

**Principal Global Investor Dynamic Risk Premia Performance Tracking Project
-Binary Classification Validation and Visualization**

by

Yurui Li

A Creative Component paper submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Major: Industrial Engineering

Program of Study Committee:
Dr. Dave Sly, Major Professor
Dr. Gary Mirka

Iowa State University

Ames, Iowa

2019

Copyright © Yurui Li, 2019. All rights reserved.

DEDICATION

I dedicate this paper to my parents Mr. Songping Li and Mrs. Xiangyun Shu for their unconditional love and support.

TABLE OF CONTENTS

	Page
LIST OF FIGURES	iv
LIST OF TABLES	v
ACKNOWLEDGMENTS	vi
ABSTRACT	vii
CHAPTER 1. INTRODUCTION	1
Objectives	3
Deliverables	4
Assumption and Constraints	4
Assumption	4
Constraints	4
CHAPTER 2. METHODOLOGY	5
Validation Metric	6
Standard Validation Metric	8
Confusion Matrix	8
Matthews Correlation Coefficient (MCC)	10
Receiver Operating Characteristic (ROC)	11
Visualization	12
Data and Dataset	12
Reports	13
Dashboard	14
CHAPTER 3. RESULTS	15
ACC and MCC	15
ROC	17
Confusion Matrix	18
Classes Distribution Plot	19
Magnitude Comparison	21
Volatility of Prediction Outcome	23
Model Comparison	24
Actual Return Aquarium	25
CHAPTER 4. CONCLUSION	27
REFERENCES	29
APPENDIEX	30

LIST OF FIGURES

	Page
Figure 1: <i>DRP Process Map</i>	3
Figure 2: <i>Outcome of Factor Prediction Model</i>	5
Figure 3 <i>Confusion Matrix</i>	8
Figure 4: <i>A Basic ROC Plot</i>	11
Figure 5: <i>Example of Final Dataset</i>	13
Figure 6: <i>DRP Dashboard</i>	14
Figure 7: <i>Overall ACC and MCC Plot</i>	16
Figure 8: <i>ROC Plot from 1996 to 2013</i>	17
Figure 9: <i>ROC Plot from 2014 to 2018</i>	18
Figure 10: <i>Confusion Matrix and FPR Plot</i>	19
Figure 11: <i>Confusion Matrix and FPR Plot from 2014 to 2018</i>	19
Figure 12: <i>Distribution Plot of Positive and Negative Classes</i>	20
Figure 13: <i>Distribution Plot of Positive and Negative Classes from 2013 to 2018</i>	21
Figure 14: <i>Magnitude Comparison</i>	22
Figure 15: <i>Magnitude Comparison from 2014 to 2018</i>	22
Figure 16: <i>Volatility of Prediction Outcome</i>	23
Figure 17: <i>Model Comparison between Different Forecast Horizon</i>	24
Figure 18: <i>Model Comparison between Different Forecast Horizon from 2013 to 2018</i>	25
Figure 19: <i>Actual Return Aquarium</i>	26
Figure 20: <i>M Code</i>	30

LIST OF TABLES

	Page
Table 1: <i>Summary of 19 Performance Measures Examined in this Study</i>	6
Table 2: <i>Summary of Confusion Matrix Related Performance Measures</i>	9
Table 3: <i>Summary of Matthews Correlation Coefficient</i>	10

ACKNOWLEDGMENTS

I would like to take this opportunity to express my gratitude to my major professor Dr. David P. Sly. He has supported me all through my graduate study. I also would like to thank my committee member Dr. Gary A. Mirka for his support and encouragement.

In addition, I would like to thank my friends, colleagues, the department faculty and staff for making my time at Iowa State University a wonderful experience.

ABSTRACT

Principal Global Investor (PGI) is a division of the Principal Financial Group headquartered in Des Moines, Iowa. PGI works to provide their customers with investment knowledge and strategic solutions to create successful outcomes. One method they use is the Dynamic Risk Premia (DRP) process. The DRP is a Random Forest-based investment model that uses historical data to predict the performance of market characteristics, known as factors. Eventually, the factors will be weighted and used to forecast stock performance.

One issue with the DRP process was that the performance of the underlying model (Random Forest) was not evaluated using standard statistical measures. The lack of standard performance measure metrics of the underlying model brought uncertainty into the DRP process. In addition, without visibility of the underlying model performance, it was hard for Principal to compare alternative solutions/algorithms to the current system.

This study defined a standard model validation metric and created a dashboard to visualize the performance for 133 factors across 13 sectors in 12 different prediction horizons from 7/26/1996 to 1/26/2018. By quantifying and visualizing the model performance, this study demonstrated that the DRP prediction model performed consistently well from 1996 to 2012. However, after 2013, when the model was first launched in production, its performance was close to a random guess. After reviewing the results of this study, Principal started to research alternative algorithms and rebuild the DRP model.

CHAPTER 1. INTRODUCTION

Principal Global Investor (PGI) is using a Dynamic Risk Premia (DRP) process to predict stock performance. For this study, the DRP process has been divided into two key sub-processes: prediction and weighing (shown in Figure 1). The prediction uses a Random Forest algorithm to predict the outperformance or underperformance of each factor. The weighing uses the Principal Component Analysis to assign weights to all factors. This project mainly focuses on the prediction process. The prediction process begins by grouping a focal company or stock by its Morgan Stanley Capital International (MSCI Inc.) group. Once the group has been established, the historical and weekly factor data of all stocks in the group are pulled from FactSet. Additional macro factor data are sourced from Bloomberg. Then, all factor data are aggregated and prepared for the factor prediction model (Random Forest) by factor data aggregation. Once the combined and smoothed factor dataset containing new and historical data is created, the factor prediction model uses those data to predict the probability of overperformance/underperformance of the factors associated with the selected group. The prediction results are run through the weighing process. Finally, the prediction results are published for portfolio creation.

One of the major issues of the process was that the performance of the underlying model (Random Forest) has not been analyzed using any standard statistical measures such as classification accuracy, precision, recall, or Area Under Curve (AUC). This caused a lack of confidence and visibility of the underlying model performance. As Figure 1 shows, the only evaluation of the DRP process (backtesting) was based on the simulated portfolio, which means it will not evaluate the process until a simulated portfolio is created. In addition, this evaluation was designed to check the portfolio performance instead of the

prediction model performance. Lacking “quality control” of the prediction model’s outcome led to deficiency in quantifying confidence level for the prediction outcomes or the entire process. These uncontrolled outcomes then became the input of the weighing process (Principal Component Analysis), which brought uncertainty to the final results. Moreover, when the backtesting indicates the simulated portfolio is underperforming, it is very difficult to identify the root of the cause.

Another issue was the lack of visibility in the prediction model’s performance. With the rapid development of data science, there might be some new algorithms that can give Principal a more accurate factor prediction. However, without the standard model validation metrics and a visual representation of the performance over time, it was impossible for Principal to compare alternative solutions or algorithms to the current system. Furthermore, the visual representation of the performance over time can also be used to separate short term noise from long term signals.

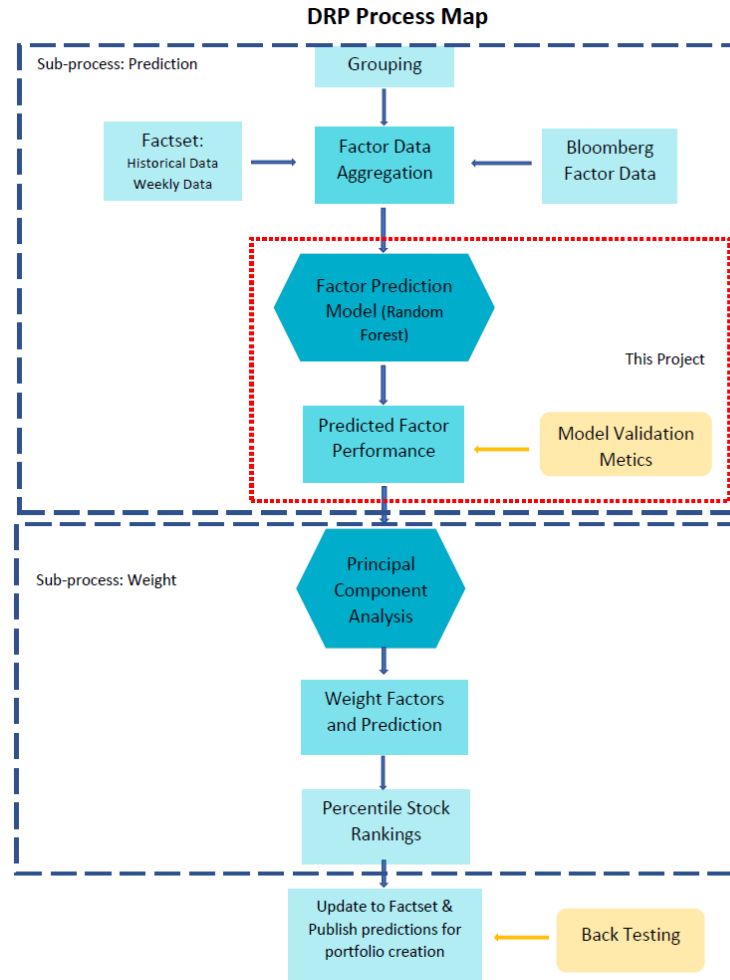


Figure 1: *DRP Process Map*

Objectives

- Define a standard model validation metric for binary classification prediction model (majorly Random Forest)
- Create a visual representation of the performance over time

Deliverables

- A graphical user interface dashboard visualizing the underlying model performance as well as each factor performance over time (Power BI)
- A comprehensive report describing the validation metric, process controls, and visualizations

Assumption and Constraints

Assumption

- Historical data are accurate and consistent
- Bloomberg factor data updates weekly
- The Factor Data Aggregation accurately aggregates FactSet and Bloomberg data

Constraints

- All analysis is based on the historical data provided by Principal
- Modification of the Random Forest Algorithm is not within the scope of the project
- The time frame for the model performance data is from 7/26/1996 to 1/26/2018 weekly

