

# Contents

<b>1</b>	<b>Introduction</b>	<b>2</b>
<b>2</b>	<b>Project Execution</b>	<b>3</b>
2.1	Work Done By Dong Xuyong . . . . .	4
2.2	Work Done By Pedro Silva . . . . .	5
2.3	Work Done By Tiago Martins . . . . .	6
2.4	The Work Methodology and Auto-Evaluation . . . . .	7
<b>3</b>	<b>Dataset Description</b>	<b>8</b>
<b>4</b>	<b>Prediction Objective</b>	<b>11</b>
4.1	Univariate Analysis . . . . .	12
4.1.1	Train-Test Split . . . . .	12
4.1.2	Growing and Rowling Window . . . . .	16
4.2	Multivariate Analysis . . . . .	19
<b>5</b>	<b>Otimization Objective</b>	<b>20</b>
<b>6</b>	<b>Development System Demonstration</b>	<b>21</b>
<b>7</b>	<b>Conclusions</b>	<b>22</b>

# 1 Introduction

Within the subject of Artificial Intelligence Techniques in Forecasting and Optimization in Business Systems, we were proposed a project aimed at using forecasting and optimization techniques in a real mind problem, in this case the distribution of drinks in a company.

Throughout this document we will first analyze the beverage sales data, then we will apply forecasting techniques to predict the number of sales of each beverage, this forecast will be divided by a univariate forecast and a multivariate forecast, for both will be used the knowledge and practices acquired in class to predict successfully. Finally, we have Optimization, which consists of finding the best values in order to maximize the sales value of each drink.

## **2 Project Execution**

This section of the document is to describe the work done by each of the elements in the group. Each description will include each of the tasks completed by each group, the effort done for this work and the time spent doing this project.

## 2.1 Work Done By Dong Xuyong

For this project the tasks made by this member where:

- Extract business objectives into features;
- Model prediction with LSTM;
- Reading the article “Understanding LSTM”;
- Run all rminer ML models;
- Parameter tuning for bud and stella;
- Pipeline for all model types;
- Analyze and run all models with all metrics for univariate variables, with different timelag combinations;
- Growing and Rowling window;
- Multivariate with VAR model and Arima with exogenous variables (precipitation and temperature);
- Weakly Naive probability implementation and experience;
- Fix the Weekly Naive template;
- Implement the GW with neural networks with multivariate series and 2 outputs and model tuning in Python;
- Analyze the optimization method.

This member of the group spent around 68 hours in this project;

## 2.2 Work Done By Pedro Silva

For this project the tasks made by this member where:

- Extract business objectives into features;
- Analysis of the exponential smoothing algorithm;
- Research in GW and RW methods;
- Search R tools;
- Formulate the validation function;
- See validation algorithms;
- Think of a strategy on how to implement this same;
- Implementation of the optimization method;
- Interface implementation.

This member of the group spent around 65 hours in this project;

## **2.3 Work Done By Tiago Martins**

For this project the tasks made by this member where:

- Extract business objectives into features;
- Analysis of Forecasting with Holt-Winters;
- Research in GW and RW methods;
- Final document development;
- See validation algorithms;
- Implementation of the optimization method;
- Interface implementation.

This member of the group spent around 60 hours in this project;

## 2.4 The Work Methodology and Auto-Evaluation

For this work, we used a work methodology where, during the various meetings we held, we usually divided them into 3 parts. The first part consisted of analyzing the results of each of the team members, in case there were tasks to be done for that meeting. The second part of the meeting consists of analyzing what needs to be done in general for the next delivery, in that same time we visualized and aligned all the tasks to be done for the next presentation. Finally, the third and final part of the meeting consisted of dividing these tasks previously defined by each of the elements of the group.

We usually have at least 2 meetings in the space from one class to another, but there were situations where more than one meeting was necessary in that space of time, mainly due to doubts in the execution of the tasks, in these situations we normally gathered everyone by video call and resolved let's get to the problem.

