVectorizer Preprocessing)
Observed days: 90
Cross validation: Shuffle X, y
DecisionTreeClassifier
RandomForestClassifier
SVM
MultinomialNB
0.0
0.2
0.4
0.6
0.8
1.0
pred_g_wn1
pred_g_wn2
pred_g_wn3
pred_g_wn4
pred_g_wn5
pred_g_wn6
pred_g_wn7

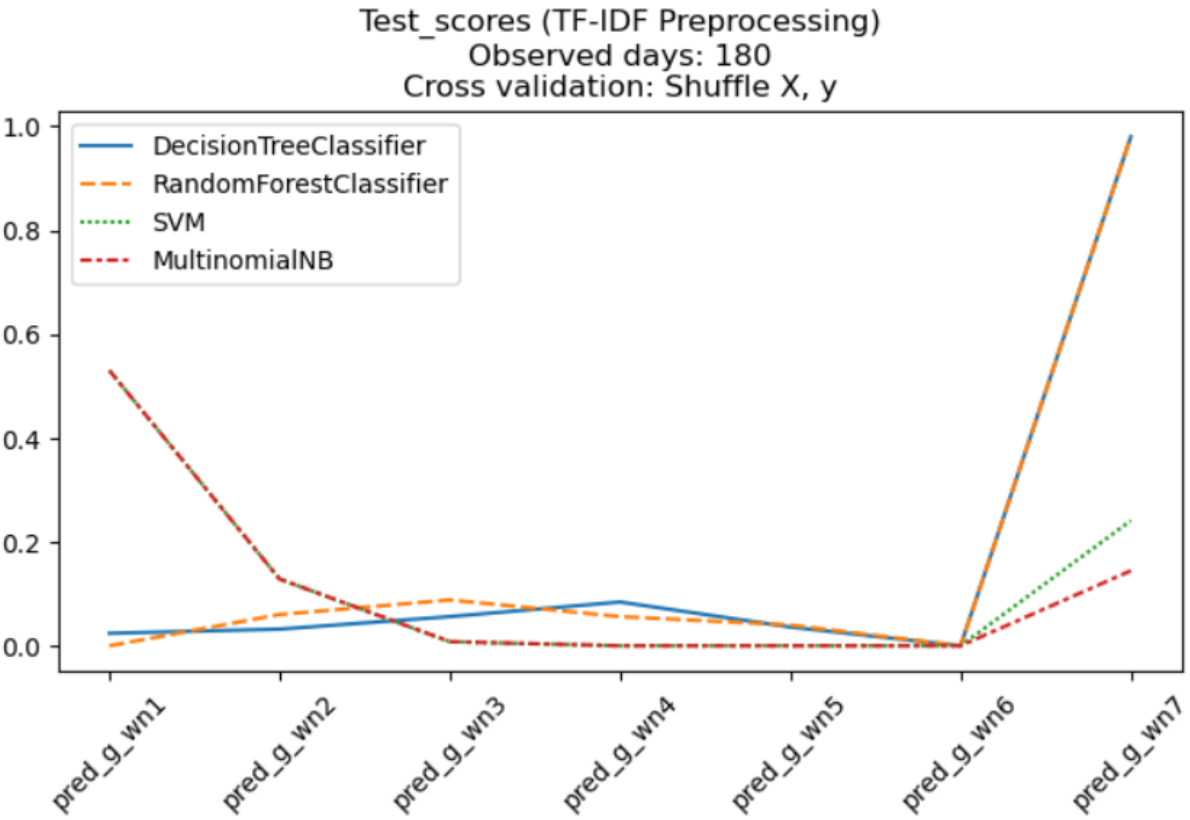

Train_scores (TF-IDF Preprocessing)
Observed days: 120
Cross validation: Shuffle X, y
DecisionTreeClassifier
RandomForestClassifier
SVM
MultinomialNB
0.0
0.2
0.4
0.6
0.8
pred_g_wn1
pred_g_wn2
pred_g_wn3
pred_g_wn4
pred_g_wn5
pred_g_wn6
pred_g_wn7

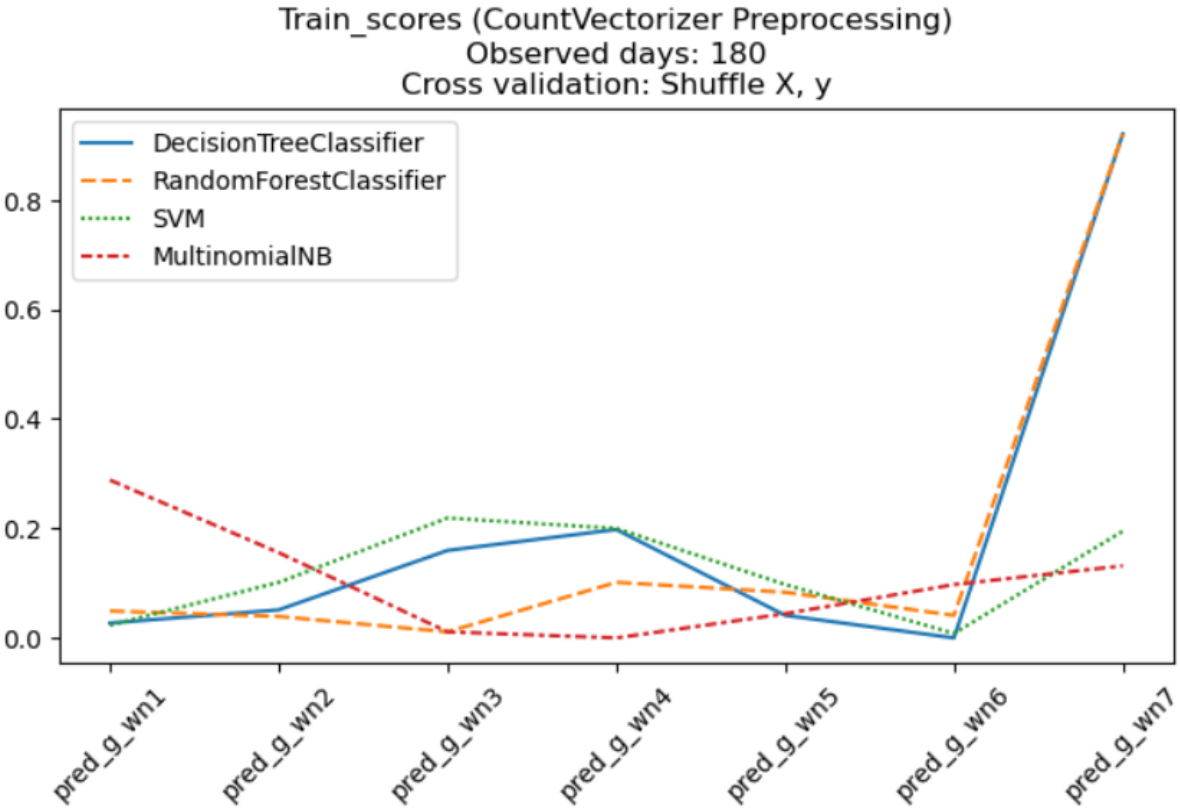

Test_scores (TF-IDF Preprocessing)
Observed days: 120
Cross validation: Shuffle X, y
DecisionTreeClassifier
RandomForestClassifier
SVM
MultinomialNB
0.0
0.2
0.4
0.6
0.8
1.0
pred_g_wn1
pred_g_wn2
pred_g_wn3
pred_g_wn4
pred_g_wn5
pred_g_wn6
pred_g_wn7