CHAPTER 2. METHODOLOGY

Random Forest is a type of supervised binary classification, which was proposed by Breiman in 2001. Random forest is a combination of tree predictors, where each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest [1]. For the DRP model, Random Forest outcome is the probability for each factor to be either 1 or 0, in which 1 means outperformance and 0 means underperformance. If the probability is higher than 0.55, the factor is classified as an overperform factor, otherwise, it is classified as an underperform factor. The final outcome of each Factor Prediction Model is stored in the Excel files shown in Figure 2. In the DRP model, there are 13 sectors containing 10 factor groups and 133 factors. The model weekly predicts the performance of each factor in each prediction horizon from Forward One Month (F1M) to Forward Twelve Month (F12M).

















































































































































 R1K STM RET_F1M.csv	 R	 R1K All RET_F11M.xlsx	 R	 R1K Reits RET_F1M.csv	 R
 R1K Reits RET_F3M.csv	 R	 R1K Reits RET_F4M.csv	 R	 R1K Reits RET_F5M.csv	 R
 R1K Reits RET_F7M.csv	 R	 R1K Reits RET_F8M.csv	 R	 R1K Reits RET_F9M.csv	 R
 R1K Reits RET_F11M.csv	 R	 R1K Reits RET_F10M.csv	 R	 R1K All RET_F9M.csv	 R
 R1K All RET_F7M.csv	 R	 R1K All RET_F6M.csv	 R	 R1K All RET_F1M.csv	 R
 R1K All RET_F3M.csv	 R	 R1K All RET_F4M.csv	 R	 R1K All RET_F5M.csv	 R
 R1K All RET_F10M.csv	 R	 R1K All RET_F12M.csv	 R	 R1K All RET_F11M.csv	 R
 R1K Energy RET_F2M.csv	 R	 R1K Energy RET_F3M.csv	 R	 R1K Energy RET_F4M.csv	 R
 R1K Energy RET_F6M.csv	 R	 R1K Energy RET_F7M.csv	 R	 R1K Energy RET_F9M.csv	 R
 R1K Energy RET_F12M.csv	 R	 R1K Energy RET_F11M.csv	 R	 R1K Energy RET_F10M.csv	 R
 R1K Bank RET_F2M.csv	 R	 R1K Bank RET_F3M.csv	 R	 R1K Bank RET_F4M.csv	 R
 R1K Bank RET_F7M.csv	 R	 R1K Bank RET_F8M.csv	 R	 R1K Bank RET_F9M.csv	 R
 R1K Bank RET_F12M.csv	 R	 R1K Bank RET_F11M.csv	 R	 R1K Bank RET_F10M.csv	 R
 R1K DivFin RET_F2M.csv	 R	 R1K DivFin RET_F3M.csv	 R	 R1K DivFin RET_F4M.csv	 R
 R1K DivFin RET_F6M.csv	 R	 R1K DivFin RET_F7M.csv	 R	 R1K DivFin RET_F9M.csv	 R
 R1K DivFin RET_F11M.csv	 R	 R1K DivFin RET_F12M.csv	 R	 R1K DivFin RET_F10M.csv	 R
 R1K Software RET_F2M.csv	 R	 R1K Software RET_F3M.csv	 R	 R1K Software RET_F4M.csv	 R
 R1K Software RET_F6M.csv	 R	 R1K Software RET_F9M.csv	 R	 R1K Software RET_F8M.csv	 R
 R1K CS RET_F1M.csv	 R	 R1K CS RET_F2M.csv	 R	 R1K CS RET_F3M.csv	 R
 R1K CS RET_F5M.csv	 R	 R1K CS RET_F6M.csv	 R	 R1K CS RET_F7M.csv	 R
 R1K CS RET_F9M.csv	 R	 R1K Software RET_F12M.csv	 R	 R1K Software RET_F11M.csv	 R
 R1K CS RET_F12M.csv	 R	 R1K CS RET_F10M.csv	 R	 R1K CS RET_F11M.csv	 R
 R1K Health RET_F1M.csv	 R	 R1K CD RET_F3M.csv	 R	 R1K Health RET_F2M.csv	 R
 R1K CD RET_F5M.csv	 R	 R1K CD RET_F6M.csv	 R	 R1K Health RET_F3M.csv	 R

Figure 2: Outcome of Factor Prediction Model

Validation Metric

There are abundant statistical performance measure methods can be used to validate binary classification; finding the most appropriate of these is the critical part of this study. In this study, nineteen different binary classification performance measure methods (show in Table 1) were researched. After considering the features of the DRP model, eight of those performance measure methods were selected to construct one standard validation metric.

For the DRP model, the ratio between positive and negative class is 13:12, which is considered as balanced. In addition, for stock selection model, both underperformance (negative) class and overperformance (positive) class are equally important. So, the performance measure should consider all classes. The majority of unselected statistical performance measures are designed for the unbalanced classes, which either positive or negative class dominates the dataset. Some of the unselected performance measures, such as Prevalence, only focus on one class that is either positive or negative.

Table 1: *Summary of 19 Performance Measures Examined in this Study*

Performance Measure	Explanation	Standard Model Validation Metric
Accuracy	Measuring the proportion of correct prediction	Selected
Class Balance Accuracy	Comparing outcomes learned from imbalanced data by weighting each class	Not selected—it is designed for imbalanced data and only focuses on one class
Confusion Matrix	Visualizing all four classes by a table format	Selected
Cost-sensitive Learning	Assigning cost-rate to different types of misclassification error	Not selected—cost-rate is arbitrarily defined
Cumulative Gain & Lift Chart	Visualizing the effectiveness of the prediction model by computing the ratio between the results obtained with and without the model	Not selected—dataset cannot support computation

Distribution of Classes	Visualizing the model's ability of distinguishing between positive and negative classes	Selected
F-measure	Computing the harmonic mean of precision and recall	Not selected— It is biased to the majority class, since it does not fully consider all four classes in the confusion matrix
F-beta measure	Generalizing the F measure as a weighted harmonic mean	Not selected— It is biased to the majority class, since it does not fully consider all four classes in the confusion matrix
False Positive Rate	Measuring the proportion of real negative cases, which were predicted positive	Selected
G-measure	Computing geometric mean of precision and recall	Not selected— It is biased to the majority class, since it does not fully consider all four classes in the confusion matrix
Kappa	Comparing model accuracy with a randomly generated accuracy	Not selected—it contains the same information as MCC
Kolmogorov-Smirnov chart	Measuring the degree of separation between the positive and negative distributions	Not selected—dataset cannot support computation
Magnitude Comparison	Comparing the magnitude (average return) of predicted result and the actual	Selected
Positive Predictive Value (also called precision)	Measuring how well the model is predicting positive class	Not selected—it is biased to the majority class, since it does not fully consider all four classes in the confusion matrix
Prevalence	Measuring the proportion of positive case among the dataset	Not selected—it only focuses on the positive cases
Receiver Operating Characteristic (ROC)	Visualizing the model performance by plotting the true positive rate as a function of false positive rate	Selected
Specificity (also called True Negative Rate)	Measuring the proportion of real negative cases that are correctly predicted negative	Not selected—it is biased to the majority class, since it does not fully consider all four classes in the confusion matrix

True Positive Rate (also called Recall or Sensitivity)	Measuring the proportion of real positives that are correctly identified as such	Selected
Volatility of Prediction Outcome	Visualizing the volatility of prediction outcome over time	Selected

Standard Validation Metric

The finalized standard validation metric contains eight performance measure methods: Accuracy (ACC), Confusion Matrix, Distribution of Classes, False Positive Rate (FPR), Magnitude Comparison, Receiver Operating Characteristic (ROC), True Positive Rate (TPR), and Volatility.

Confusion Matrix

The Confusion Matrix is a specific table layout that visualizes the performance of a supervised learning algorithm [2]. As shown in Figure 3, the instances in predicted classes are represented by each column, while the instances in actual (known) classes are represented by each row [3]. The green cells represent correct predictions, i.e., true positives and true negatives, and the red cells represent incorrect predictions, i.e., false negatives and false positives.

		Prediction Class	
		P	N
Actual Class	P	True Positives (TP)	False Negatives (FN)
	N	False Positives (FP)	True Negatives (TN)

Figure 3 *Confusion Matrix*

The Confusion Matrix provides a direct visualization of the classifier behavior in each class. A variety of performance measures can be calculated based on the four cells of the confusion matrix. In this study, three of the most commonly used performance measures were selected as parts of the standard validation metric. They are Accuracy (ACC), the True Positive Rate (TPR; also called Recall or Sensitivity), and the False Positive Rate (FPR). ACC is the proportion of all true positive and true negative results to the whole dataset. It is a general indicator of how accurate a classifier is. Next, the TPR is the proportion of real positive cases that are correctly predicted to be positive to the dataset. It indicates how often the model predicts positive when the actual is positive. Finally, the FPR is the proportion of real negative cases that are correctly predicted to be positive to the dataset. It illustrates how often the model predicts positive when the actual is negative. FPR is one of the most important performance measures for Principal, because purchasing an underperforming stock has a higher negative impact on a portfolio than not purchasing an overperforming stock. Table 2 provides a summary of these performance measure methods and the formula for each measure.

Table 2: *Summary of Confusion Matrix Related Performance Measures*

Performance Measure	Definition	Question answered	Formula
Accuracy (ACC)	The proportion of True Results (both true positives and true negatives) among the dataset	Overall, how often is the classifier correct?	$ACC = \frac{TP + TN}{TP + FP + TN + FN}$
True Positive Rate (TPR)	The proportion of Real Positive cases that are correctly Predicted Positive	When the actual case is positive, how often does the classifier predict positive?	$TPR = \frac{TP}{TP + FN}$

False Positive Rate (FPR)	The proportion of Real Negatives that occurs as Predicted Positive	When the actual case is negative, how often does the classifier predict positive?	$FPR = \frac{FP}{FP + TN}$
---------------------------	--	---	----------------------------

Matthews Correlation Coefficient (MCC)

Matthews Correlation Coefficient (MCC) is a correlation coefficient between actuals and predictions. Since the values of all four quadrants of a confusion matrix are involved, MCC is considered as a balanced measure [4]. In this study, MCC was selected rather than F1 score, because the widely used statistical measures, accuracy and F1 score can both be misleading since none of them fully considers all four classes of the confusion matrix in the final score computation [5]. As shown in Table 3, MCC is the geometric mean of informedness and markedness [6]. It generally varies between -1 and +1. Positive one means actual value equals the prediction, zero means the model is randomly predicting the actuals, negative one means there is a total negative correlation between actual value and the prediction.