For the Auto-Evaluation our group thinks this the grade that each element of the group deserves:

DONG	PEDRO	TIAGO
16	16	16

Figure 1: Group Auto-Evaluation

### 3 Dataset Description

For this project, we were provided with an excel file called "bebidas.xlsx", within which are the daily sales records of each of the two beverages made available by the company in question, within that excel file there are still other relevant data, which will be detailed afterwards.

In the following image we have a print of the columns of the dataset mentioned above:

	A	B	C	D	E	F	G
1	DATA	DIA_SEMANA	PRECIPITACAO	TEMP_MAX	STELLA	BUD	
2	01/02/2019	4	6,8	30,1	53	71	
3	01/03/2019	5	0	32,9	106	235	
4	01/04/2019	6	14,2	31,8	218	42	
5	01/05/2019	7	3	27,7	180	110	
6	01/06/2019	1	0,6	29	69	15	
7	01/07/2019	2	0	31,6	18	8	
8	01/08/2019	3	0	33,2	61	10	
9	01/09/2019	4	0	31,1	38	6	
10	01/10/2019	5	0	33,2	545	26	

Figure 2: Project Dataset

The following dataset is composed of 6 columns, they being:

- DATA: This column represents the date the records are from;
- DIA\_SEMANA: This column represents the day of the week, where 1 is Sunday, 2 is Monday, 3 is Tuesday, 4 is Wednesday, 5 is Thursday, 6 is Friday and 7 is Saturday;
- PRECIPITACAO: This column represents the total of precipitation in mm in that day;
- TEMP\_MAX: This column represents the daily maximum temperature in Celcius from that day;
- STELLA: This column represents the number of STELLA drinks that were sold in that day;
- BUD: This column represents the number of BUD drinks that were sold in that day.

To get a nice description of the values of each column we calculate the minimum value, the median, the mean, the max and other important values for each column, the next image represents the values that we got:

DATA	DIA_SEMANA	PRECIPITACAO	TEMP_MAX	STELLA	BUD
Min. :2019-01-02	Min. :1.000	Min. : 0.000	Min. :21.40	Min. : 0.0	Min. : 0.0
1st Qu.:2019-07-03	1st Qu.:2.000	1st Qu.: 0.000	1st Qu.:29.20	1st Qu.: 13.0	1st Qu.: 22.0
Median :2020-01-01	Median :4.000	Median : 0.000	Median :30.90	Median : 47.0	Median : 58.0
Mean :2020-01-01	Mean :4.001	Mean : 3.669	Mean :31.13	Mean : 105.4	Mean : 101.4
3rd Qu.:2020-07-01	3rd Qu.:6.000	3rd Qu.: 1.550	3rd Qu.:32.80	3rd Qu.: 128.8	3rd Qu.: 125.8
Max. :2020-12-31	Max. :7.000	Max. :66.000	Max. :40.50	Max. :1335.0	Max. :1280.0

Figure 3: Dataset Statistics



These two diagrams represent the sales of each of the drinks provided in the dataset:

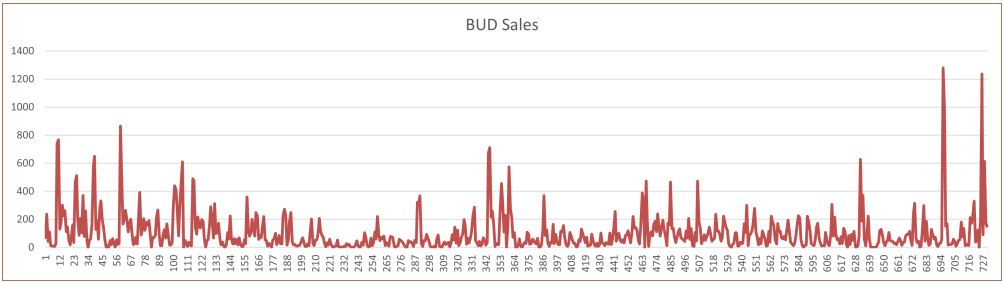


Figure 4: BUD Sales

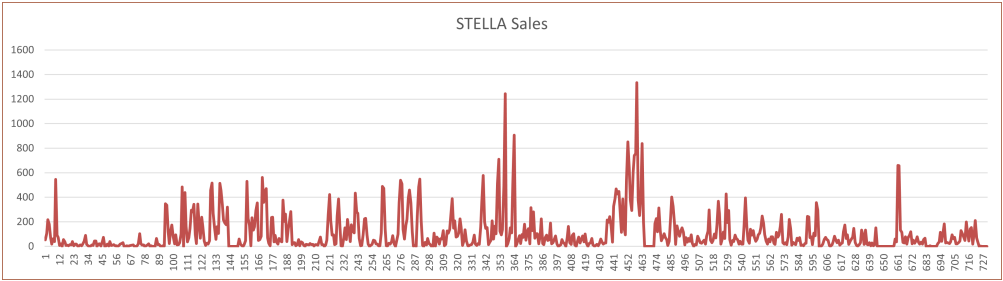


Figure 5: STELLA Sales

The next diagrams represent the outliers of sales data in relation to the average sales of the respective dring (BUD and STELLA):

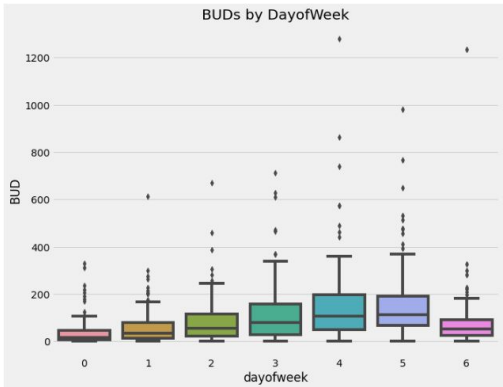


Figure 6: BUD Outliers

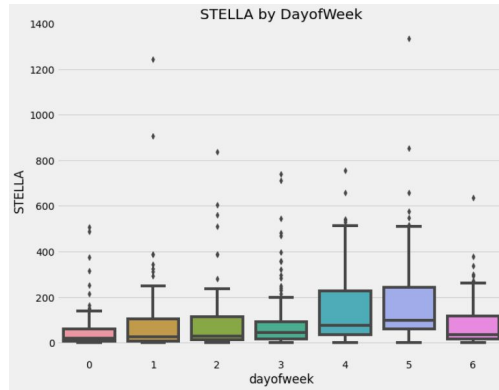


Figure 7: STELLA Outliers

## 4 Prediction Objective

One of the tasks for this project is the prediction of the number of sales of each one of the drinks (STELLA and BUD). To do that task there are two main modules that we can use, Univariate Analysis and Multivariate Analysis.

Univariate Analysis consist in examining the relationship between a single column of data, that means that for this project we will create 2 univariate predictions, one for STELLA and the other for BUD. Also each one of those predictions will include multiple prediction methods, and the objective is to find the prediction method with the lowest error rate, for each of the drinks.

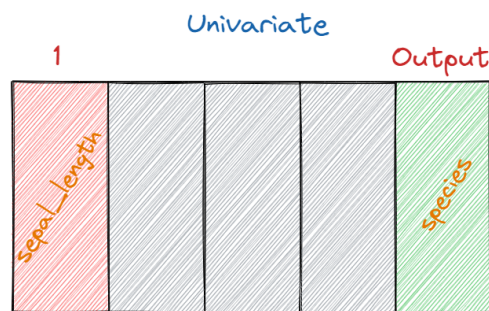


Figure 8: Univariate Analysis

Multivariate Analysis is similar to the univariate analysis the main difference between the two is that multivariate analysis does not focus in only one column of data, but instead use the data from multiple columns of data to predict. The data that will be included for the prediction of the sales of each drink will include of course the sales of each respective drink, the precipitation and the maximum temperature on that day.