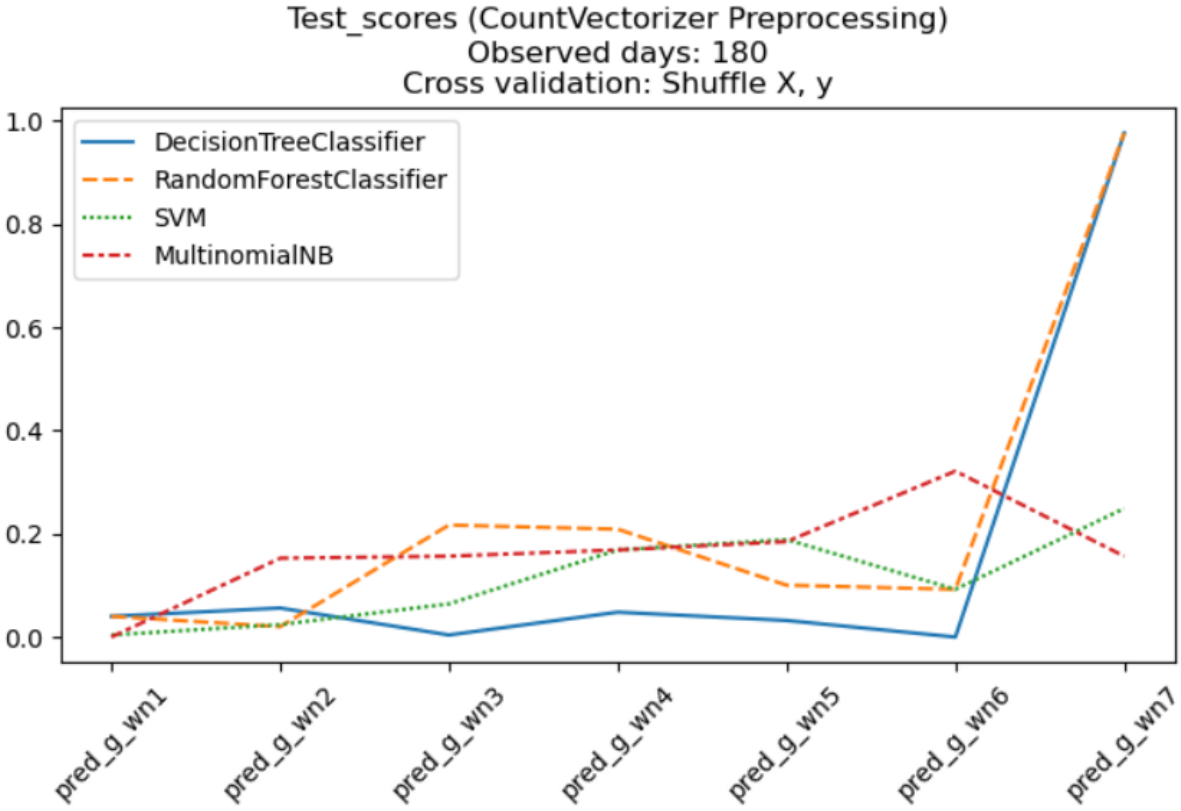

Train_scores (CountVectorizer Preprocessing)
Observed days: 120
Cross validation: Shuffle X, y
DecisionTreeClassifier
RandomForestClassifier
SVM
MultinomialNB
0.0
0.2
0.4
0.6
0.8
pred_g_wn1
pred_g_wn2
pred_g_wn3
pred_g_wn4
pred_g_wn5
pred_g_wn6
pred_g_wn7

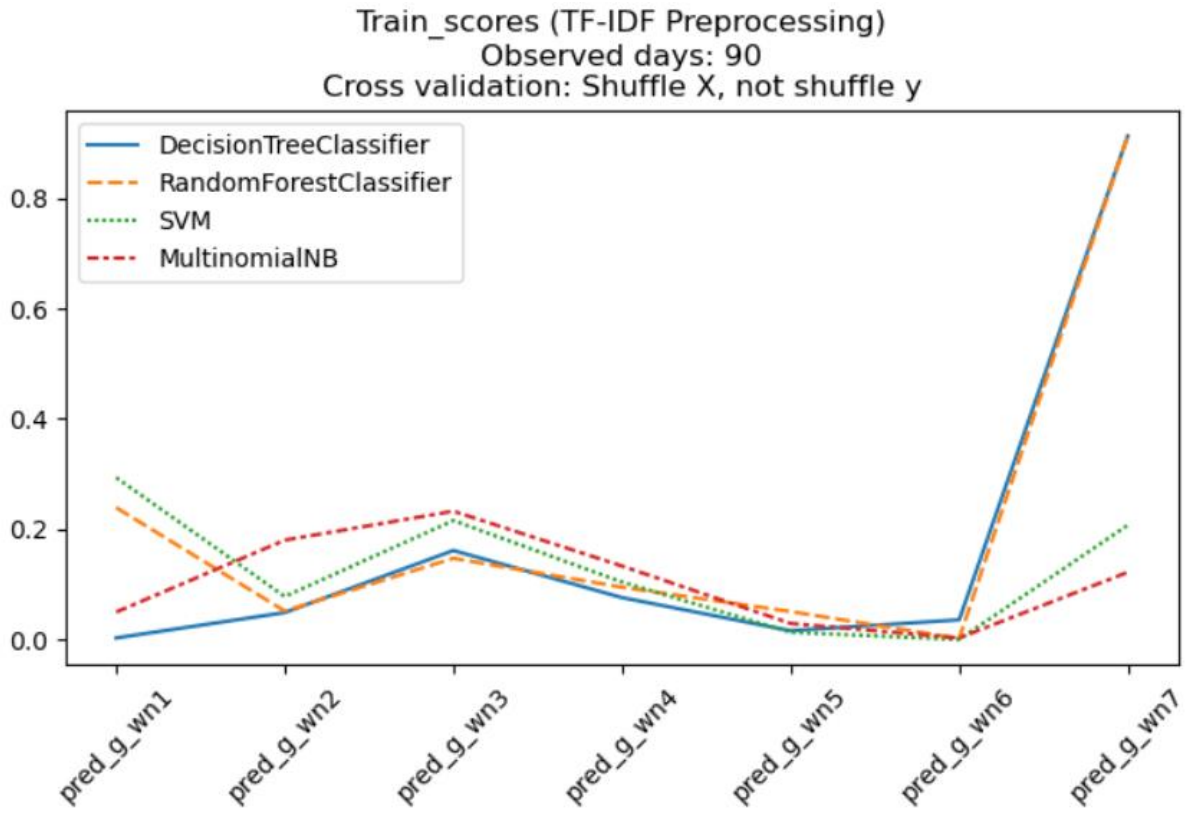

Test_scores (CountVectorizer Preprocessing)
Observed days: 120
Cross validation: Shuffle X, y
DecisionTreeClassifier
RandomForestClassifier
SVM
MultinomialNB
0.0
0.2
0.4
0.6
0.8
1.0
pred_g_wn1
pred_g_wn2
pred_g_wn3
pred_g_wn4
pred_g_wn5
pred_g_wn6
pred_g_wn7