Table 3: *Summary of Matthews Correlation Coefficient*

Matthews Correlation Coefficient		
$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$		
<i>Definition</i>	<i>Advantage</i>	<i>Interpretation</i>
<p>It is the geometric mean of informedness and markedness</p> <ul style="list-style-type: none"> “Informedness, Kappa_I, is the probability of an informed decision” [6] <p>$Informedness = TPR + TNP - 1$</p>	<p>It is a balanced measure that considers all the four quadrants of a confusion matrix</p>	<p>It generally varies between -1 and +1</p> <ul style="list-style-type: none"> 1 indicates there is a perfect agreement between actuals and predictions

<ul style="list-style-type: none"> • “Markedness, Kappa_k, is the probability of a decision variable being marked by the real class” [6] $Markedness = PPV + NPV - 1$ 		<ul style="list-style-type: none"> • 0 indicates the prediction is random with respect to the actuals • -1 indicates there is a total disagreement between actuals and predictions
--	--	--

Receiver Operating Characteristic (ROC)

Receiver operating characteristics (ROC) is a type of graph that organizes and visualizes the performance of classifiers. Nowadays, it is increasingly used in machine learning and data mining research. In a ROC plot, X-axis is the false positive rate (FPR), Y-axis is the true positive rate (TPR). The diagonal line ($y = x$) means the classifier is randomly guessing a class. The best possible outcome of a classifier will generate a point at the top-left corner (0,1) in ROC space [7]. The green area shown in Figure 4 represents the classifier performance better than a random guess, the red area means it is worse than a random guess.

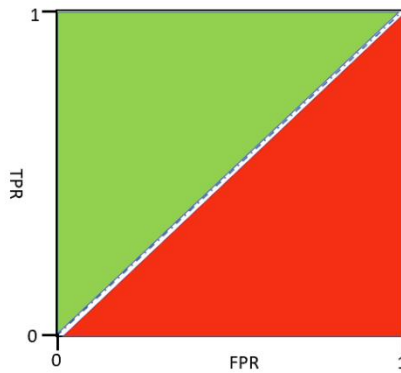


Figure 4: A Basic ROC Plot

Visualization

One of the primary objectives of this study has been to visualize the DRP model performance over time. A secondary goal is to enable users to interact with the graphs. Visualization and interaction of a complex machine learning algorithm output, such as Random Forest, are generally easier to interpret than numerical output. A business analytics tool—Power BI [8] was chosen to implement visualization and interaction. Power BI is a software that can be used to visualize data and share insights across an organization [9]. There are four main components of Power BI: data, datasets, reports, and dashboards [10].

Data and Dataset

In this study, each of the 133 factors in the data was originally in two different Excel files. One file describes the weekly DRP model prediction for the factor in 12 different prediction horizons (F1M—F12M). Another file describes the actual return for the factor. The data are available on a weekly basis from 7/26/1996 to 1/26/2018. The model prediction file was then merged with the corresponding actual return file using RStudio to form the final dataset (shown in Figure 5) that was used to create Power BI reports.

FACTORS	PERIOD	Horizon	Sector	Actuals	Actuals_b	Pred	Prob1	Prob2	Count1	Count2	Group
ACCX_F11	07/26/1996	F10M	Hardware	-0.56246	0	0	0.85124	0.14876	103	18	Quality
ACCX_F11	07/26/1996	F11M	Hardware	2.53221	1	0	0.876033	0.123967	106	15	Quality
ACCX_F11	07/26/1996	F12M	Hardware	9.57073	1	1	0.380165	0.619835	46	75	Quality
ACCX_F11	07/26/1996	F1M	Hardware	-0.70526	0	0	0.818182	0.181818	99	22	Quality
ACCX_F11	07/26/1996	F2M	Hardware	-0.01782	0	0	0.950413	0.049587	115	6	Quality
ACCX_F11	07/26/1996	F3M	Hardware	-3.07721	0	0	0.867769	0.132231	105	16	Quality
ACCX_F11	07/26/1996	F4M	Hardware	-8.34896	0	0	0.867769	0.132231	105	16	Quality
ACCX_F11	07/26/1996	F5M	Hardware	-4.04189	0	0	0.900826	0.099174	109	12	Quality
ACCX_F11	07/26/1996	F6M	Hardware	1.12994	1	0	0.950413	0.049587	115	6	Quality
ACCX_F11	07/26/1996	F7M	Hardware	0.8411	1	0	0.842975	0.157025	102	19	Quality
ACCX_F11	07/26/1996	F8M	Hardware	2.05544	1	1	0.429752	0.570248	52	69	Quality
ACCX_F11	07/26/1996	F9M	Hardware	1.43731	1	0	0.859504	0.140496	104	17	Quality
ACCX_F11	08/02/1996	F10M	Hardware	7.86716	1	0	0.826446	0.173554	100	21	Quality
ACCX_F11	08/02/1996	F11M	Hardware	7.02912	1	0	0.53719	0.46281	65	56	Quality
ACCX_F11	08/02/1996	F12M	Hardware	8.29036	1	1	0.446281	0.553719	54	67	Quality
ACCX_F11	08/02/1996	F1M	Hardware	-0.88849	0	0	0.677686	0.322314	82	39	Quality
ACCX_F11	08/02/1996	F2M	Hardware	2.0153	1	0	0.710744	0.289256	86	35	Quality
ACCX_F11	08/02/1996	F3M	Hardware	-1.27632	0	0	0.677686	0.322314	82	39	Quality
ACCX_F11	08/02/1996	F4M	Hardware	-1.3121	0	0	0.801653	0.198347	97	24	Quality
ACCX_F11	08/02/1996	F5M	Hardware	1.64352	1	0	0.818182	0.181818	99	22	Quality
ACCX_F11	08/02/1996	F6M	Hardware	8.27921	1	0	0.545455	0.454545	66	55	Quality
ACCX_F11	08/02/1996	F7M	Hardware	5.15282	1	0	0.553719	0.446281	67	54	Quality
ACCX_F11	08/02/1996	F8M	Hardware	6.30939	1	0	0.520661	0.479339	63	58	Quality
ACCX_F11	08/02/1996	F9M	Hardware	8.94952	1	0	0.520661	0.479339	63	58	Quality
ACCX_F11	08/09/1996	F10M	Hardware	7.92633	1	1	0.371901	0.628099	45	76	Quality
ACCX_F11	08/09/1996	F11M	Hardware	11.02169	1	1	0.256198	0.743802	31	90	Quality
ACCX_F11	08/09/1996	F12M	Hardware	14.28331	1	1	0.239669	0.760331	29	92	Quality
ACCX_F11	08/09/1996	F1M	Hardware	2.05661	1	0	0.727273	0.272727	88	33	Quality
ACCX_F11	08/09/1996	F2M	Hardware	7.37093	1	1	0.396694	0.603306	48	73	Quality
ACCX_F11	08/09/1996	F3M	Hardware	-1.36536	0	0	0.793388	0.206612	96	25	Quality
ACCX_F11	08/09/1996	F4M	Hardware	4.64444	1	0	0.859504	0.140496	104	17	Quality
ACCX_F11	08/09/1996	F5M	Hardware	6.29541	1	1	0.446281	0.553719	54	67	Quality

Figure 5: *Example of Final Dataset*

Reports

Two Power BI reports were created in this study to visualize the performance of the DRP model. Each report contains different performance measure visualizations. The first report includes the visualization of MCC, ACC, ROC, model performance on different forecast horizon, actual return over time, and actual factor return by sector. The second report visualizes the distribution of positive and negative prediction, the magnitude of predicted return, and the volatility of prediction and actual value over time. Each tab on the reports has different filters that can help users interact with the report by filtering certain information shown in the graphs.

Dashboard

A Power BI dashboard is one page of visualizations that enables the user to tell a story of the data and navigate through different reports. One dashboard (show in Figure 6) was created in this study by showing the highlights of each report. Each graph on the DRP dashboard is clickable. The user is able to navigate to different report tabs by clicking those graphs.