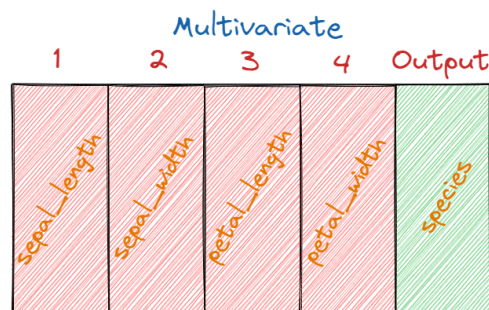


Figure 9: Multivariate Analysis

## 4.1 Univariate Analysis

For the Univariate Analysis there are multiple prediction methods that we will use, we will use mainly the machine learning and forecast methods.

Our objective was to predict the last 20 weeks of each one of the drinks, for that we will use two methods for training the machine, they being the train-test split and the Growing and Rowling window split.

### 4.1.1 Train-Test Split

In the Train-test split as we mention above, we will use Machine Learning prediction methods and Forecast prediction methods to predict the last 20 weeks of the sales. For each methods we create a script that run all of them and later will show the best methods. To determine the best method we will calculate the error rate from each method and of course the method with the lowest error rate will be considered the best prediction method.

For the Machine Learning prediction methods we will use the following:

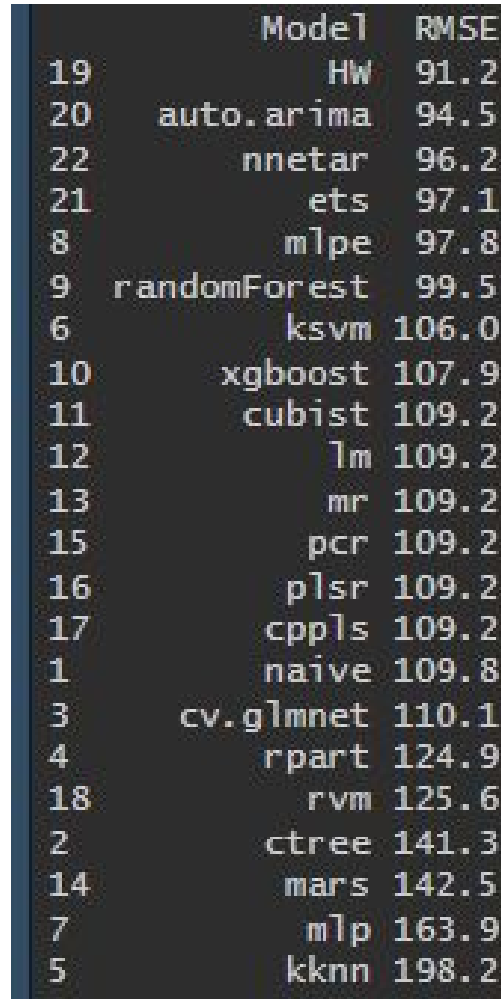
- "naive";
- "ctree";
- "cv.glmnet";
- "kkn";
- "mlp";
- "randomForest";
- "xgboost";
- "cubist";
- "lm";
- "mars";
- "pcr";
- "plsr";
- "cppls";
- "rvm".

For the Forecast predictions methods we will sue the following:

- "Holt-Winters";
- "auto.arima";
- "ets";

•"nnetar".

After running the script that was developed by our group, we determine that the best prediction method was Holt Winters for the BUD drink, and mars for the STELLA drink, as they were the methods with the lowest error rate from all the Machine Learning predictions methods and all the Forecast predictions methods. The next image will allow us to see all the error rates from all the predictions methods.