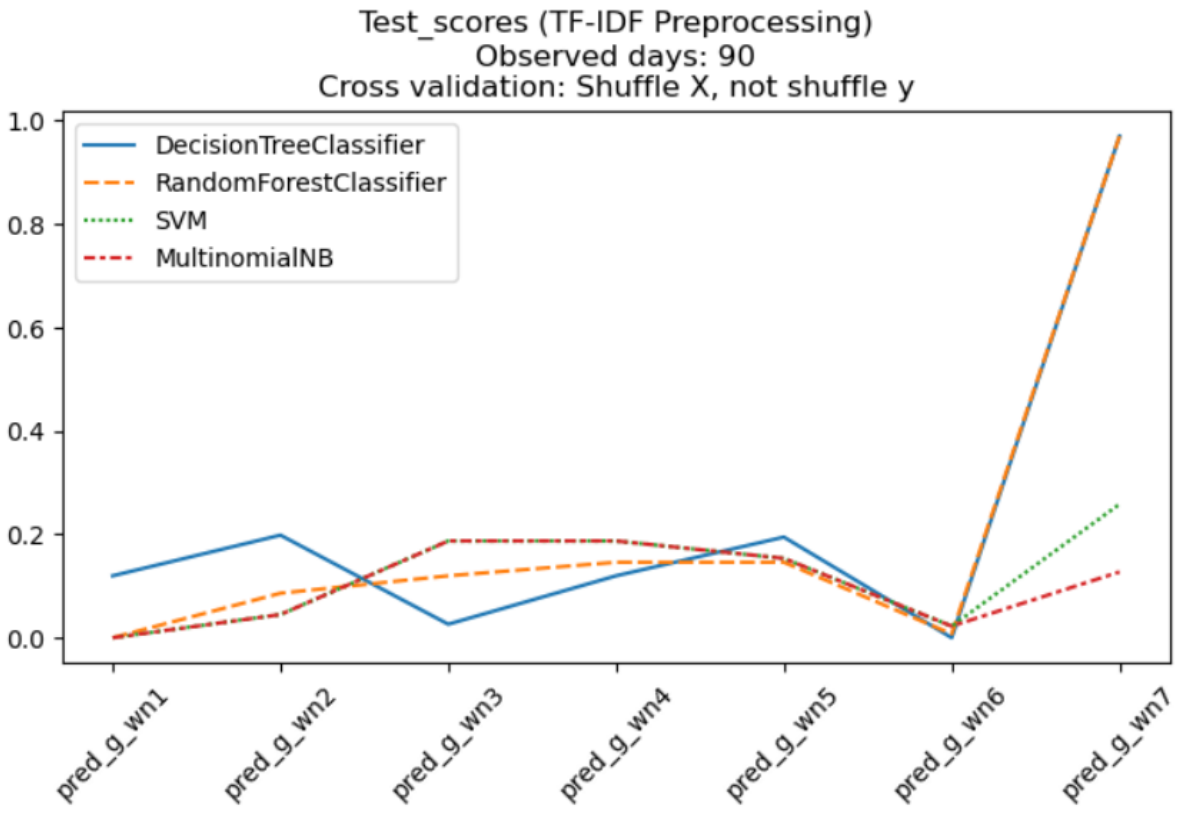

Train_scores (TF-IDF Preprocessing)
Observed days: 180
Cross validation: Shuffle X, y
DecisionTreeClassifier
RandomForestClassifier
SVM
MultinomialNB
0.0
0.2
0.4
0.6
0.8
pred_g_wn1
pred_g_wn2
pred_g_wn3
pred_g_wn4
pred_g_wn5
pred_g_wn6
pred_g_wn7

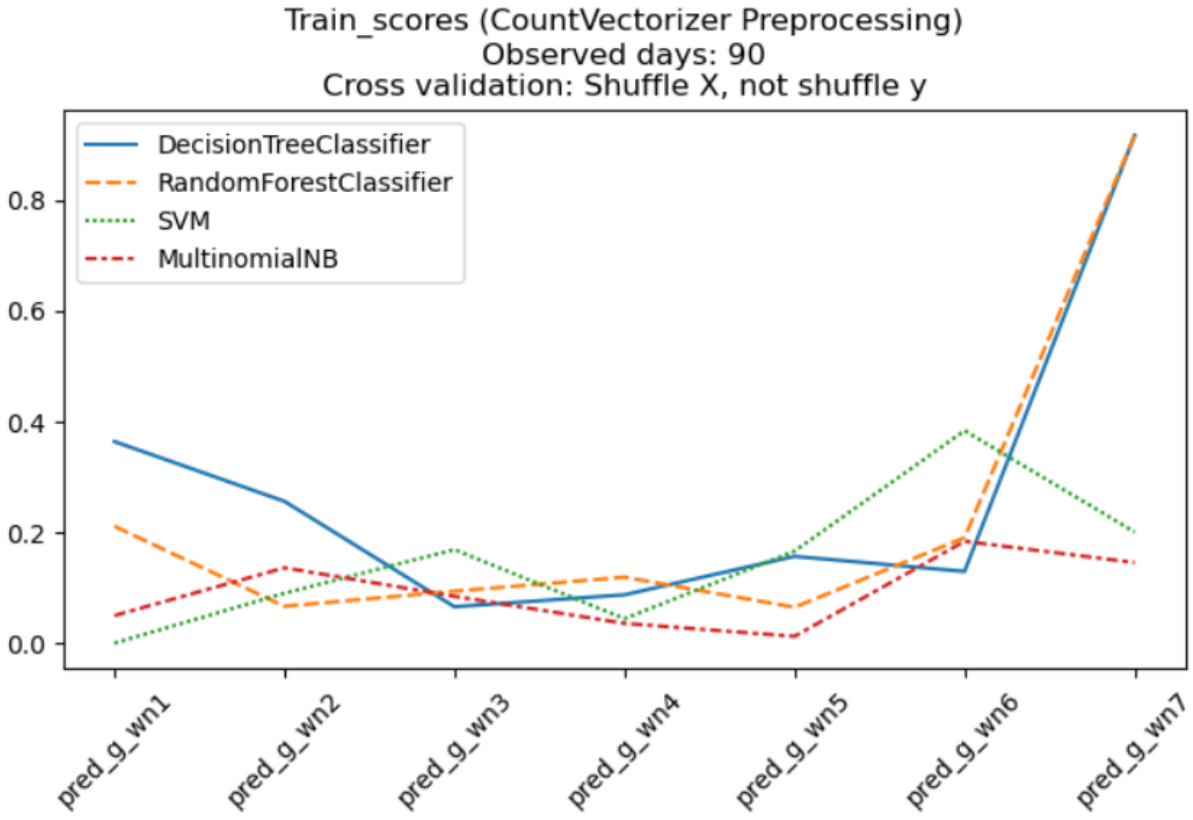

Test_scores (TF-IDF Preprocessing)
Observed days: 180
Cross validation: Shuffle X, y
DecisionTreeClassifier
RandomForestClassifier
SVM
MultinomialNB
0.0
0.2
0.4
0.6
0.8
1.0
pred_g_wn1
pred_g_wn2
pred_g_wn3
pred_g_wn4
pred_g_wn5
pred_g_wn6
pred_g_wn7