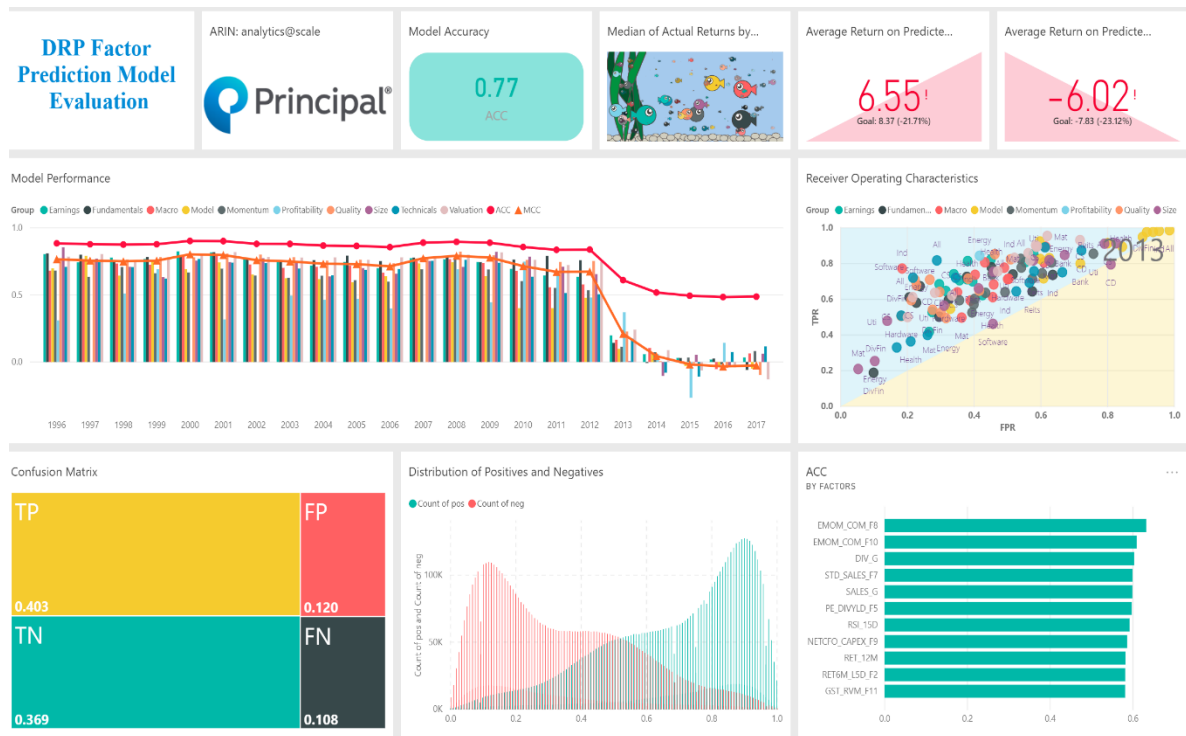


Figure 6: *DRP Dashboard*

CHAPTER 3. RESULTS

In order to better capture the behavior of DRP model, the model timeline was divided into two parts—Research and Production. From 1996 to 2013 the model was in the research stage, which means the DRP model was not used to select stocks in the real production of forming a portfolio. This timeframe was called “Research”. After 2013, the model was launched in production, which means the portfolio managers were using it to select stocks. The timeline after 2013 was called “Production”. After visualizing the standard validation metric in Power BI, this study was able to uncover the insight of DRP prediction model. Overall, the model was able to accurately predict stock performance in the research timeline (from 1996 to 2013). However, in the production timeline (after 2013), the model’s performance was close to random guess, which indicates the DRP prediction model completely failed to predict stock performance.

ACC and MCC

Figure 7 is the overall accuracy and MCC. The X-axis is a timeline, Y-axis is the model accuracy and MCC. The graph provides an overall evaluation of the DRP factor prediction model. Users have the ability to select different forecast horizon, timeline, sector, and factor group.

As shown in Figure 7, from 1996 to 2018, the DRP model had an average accuracy above 77%, which indicates overall it was a great prediction model. However, in the production timeline, the accuracy dropped to 50%, which implies the model was not able to forecast stock performance. When focusing on MCC, the overall average was 0.586, which indicates there was a strong positive relationship between actuals and predictions. The

positive relationship between actuals and predictions was even stronger in research timeline since MCC was 0.718. However, in the production timeline the average MCC is -0.01, which indicates there is a negligible relationship between actuals and predictions. In other words, the DRP prediction model was randomly guessing the stock performance after 2013.

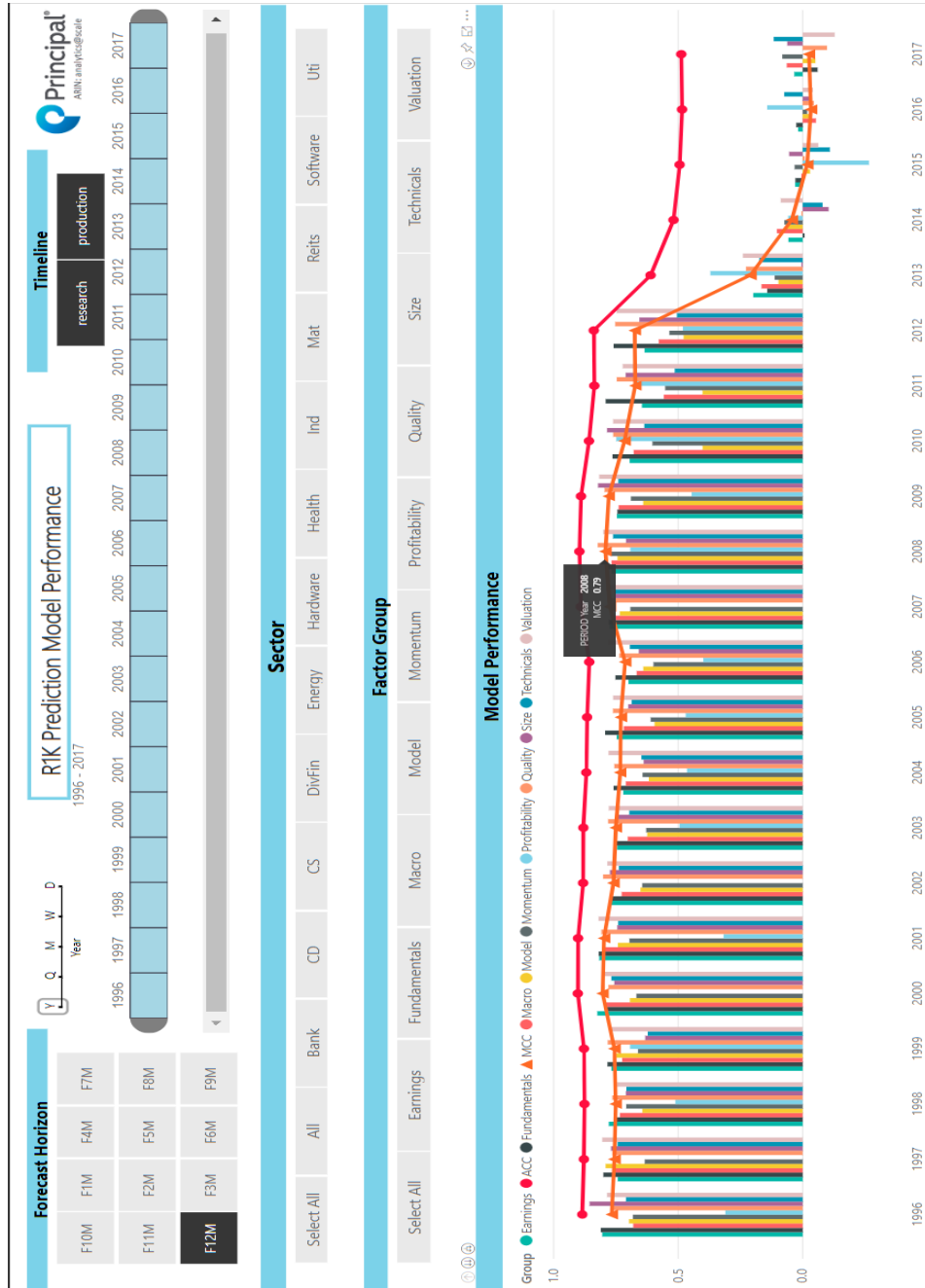


Figure 7: Overall ACC and MCC Plot

ROC

Figure 8 and Figure 9 are the ROC plots. True Positive Rate is plotted on the Y-axis and False Positive Rate is plotted on the X-axis. The best possible prediction would yield a point in the upper left corner or coordinate (0,1). Any point in the blue area is better than a random guess, which would give a point along a diagonal line. Points below the line (in the yellow area) represent worse than random. Another dynamic timeline was added into the graph, it shows how the model performance changes along the timeline.

In the research timeline (show in Figure 8), the majority of points were in the blue area, which means the model was correctly predicting the stock performance. However, when focusing on production timeline in Figure 9, the points were scattered in the ROC space, which indicates the model was randomly guessing stock performance.