	Model	RMSE
19	HW	91.2
20	auto.arima	94.5
22	nnetar	96.2
21	ets	97.1
8	mlpe	97.8
9	randomForest	99.5
6	ksvm	106.0
10	xgboost	107.9
11	cubist	109.2
12	lm	109.2
13	mr	109.2
15	pcr	109.2
16	pls	109.2
17	cppls	109.2
1	naive	109.8
3	cv.glmnet	110.1
4	rpart	124.9
18	rvm	125.6
2	ctree	141.3
14	mars	142.5
7	mlp	163.9
5	kknn	198.2

Figure 10: BUD - Machine Learning and Forecast Results

As we can see from this image, the best prediction method for the BUD drinks was Holt Winters, also the second best and the third best methods were auto.arima and nnetar. Just a side note we can see that the best 3 methods come from Forecast predictions.

	Model	RMSE
14	mars	185.8
1	naive	186.5
12	lm	186.5
13	mr	186.5
15	pcr	186.5
16	pls	186.5
17	cppls	186.5
3	cv.glmnet	186.6
9	randomForest	186.7
2	ctree	187.3
4	rpart	187.4
7	mlp	187.9
8	mlpe	189.2
18	rvm	190.7
21	ets	190.8
20	auto.arima	194.4
11	cubist	198.2
10	xgboost	201.0
5	kknn	201.3
6	ksvm	204.3
22	nnetar	206.1
19	HW	206.7

Figure 11: STELLA - Machine Learning and Forecast Results

As we can see from this image, the best prediction method for the STELLA drinks was mars, also the second best and the third best were naive and lm. Just a side note we can see that the best3 methods where form Machine Learning predictions.

For demonstration purposes here are the graphics of each of the predictions made by the best method for BUD and STELLA:

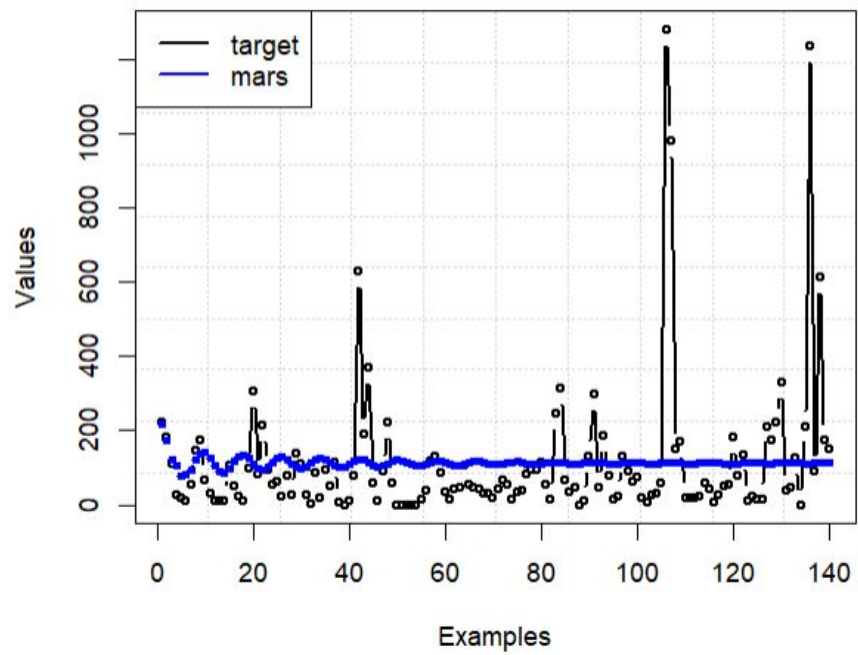


Figure 12: BUD - Machine Learning and Forecast HW Graph

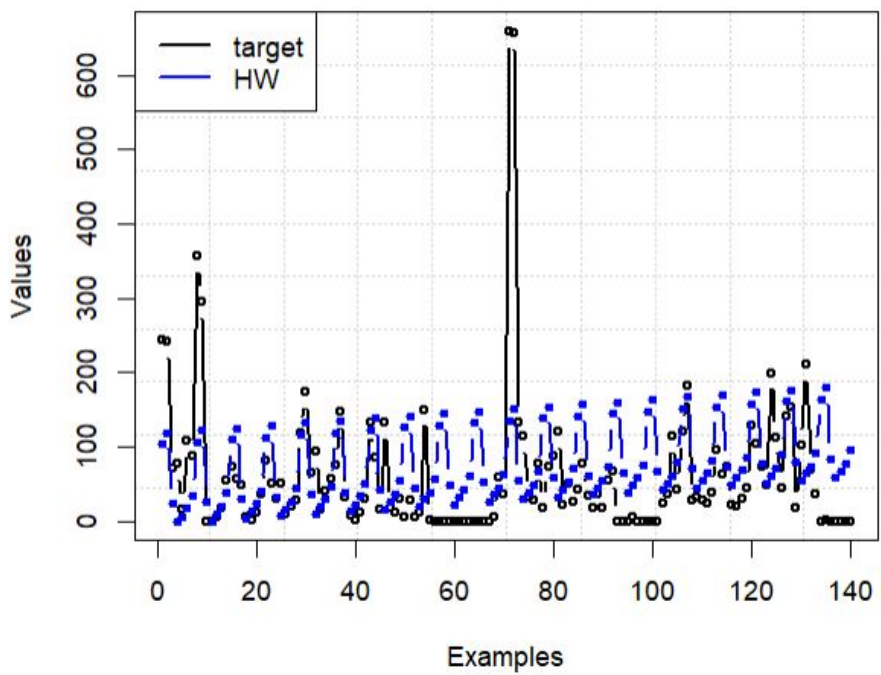


Figure 13: STELLA - Machine Learning and Forecast mars Graph

#### 4.1.2 Growing and Rowling Window

Similar to above for this split we will use the same predictions methods, but we will also add the Weekly naive method.

```
> bud_metrics
  X      Model      MSE
1 16      lm 21093.14
2 17      mr 21093.14
3 19      pcr 21093.14
4 20      pls 21093.14
5 21      cppls 21093.14
6 10      ksvm 21807.55
7 13 randomForest 22132.54
8 15      cubist 22925.25
9 18      mars 23605.71
10 5      naive 23827.39
11 7      cv.glmnet 24078.55
12 14      xgboost 24822.12
13 6      ctree 25071.61
14 12      mlpe 25989.65
15 9      kkn 25993.33
16 11      mlp 26054.86
17 8      rpart 27084.84
18 22      rvm 28047.55
19 3      ets 36382.69
20 2      auto.arima 37431.72
21 4      nnetar 40482.54
22 1      HW 41338.73
23 23 weekly_naive 46756.09
```

Figure 14: BUD - Growing and Rowling Window Results

As we can see from this image, the best prediction method for the BUD drinks was lm, also the second best and the thir best methods were mr and pcr.



```

> stella_metrics
  X      Model      MSE
1 16      lm 5194.414
2 17      mr 5194.414
3 19      pcr 5194.414
4 20      plsr 5194.414
5 21      cppls 5194.414
6 18      mars 5880.614
7 12      mlpe 6323.571
8 15      cubist 6458.479
9 10      ksvm 6466.993
10 11      mlp 7026.807
11 14      xgboost 8163.121
12 13 randomForest 8236.414
13 2      auto.arima 8652.407
14 8      rpart 9006.429
15 6      ctree 9236.121
16 1      HW 9522.021
17 22      rvm 9697.829
18 9      kknk 10160.043
19 3      ets 10823.307
20 7      cv.glmnet 10852.136
21 5      naive 11267.979
22 23 weekly_naive 16219.343
23 4      nnetar 20830.221

```

Figure 15: STELLA - Growing and Rowling Window Results

As we can see from this image, the best prediction method for the STELLA drinks was lm, also the second best and the third best were mr and pcr.