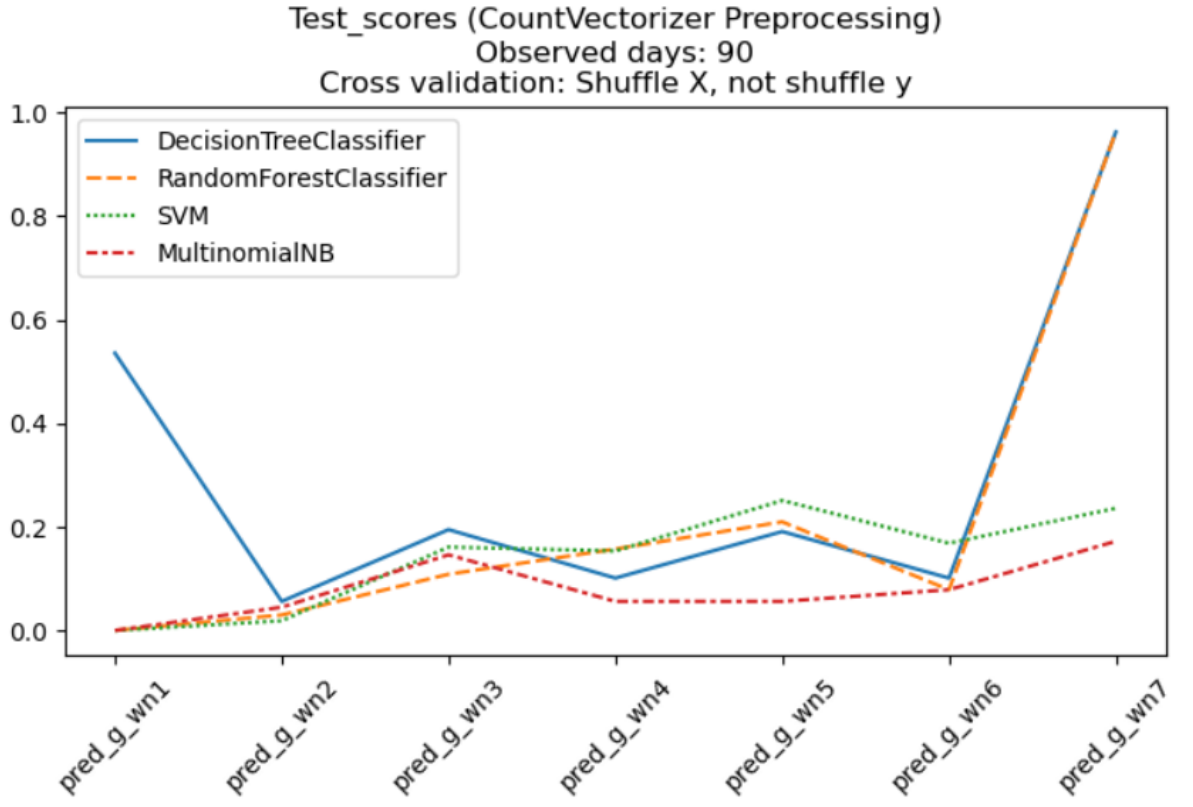

Train_scores (CountVectorizer Preprocessing)
Observed days: 180
Cross validation: Shuffle X, y
DecisionTreeClassifier
RandomForestClassifier
SVM
MultinomialNB
0.0
0.2
0.4
0.6
0.8
pred_g_wn1
pred_g_wn2
pred_g_wn3
pred_g_wn4
pred_g_wn5
pred_g_wn6
pred_g_wn7

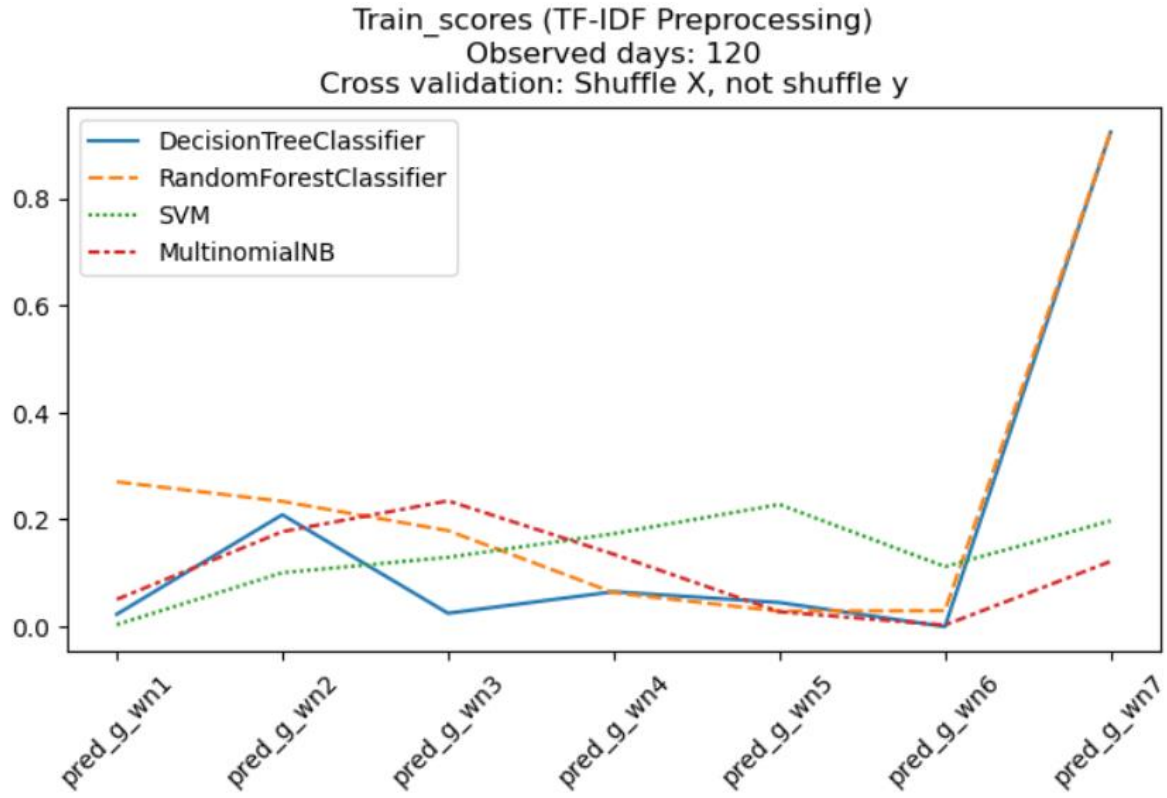

Test_scores (CountVectorizer Preprocessing)
Observed days: 180
Cross validation: Shuffle X, y
DecisionTreeClassifier
RandomForestClassifier
SVM
MultinomialNB
0.0
0.2
0.4
0.6
0.8
1.0
pred_g_wn1
pred_g_wn2
pred_g_wn3
pred_g_wn4
pred_g_wn5
pred_g_wn6
pred_g_wn7