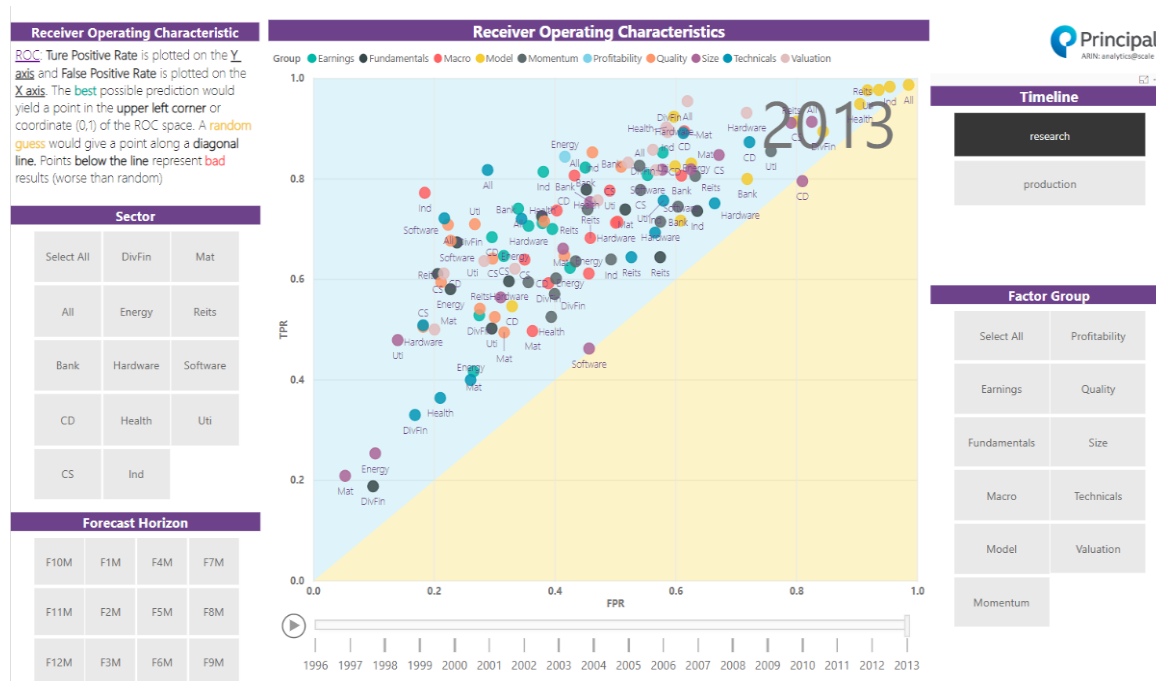


Figure 8: ROC Plot from 1996 to 2013

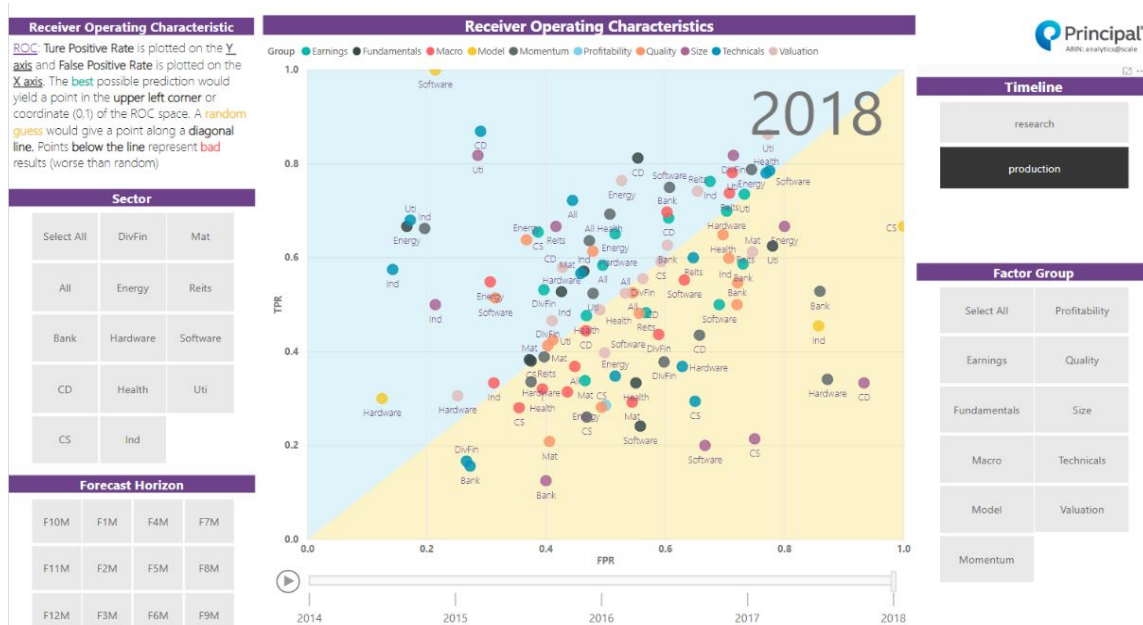


Figure 9: ROC Plot from 2014 to 2018

Confusion Matrix

As shown on the confusion matrix page in Figure 10, users can filter different forecast horizons and years to drill down the data shown in the plot. The table next to the confusion matrix provides detailed accuracy information by sector and factor group. The plot at the bottom is showing one of the most important performance measures—FPR. This measure is important because investing in an underperformed stock will harm the portfolio more than not investing in an overperformed stock. The goal of the DRP model is to keep the FPR as low as possible.

In the research timeline in Figure 10, the model accuracy was 0.77, and the FPR line was low. However, in the production timeline (show in Figure 11) the average model accuracy and FPR were close to 0.5, which indicates the model failed to predict stock performance.

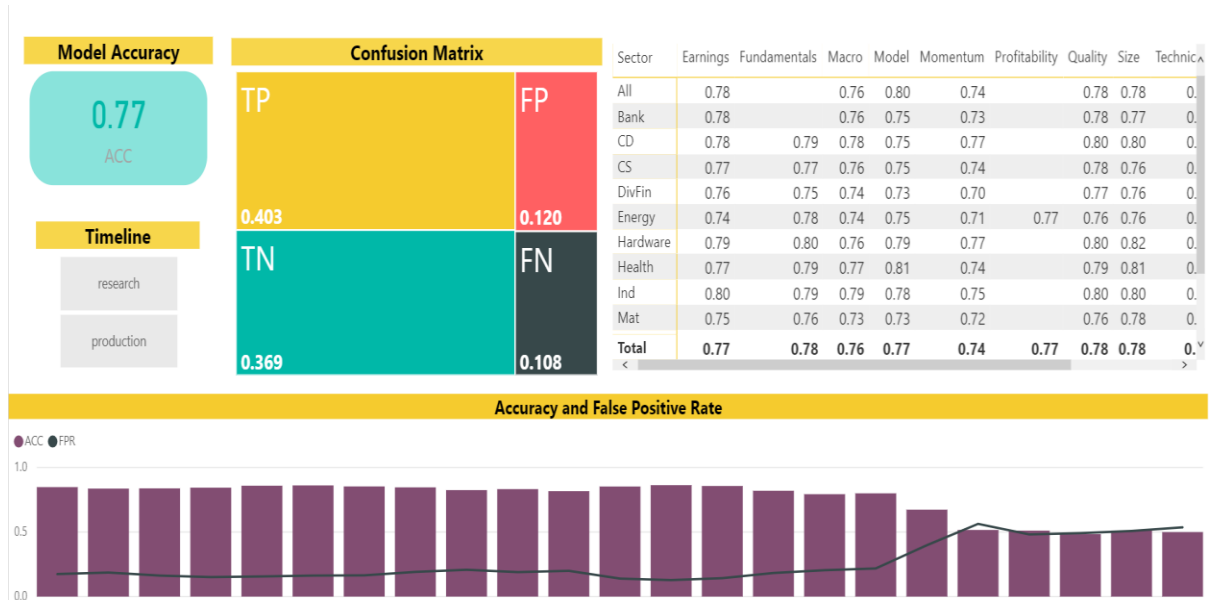


Figure 10: Confusion Matrix and FPR Plot

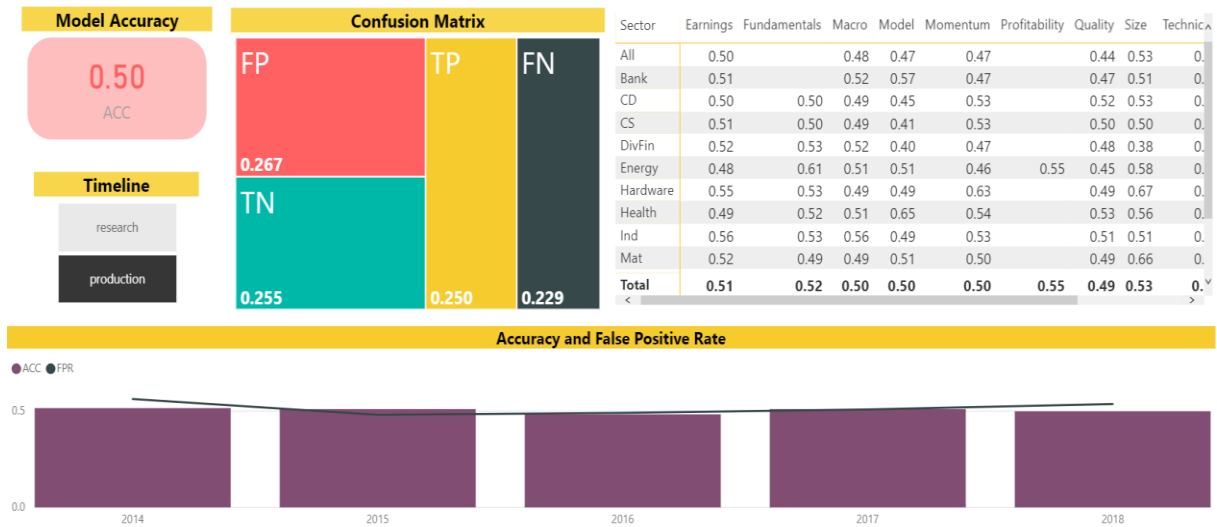


Figure 11: Confusion Matrix and FPR Plot from 2014 to 2018

Classes Distribution Plot

Figure 12 is a distribution plot, which provides a direct view of how the model distinguishes outperformance (positive) factors and underperformance (negative) ones. The predicted outperformance probability is plotted on the X-axis. The yellow line is the cutoff point (0.55), which means if the outperformance probability is higher than 0.55, the model will classify it as the outperformed factor. Red distribution is the count of actual underperformance factors. Green distribution is the count of actual outperformance factors. The overlapping part is the misclassification class.

In the research timeline (show in Figure 12), the model was able to distinguish positive class and negative class. However, as Figure 13 shows, in the production timeline, the positive and negative classes were completely overlapped with each other. The model failed to distinguish two classes.