For demonstration purposes here are the graphics of each of the predictions made by the best method for BUD and STELLA:

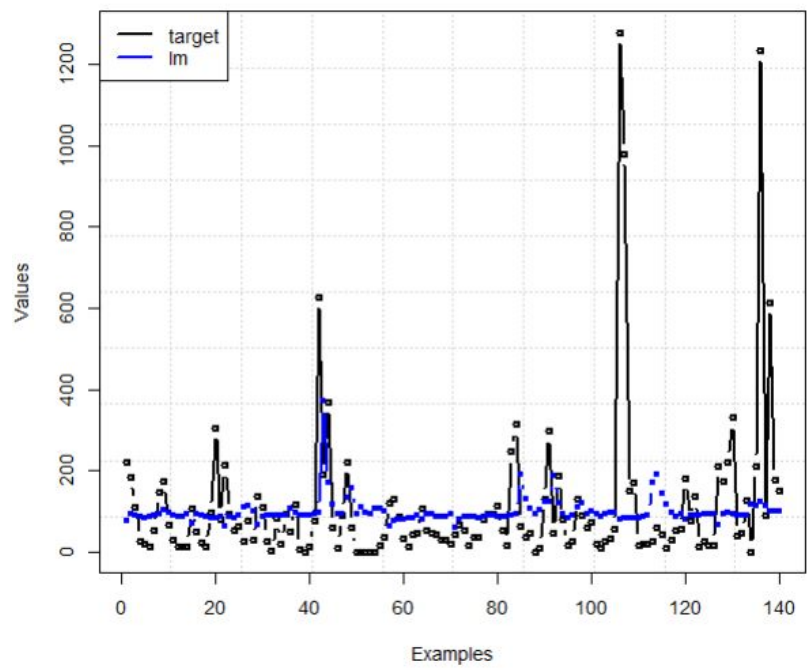


Figure 16: BUD - Growing and Rowling Window Im Graph

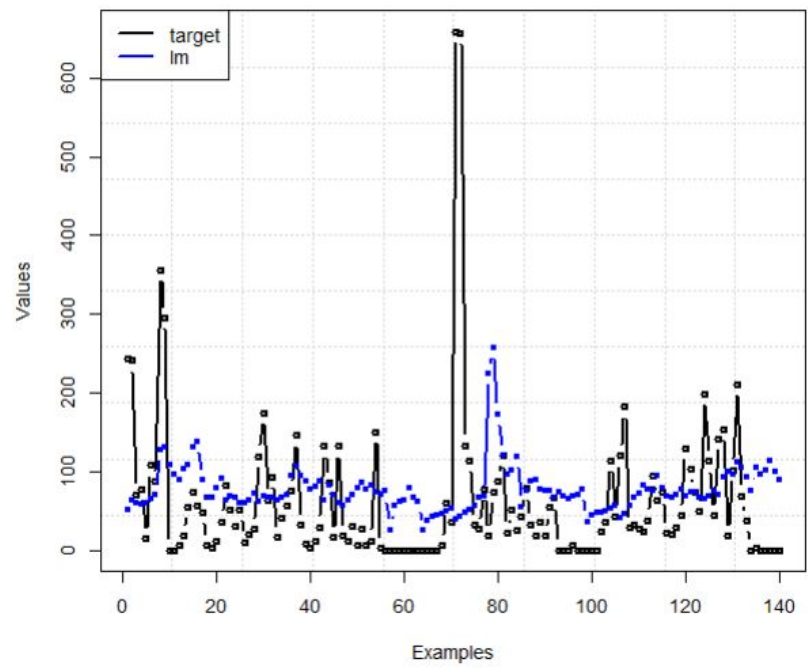


Figure 17: STELLA - Growing and Rowling Window Im Graph

## 4.2 Multivariate Analysis

## 5 Optimization Objective

## **6 Development System Demonstration**

Where is the link for the video of the demonstration: "link do youtube".

## 7 Conclusions