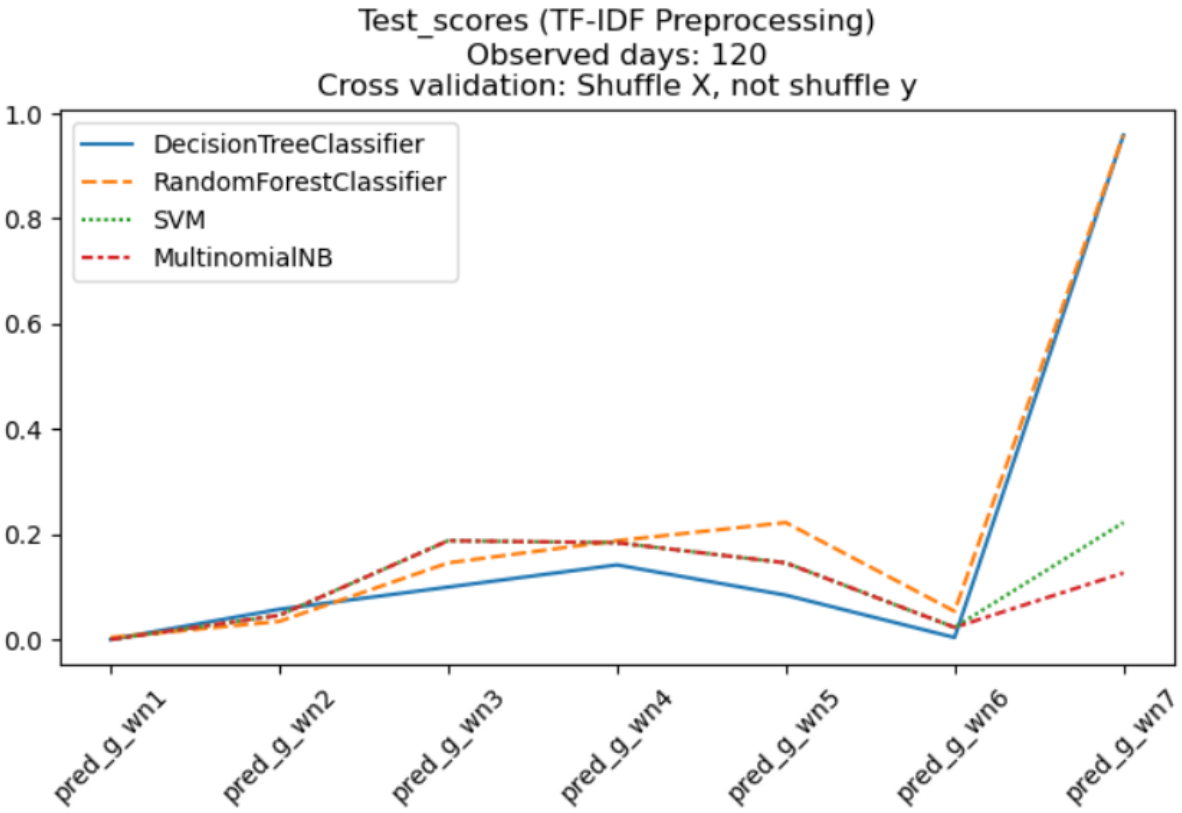

Train_scores (TF-IDF Preprocessing)
Observed days: 90
Cross validation: Shuffle X, not shuffle y
DecisionTreeClassifier
RandomForestClassifier
SVM
MultinomialNB
0.0
0.2
0.4
0.6
0.8
pred_g_wn1
pred_g_wn2
pred_g_wn3
pred_g_wn4
pred_g_wn5
pred_g_wn6
pred_g_wn7

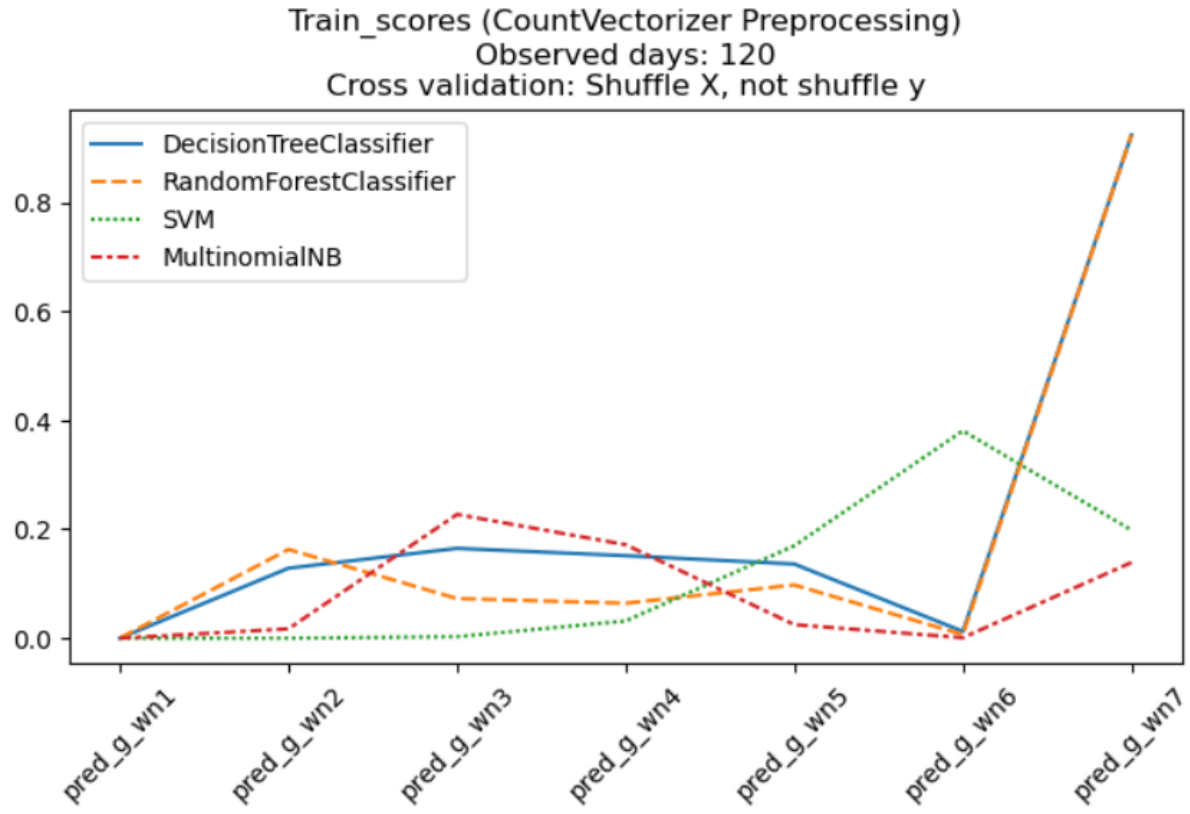

Test_scores (TF-IDF Preprocessing)
Observed days: 90
Cross validation: Shuffle X, not shuffle y
DecisionTreeClassifier
RandomForestClassifier
SVM
MultinomialNB
0.0
0.2
0.4
0.6
0.8
1.0
pred_g_wn1
pred_g_wn2
pred_g_wn3
pred_g_wn4
pred_g_wn5
pred_g_wn6
pred_g_wn7