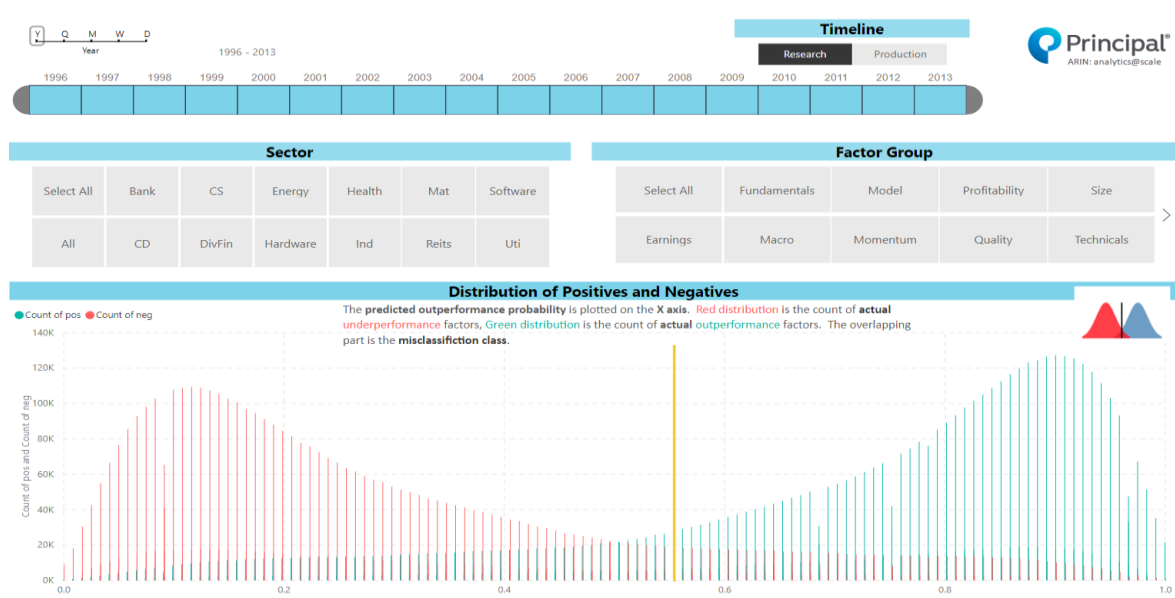


Figure 12: *Distribution Plot of Positive and Negative Classes*

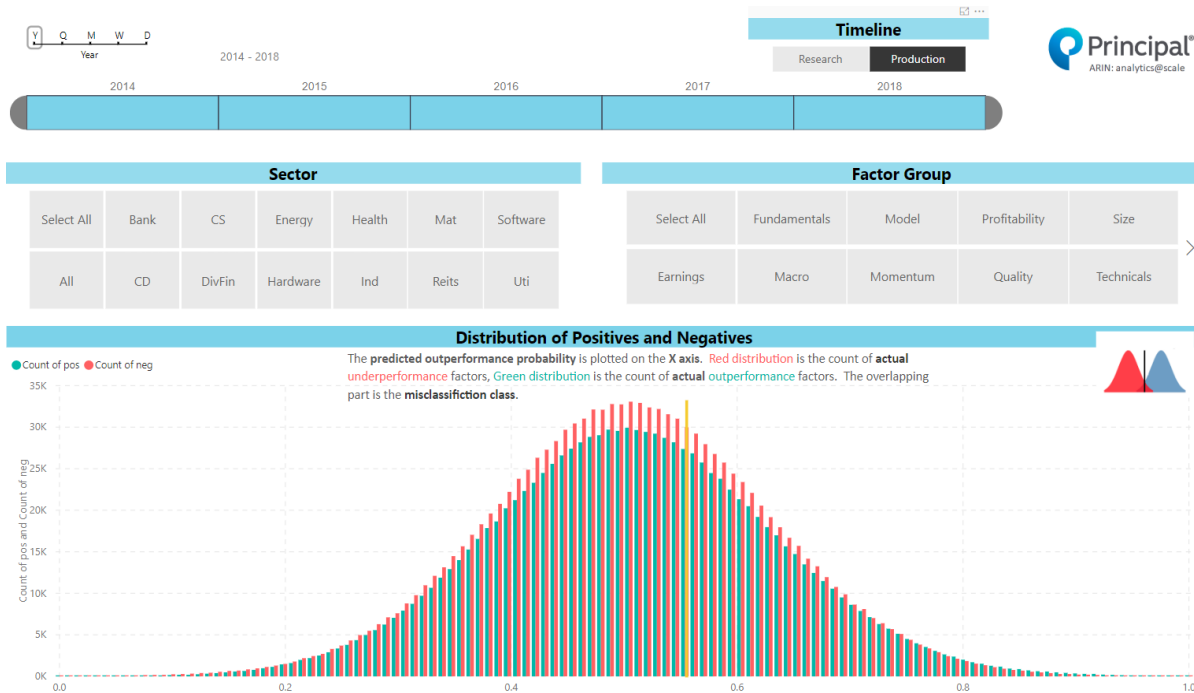


Figure 13: *Distribution Plot of Positive and Negative Classes from 2013 to 2018*

Magnitude Comparison

The magnitude comparison plot (shown in Figure 14) compares the average return on predicted underperformance or outperformance factors with the goal: average return on actual underperformance or outperformance factors. The expected actual return for the underperformance factors is negative, and for the outperformance factors is positive. The plot also shows the percentage of the prediction diverting from the goal. These comparisons provide a view to evaluating the magnitude of the model results.

In the research timeline in Figure 14, the deviation between prediction and actuals was approximately 24% for both underperformance and outperformance factors, which indicates the prediction was only 24% off target. However, in the production timeline shown in Figure 15, the magnitude of both underperformance and outperformance prediction was

approximately 100% off the actual, which indicates the DRP model failed to predict stock performance.

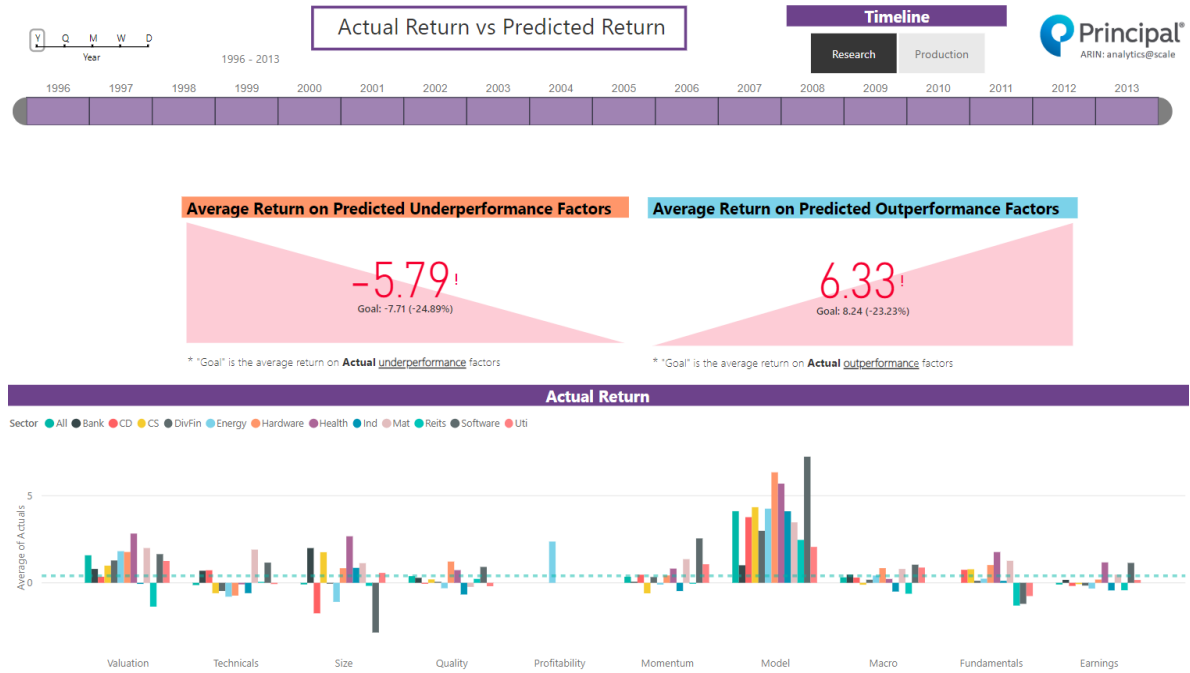


Figure 14: *Magnitude Comparison*

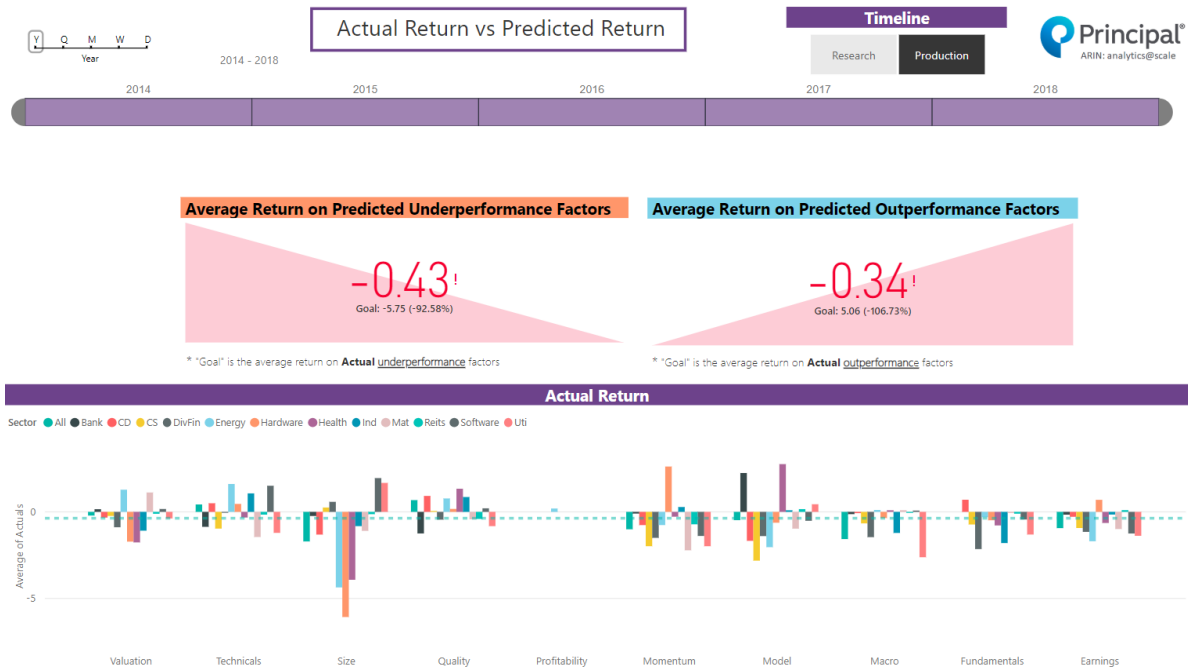


Figure 15: *Magnitude Comparison from 2014 to 2018*

Volatility of Prediction Outcome

Figure 16 shows two volatility plots. The top one is prediction volatility plot with the Y-axis indicating the predicted outperformance probability. The bottom one is the actual volatility plot with the Y-axis showing the median value of normalized actual returns. These two plots visualized how the prediction and actual fluctuate over time. Ideally, the amount of volatility will be the same in both research and production timeline.

As prediction volatility plot shows, the model fluctuated a lot in the research timeline. After 2013, the model suddenly fluctuated less. It indicates that some parameters in the model were very sensitive in the research timeline but stopped working in the production timeline. Since the algorithm modification is not within the scope of this study, another project or research is needed to identify the parameter and modify the Random Forest.

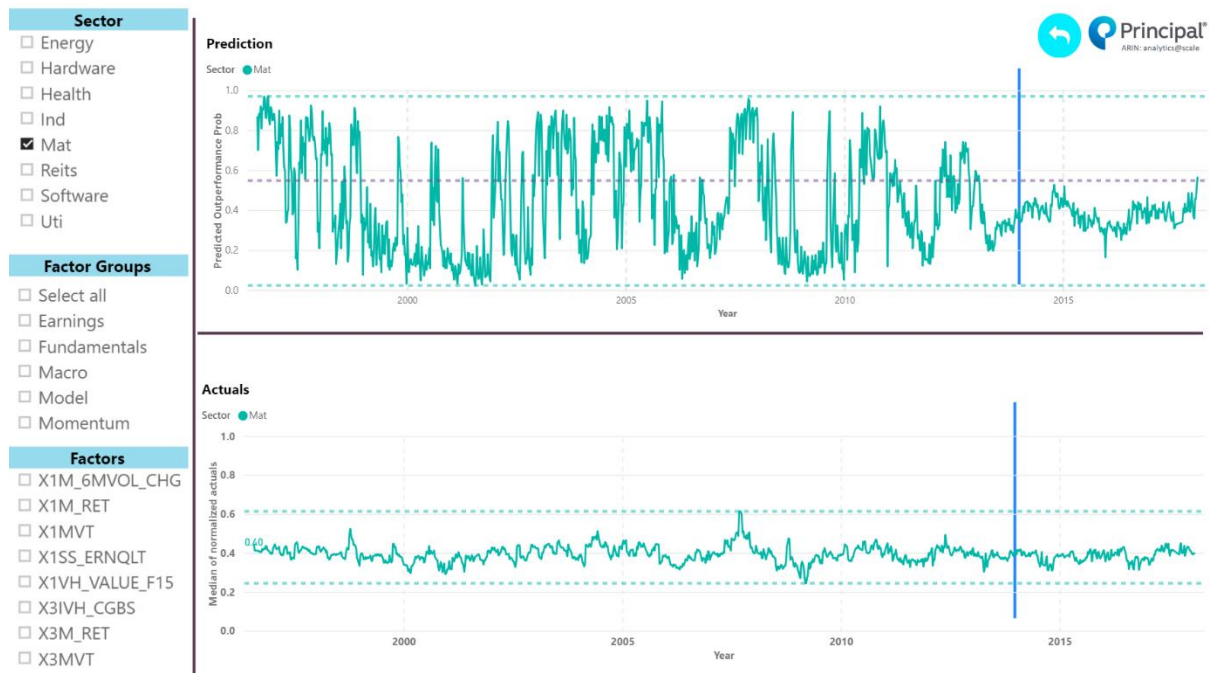


Figure 16: *Volatility of Prediction Outcome*

Model Comparison

Although the overall performance of the DRP prediction model is not ideal, the most important forecast horizons for Principal are F6M and F12M. A ribbon chart was designed to visualize the model performance on each forecast horizon. As shown in Figure 17, forecast horizons are ordered based on the highest accuracy in each factor group. The ribbon chart provides a direct view of the model accuracy among different forecast horizons.

In the research timeline, it was clear that the DRP model was better at predicting longer forecast horizons. However, in the production timeline shown in Figure 18, there was no clear evidence about which forecast horizon the model is able to predict more accurate.

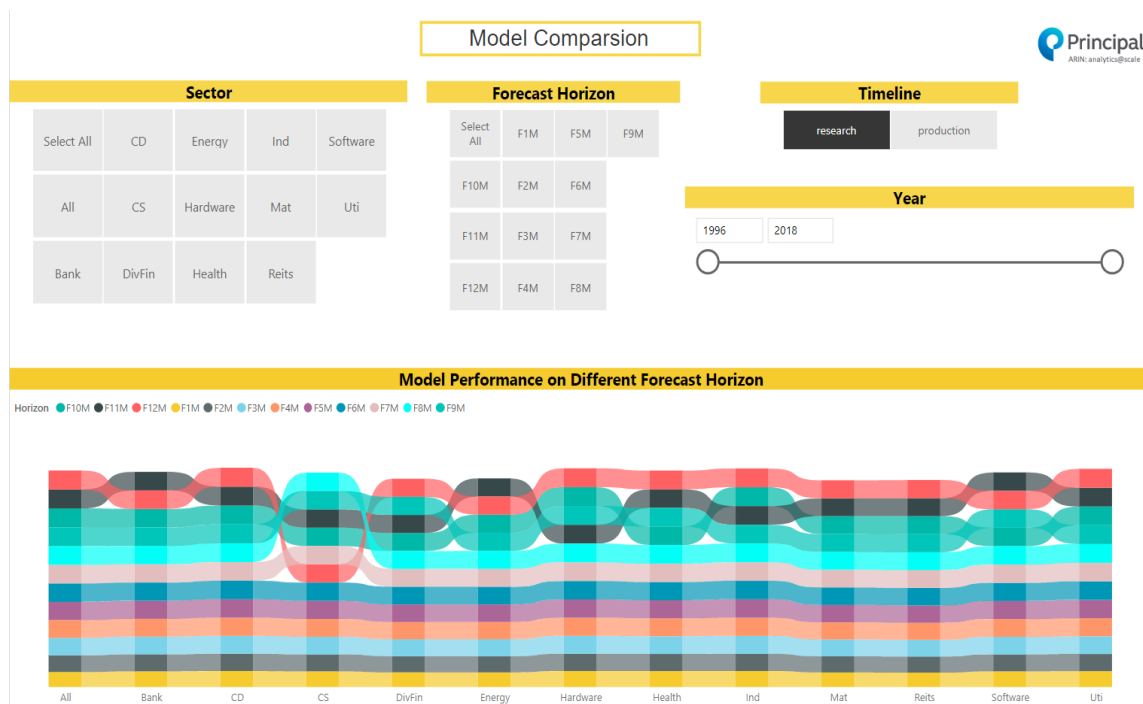


Figure 17: *Model Comparison between Different Forecast Horizon*

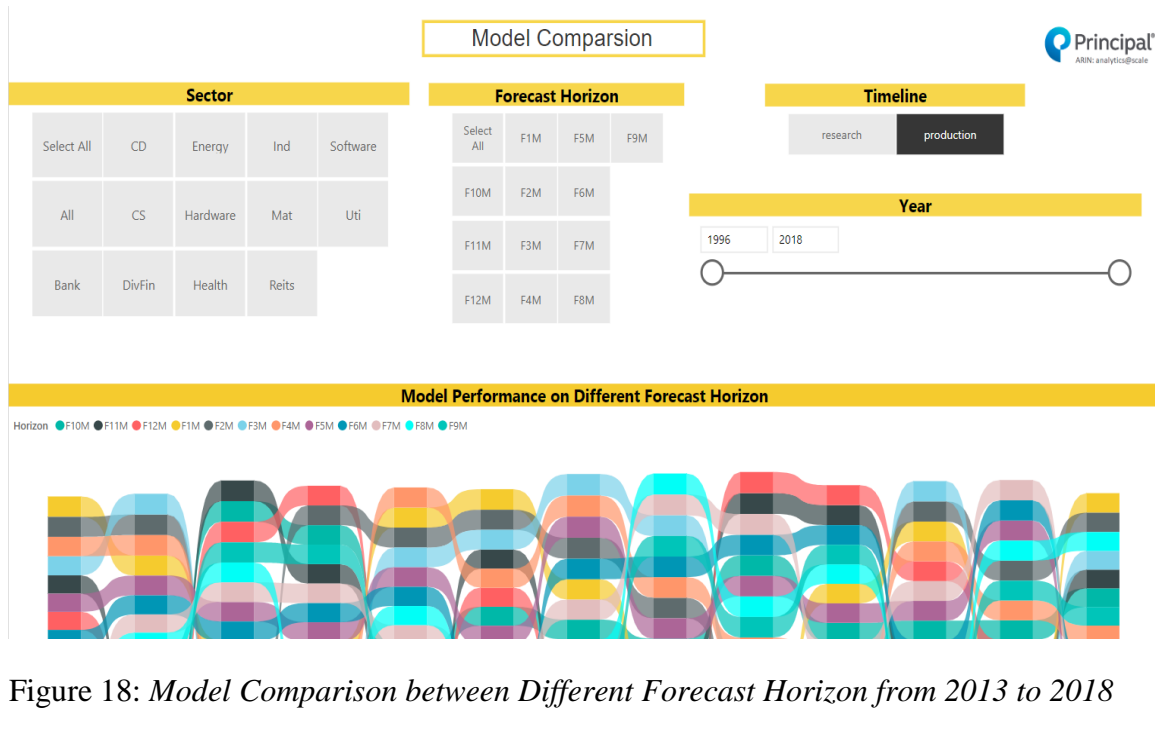


Figure 18: *Model Comparison between Different Forecast Horizon from 2013 to 2018*

Actual Return Aquarium

An actual return aquarium was added to the DRP dashboard (show in Figure 19). The data plotted in the aquarium are the actual returns for each factor. This visualization is not a validation of the DRP model, but it provides insight into the actual return and enables users to easily interpret the data. Each fish represents a factor, where the bigger size indicates higher actual returns. “Dead fish” represent negative returns, and eventually, all “dead fish” will be at the top of the aquarium. By looking at the fish tank, users are able to tell a story beyond the data itself.