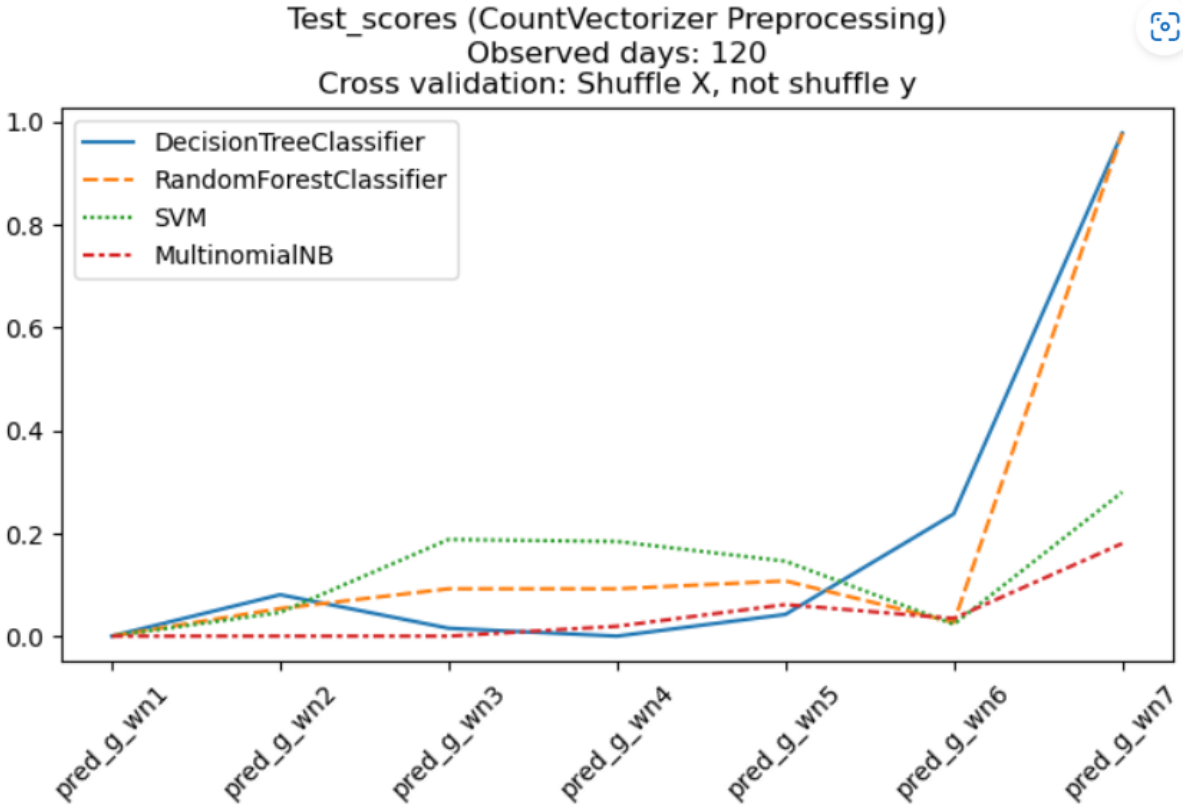

Train_scores (CountVectorizer Preprocessing)
Observed days: 90
Cross validation: Shuffle X, not shuffle y
DecisionTreeClassifier
RandomForestClassifier
SVM
MultinomialNB
0.0
0.2
0.4
0.6
0.8
pred_g_wn1
pred_g_wn2
pred_g_wn3
pred_g_wn4
pred_g_wn5
pred_g_wn6
pred_g_wn7

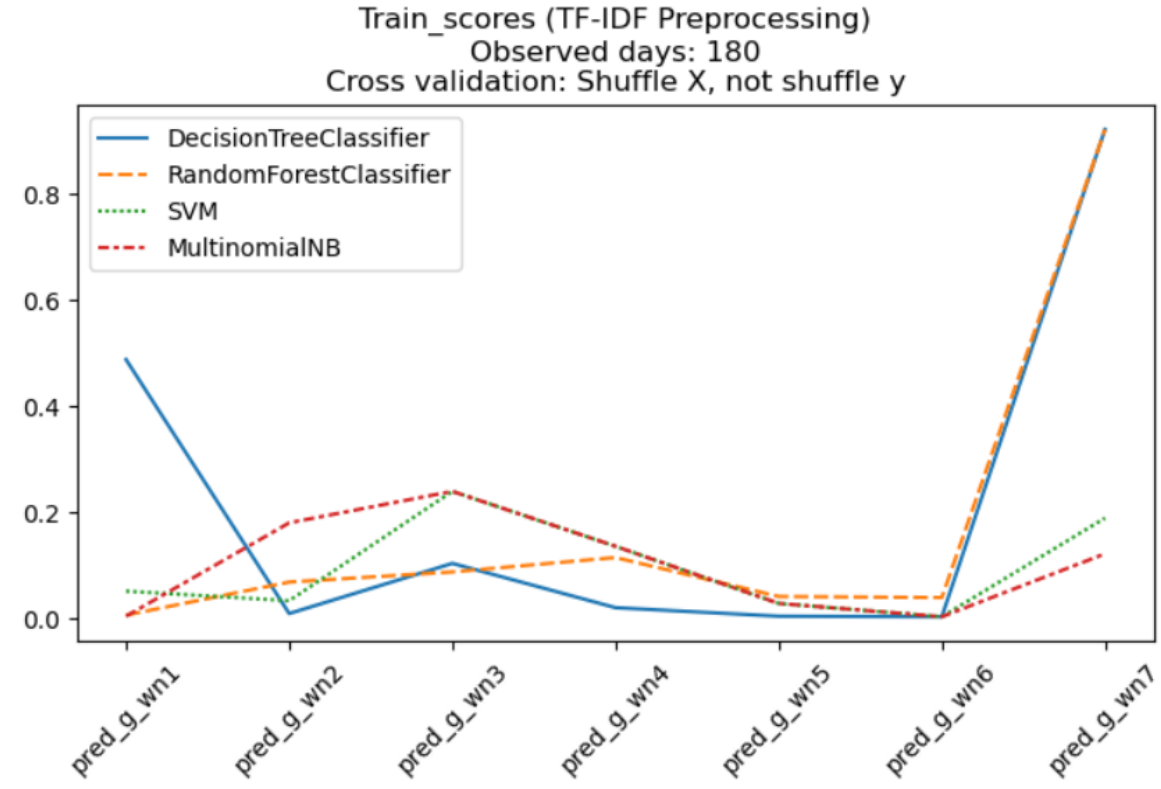

Test_scores (CountVectorizer Preprocessing)
Observed days: 90
Cross validation: Shuffle X, not shuffle y
DecisionTreeClassifier
RandomForestClassifier
SVM
MultinomialNB
0.0
0.2
0.4
0.6
0.8
1.0
pred_g_wn1
pred_g_wn2
pred_g_wn3
pred_g_wn4
pred_g_wn5
pred_g_wn6
pred_g_wn7