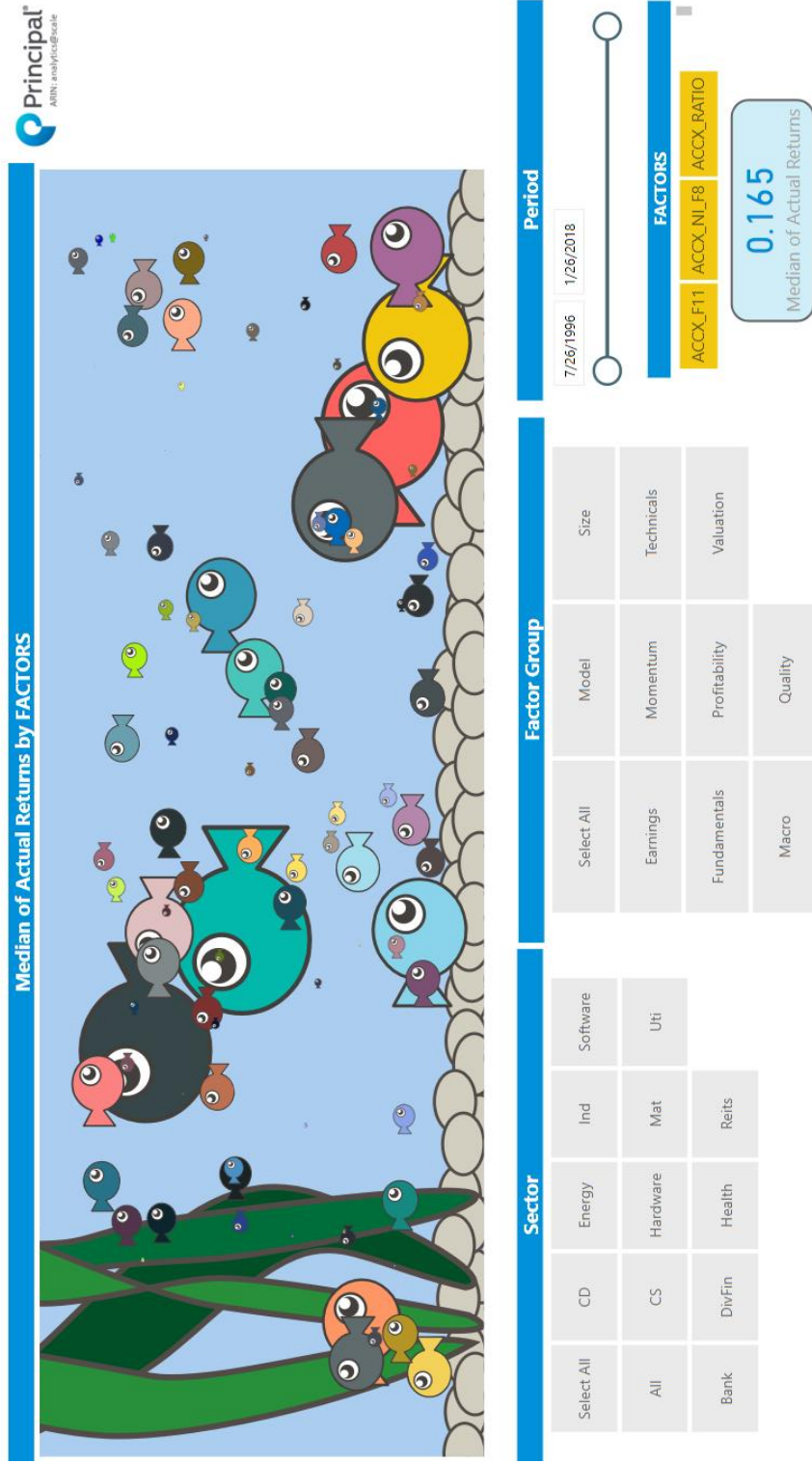


Figure 19: Actual Return Aquarium

CHAPTER 4. CONCLUSION

This study researched nineteen different binary classification performance measure methods and selected eight of them to construct a standard validation metric for the Principal Dynamic Risk Premia (DRP) prediction model. The selected performance measure methods were: Accuracy (ACC), Confusion Matrix, Distribution of Classes, False Positive Rate (FPR), Magnitude Comparison, Receiver Operating Characteristic (ROC), True Positive Rate (TPR), and Volatility. The standard validation metric quantified the confidence level of the prediction outcome from multiple aspects and controlled the input data for the weighing model in the DRP process since the outcome of the prediction model is the input of the weighing model.

After finalizing the standard validation metric, a Power BI dashboard allowing users to interact with visualizations was created to visualize the results of the standard validation metric over time. The visualizations allow for a straightforward interpretation of the model performance and a comparison between the current algorithm and alternative binary classification algorithms. In addition, this validation metric and the dashboard can also be applied to any balanced binary classifiers.

The results of this study demonstrated that the performance of the current DRP prediction model was outstanding from 1996 to 2012. However, after 2013, when it was launched in production, the performance was close to a random guess. In other words, selecting stocks based on the DRP prediction model after 2013 was the same as selecting stocks by flipping a coin. This study clearly quantified the DRP model performance over time and has enabled Principal to compare alternative algorithms using the same standard validation metric. After reviewing the results of this study, Principal decided to launch a new


project called DRP 2.0 to research potential algorithms and recreate the DRP prediction model. This study will be continuously used to validate the new DRP prediction model.

REFERENCES

- [1] L. Breiman, *Machine Learning*, vol. 45, no. 1, pp. 5-32, 2001. [Abstract]. Available: 10.1023/a:1010933404324.
- [2] S. Stehman, "Selecting and interpreting measures of thematic classification accuracy", *Remote Sensing of Environment*, vol. 62, no. 1, pp. 77-89, 1997. Available: 10.1016/s0034-4257(97)00083-7.
- [3] K. Horak, J. Klecka, O. Bostik and D. Davidek, "Classification of SURF image features by selected machine learning algorithms", *2017 40th International Conference on Telecommunications and Signal Processing (TSP)*, 2017. Available: 10.1109/tsp.2017.8076064.
- [4] S. Das and U. Cakmak, *Hands-on automated machine learning*, 1st ed. Birmingham UK: PACKT PUBLISHING, 2018, pp. 54-56.
- [5] D. Chicco, "Ten quick tips for machine learning in computational biology", *BioData Mining*, vol. 10, no. 1, 2017. Available: 10.1186/s13040-017-0155-3.
- [6] D. Powers, "Evaluation: from Precision, Recall and F-measure to ROC, Informedness, Markedness and Correlation", *Journal of Machine Learning Technologies*, vol. 2, no. 1, pp. 37-63, 2011. Available: https://bioinfopublication.org/files/articles/2_1_1_JMLT.pdf.
- [7] T. Fawcett, "An introduction to ROC analysis", *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861-874, 2006. Available: 10.1016/j.patrec.2005.10.010.
- [8] Microsoft, USA. 2016. Power BI. Available: <https://powerbi.microsoft.com/en-us/>.
- [9] "What is Power BI | Microsoft Power BI", *Powerbi.microsoft.com*, 2019. [Online]. Available: <https://powerbi.microsoft.com/en-us/what-is-power-bi/>.
- [10] V. Negru, "POWER BI: EFFECTIVE DATA AGGREGATION", *Quaestus*, no. 13, pp. 146-152, 2018.

APPENDIEX

The appendix is showing the M code used in transforming the Excel dataset structure to Power BI report data structure.

 Advanced Editor

consolidatedData

```
let
    Source = Csv.Document(File.Contents("C:\Users\1868598\Documents\Data\consolidatedData.csv"),[Delimiter=";", Columns=12, Encoding=1252, QuoteStyle=QuoteStyle.None]),
    #"Promoted Headers" = Table.PromoteHeaders(Source, [PromoteAllScalars=true]),
    #"Changed Type" = Table.TransformColumnTypes(#"Promoted Headers",{{{"FACTORS", type text}, {"PERIOD", type date}, {"Horizon", type text}, {"Sector", type text}, {"Actuals", type number}, {"Actuals_binary", Int64.Type}, {"Pred", Int64.Type}, {"Prob1", type number}, {"Prob2", type number}, {"Count1", Int64.Type}, {"Count2", Int64.Type}, {"Group", type text}}}},
    #"Added Conditional Column" = Table.AddColumn(#"Changed Type", "T", each if [Actuals_binary] = [Pred] then "T" else null),
    #"Added Conditional Column1" = Table.AddColumn(#"Added Conditional Column", "F", each if [Actuals_binary] <> [Pred] then "F" else null),
    #"Added Custom" = Table.AddColumn(#"Added Conditional Column1", "CM", each if [Actuals_binary] = 0 and [Pred] = 1
then
    "FP"
else if [Actuals_binary] = 0 and [Pred] = 0
then "TN"
else if [Actuals_binary] = 1 and [Pred] = 0
then "FN"
else "TP"),
    #"Filtered Rows" = Table.SelectRows(#"Added Custom", each [Actuals] <> null and [Actuals] <> ""),
    #"Filtered Rows1" = Table.SelectRows(#"Filtered Rows", each [Actuals_binary] <> null and [Actuals_binary] <> ""),
    #"Added Conditional Column2" = Table.AddColumn(#"Filtered Rows1", "TP", each if [CM] = "TP" then "TP" else null),
    #"Added Conditional Column3" = Table.AddColumn(#"Added Conditional Column2", "TN", each if [CM] = "TN" then "TN" else null),
    #"Added Conditional Column4" = Table.AddColumn(#"Added Conditional Column3", "FP", each if [CM] = "FP" then "FP" else null),
    #"Added Conditional Column5" = Table.AddColumn(#"Added Conditional Column4", "FN", each if [CM] = "FN" then "FN" else null),
    #"Filtered Rows2" = Table.SelectRows(#"Added Conditional Column5", each [Actuals_binary] <> null and [Actuals_binary] <> ""),
    #"Filtered Rows3" = Table.SelectRows(#"Filtered Rows2", each [Actuals] <> null and [Actuals] <> ""),
    #"Added Custom1" = Table.AddColumn(#"Filtered Rows3", "id", each Text.From([Pred]) & ", " & Text.From([Actuals_binary])),
    #"Reordered Columns" = Table.ReorderColumns(#"Added Custom1",{"id", "FACTORS", "PERIOD", "Horizon", "Sector", "Actuals", "Actuals_binary", "Pred", "Prob1", "Prob2", "Count1", "Count2", "Group", "T", "F", "CM", "TP", "TN", "FP", "FN"}),
    #"Removed Columns" = Table.RemoveColumns(#"Reordered Columns",{"Count1", "Count2"}),
    #"Added Conditional Column6" = Table.AddColumn(#"Removed Columns", "Timeline", each if [PERIOD] < #date(2014, 1, 3) then "research" else "production")
in
    #"Added Conditional Column6"
```

Figure 20: *M Code*