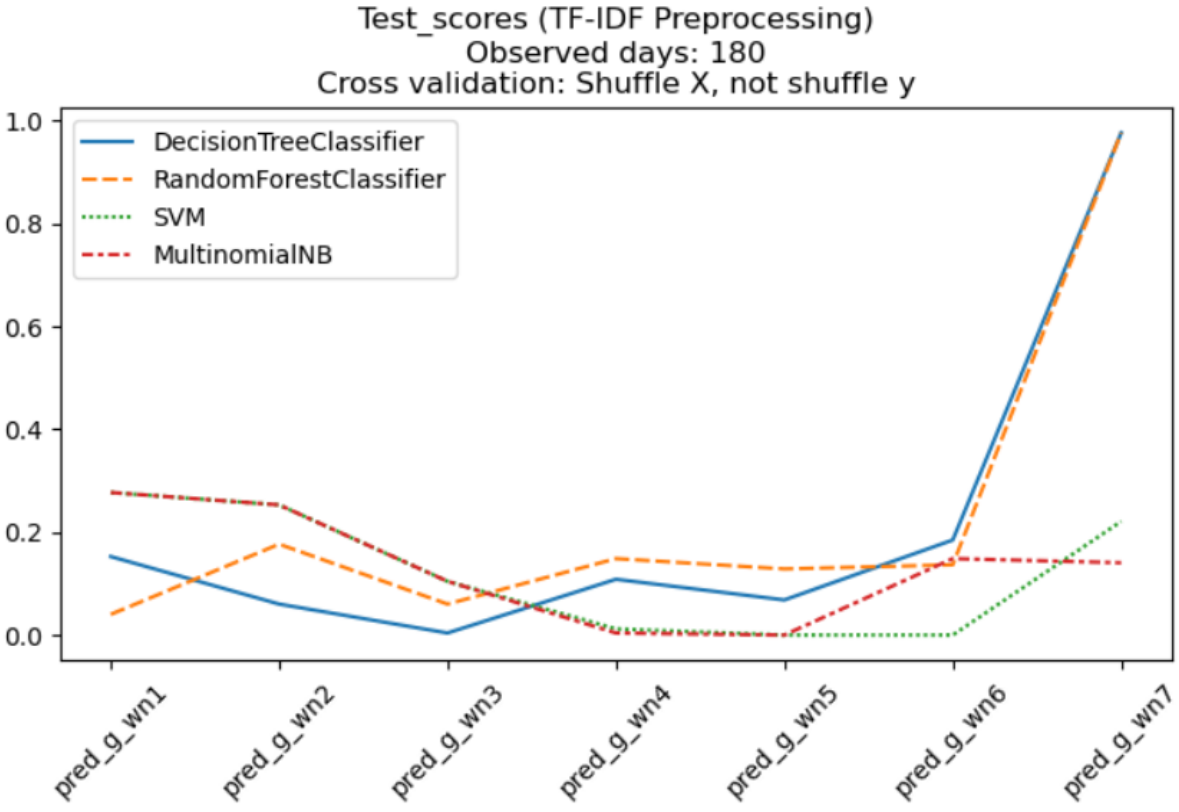

Train_scores (TF-IDF Preprocessing)
Observed days: 120
Cross validation: Shuffle X, not shuffle y
DecisionTreeClassifier
RandomForestClassifier
SVM
MultinomialNB
0.0
0.2
0.4
0.6
0.8
pred_g_wn1
pred_g_wn2
pred_g_wn3
pred_g_wn4
pred_g_wn5
pred_g_wn6
pred_g_wn7

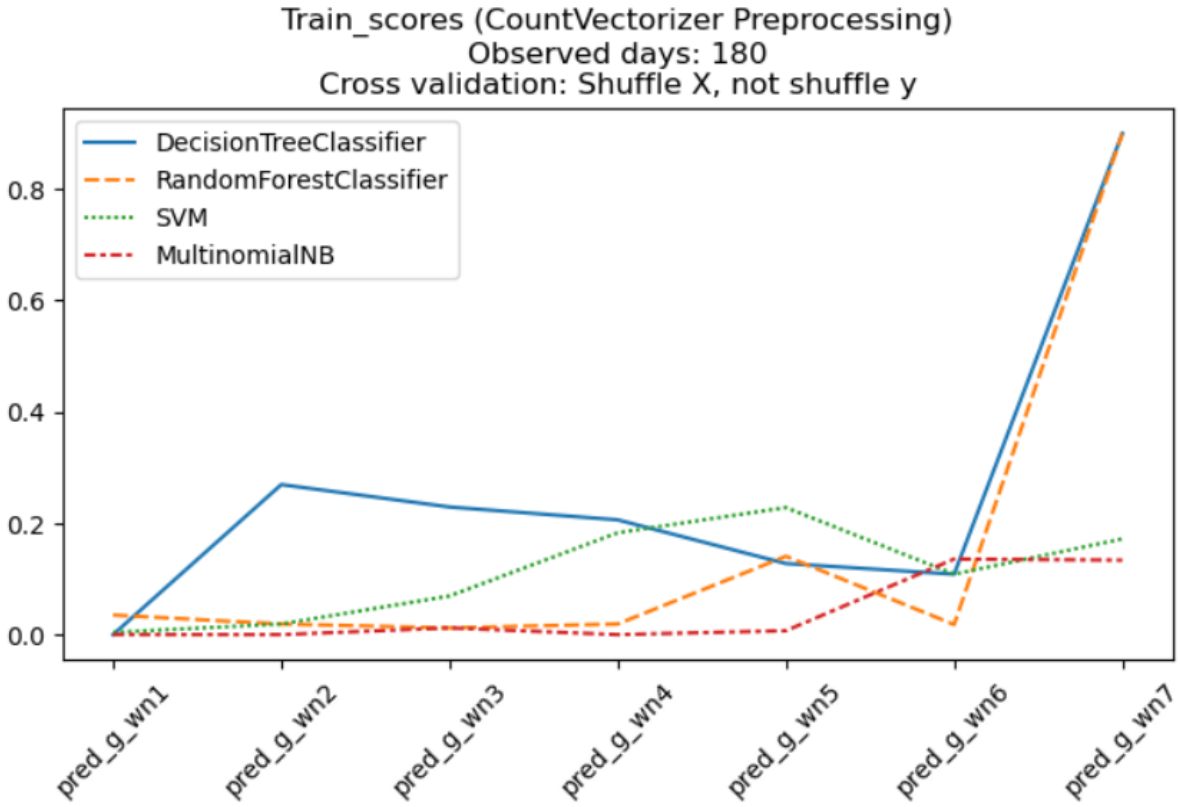

Test_scores (TF-IDF Preprocessing)
Observed days: 120
Cross validation: Shuffle X, not shuffle y
DecisionTreeClassifier
RandomForestClassifier
SVM
MultinomialNB
0.0
0.2
0.4
0.6
0.8
1.0
pred_g_wn1
pred_g_wn2
pred_g_wn3
pred_g_wn4
pred_g_wn5
pred_g_wn6
pred_g_wn7

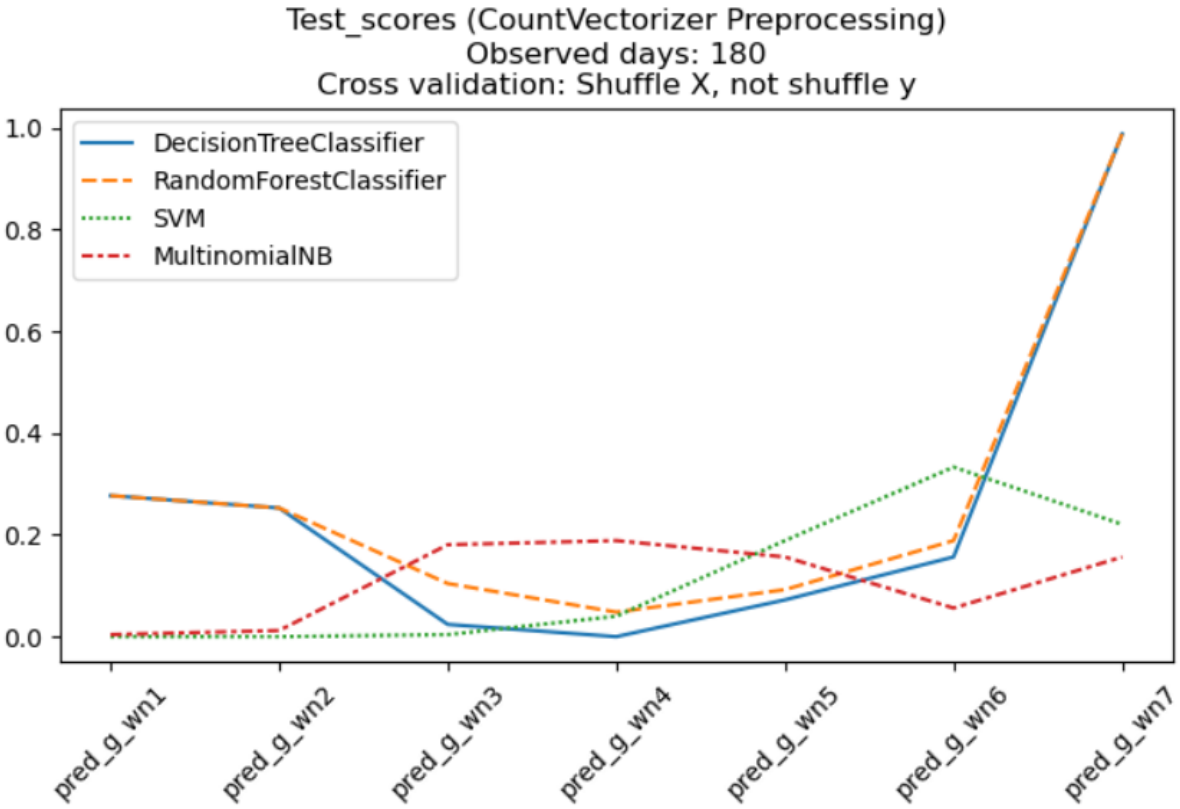

Train_scores (CountVectorizer Preprocessing)
Observed days: 120
Cross validation: Shuffle X, not shuffle y
DecisionTreeClassifier
RandomForestClassifier
SVM
MultinomialNB
0.0
0.2
0.4
0.6
0.8
pred_g_wn1
pred_g_wn2
pred_g_wn3
pred_g_wn4
pred_g_wn5
pred_g_wn6
pred_g_wn7

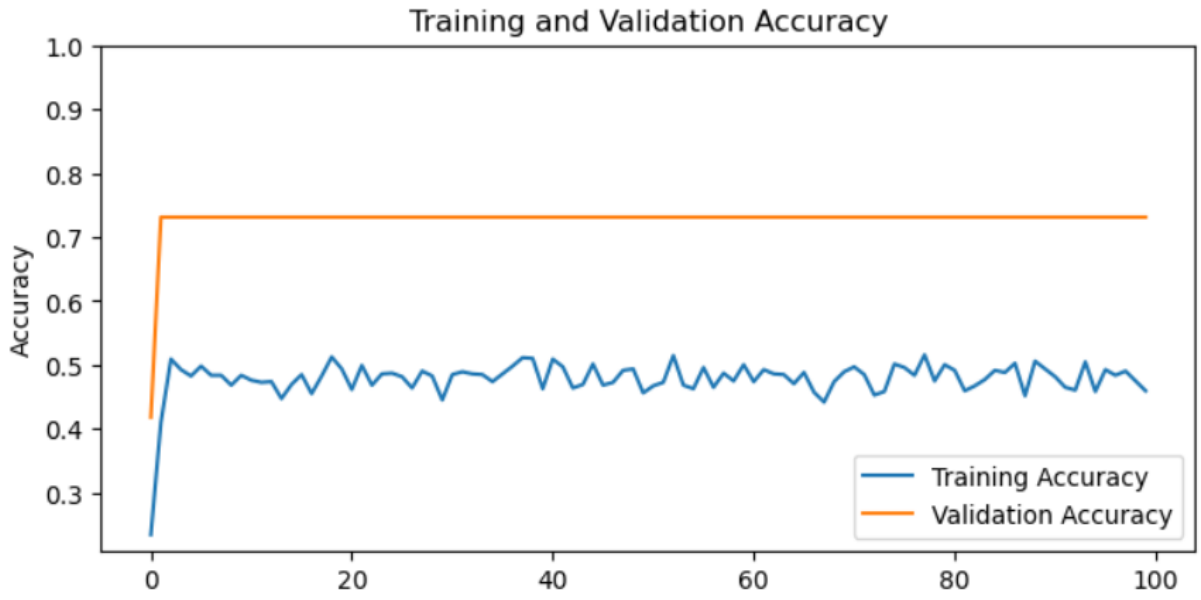

Test_scores (CountVectorizer Preprocessing)
Observed days: 120
Cross validation: Shuffle X, not shuffle y
DecisionTreeClassifier
RandomForestClassifier
SVM
MultinomialNB
0.0
0.2
0.4
0.6
0.8
1.0
pred_g_wn1
pred_g_wn2
pred_g_wn3
pred_g_wn4
pred_g_wn5
pred_g_wn6
pred_g_wn7


Train_scores (TF-IDF Preprocessing)
Observed days: 180
Cross validation: Shuffle X, not shuffle y
DecisionTreeClassifier
RandomForestClassifier
SVM
MultinomialNB
0.0
0.2
0.4
0.6
0.8
pred_g_wn1
pred_g_wn2
pred_g_wn3
pred_g_wn4
pred_g_wn5
pred_g_wn6
pred_g_wn7


Test_scores (TF-IDF Preprocessing)
Observed days: 180
Cross validation: Shuffle X, not shuffle y
DecisionTreeClassifier
RandomForestClassifier
SVM
MultinomialNB
0.0
0.2
0.4
0.6
0.8
1.0
pred_g_wn1
pred_g_wn2
pred_g_wn3
pred_g_wn4
pred_g_wn5
pred_g_wn6
pred_g_wn7


Train_scores (CountVectorizer Preprocessing)
Observed days: 180
Cross validation: Shuffle X, not shuffle y
DecisionTreeClassifier
RandomForestClassifier
SVM
MultinomialNB
0.0
0.2
0.4
0.6
0.8
pred_g_wn1
pred_g_wn2
pred_g_wn3
pred_g_wn4
pred_g_wn5
pred_g_wn6
pred_g_wn7


Test_scores (CountVectorizer Preprocessing)
Observed days: 180
Cross validation: Shuffle X, not shuffle y
DecisionTreeClassifier
RandomForestClassifier
SVM
MultinomialNB
0.0
0.2
0.4
0.6
0.8
1.0
pred_g_wn1
pred_g_wn2
pred_g_wn3
pred_g_wn4
pred_g_wn5
pred_g_wn6
pred_g_wn7



Data Answer:

The model still needs fed with more data to it to upgrade its accuracy score. The unbalance data is a big concern for model's good performance.

Business Answer

Predicting groups in next winning numbers has not yet reached much accuracy score, just about 10% - 30%.

Those models are limited in offering its diverse options for picking groups in each winning number.

Response to stakeholders

More time is needed to collect more historical winning numbers. The chance of getting the lowest prize with just 3 correct winning numbers is 19/267 cases which is about 7% potential chance of winning $10.

Further Model Development

The current dataset needs to be expanded with more observations. More factors such as days of withdrawing numbers and the values of periodic prize pool could be added in the model. Unbalance data problem should be addressed before running models.

References

Kaggle. (n.d.). Singapore Lottery Numbers. Retrieved from https://www.kaggle.com/datasets/calven22/singapore-lottery-numbers

Israeli Lotto Prediction using LSTM. (2021). Kaggle. Retrieved from https://www.kaggle.com/code/emfhal/israeli-lotto-prediction-using-lstm/

Lotto Prediction. (2020). Kaggle. Retrieved from https://www.kaggle.com/code/gogo827jz/lotto-prediction/notebook#Predict-the-Future-Draw-on-2020/Aug/26