

UNIVERSITY OF MORATUWA

**Predictive Model for Gap Reduction
Between Web Analytics and Business
Strategy**

by

P.H.A. Nissanka (158230N)

A thesis submitted in partial fulfillment of the requirements for the
Degree of MSc in Computer Science specializing in Cloud Computing

in the
Faculty of Engineering
Department of Computer Science and Engineering

28th February 2019

Declaration

I declare that this is my own work and contains no material that has been published previously in whole or in part for the fulfillment of any degree program. All the referenced materials have been acknowledged in text.

Student:

Supervisor:

P.H.A. Nissanka (158230N)

Dr. Shantha Fernando

Abstract

Digital marketing and web analytics are two distinct areas that have captured the attention of many industrial firms. There are a lot of tools developed and a lot of studies carried out in each area separately. But still, a firm's ability to harness web analytics to optimize digital marketing elements is limited. This work focuses on evaluating previous work in each of these areas and combine them to build a model that would define the relationship between digital marketing and web analytics. Data captured through each area is expected to be analyzed in the form of a time series forecasting problem. Time series forecasting is a very popular area that captured a lot of firm's attention in recent years. This is due to the fact that most real-world problems are linked to a temporal component, and thus can be considered as a time series. Furthermore, this work utilizes cloud services for building and running the learning models.

Acknowledgements

I would like to express my sincere thank to my parents and my beloved wife for the continuous support given to manage my studies. And I would like to thank Dr. Shantha Fernando for the guidance and support given throughout the research to make it a success. Furthermore, I would like to thank Mr.Ashan Fernando for the encouragement and knowledge shared to continue this research until the end. Finally, my sincere appreciation goes to my colleagues at ShoutOUT Labs for the support given to manage my research work with office work.

Contents

Declaration	i
Abstract	ii
Acknowledgements	iii
List of Figures	vi
List of Tables	viii
Abbreviations	ix
1 Introduction	1
1.1 Background	1
1.2 Problem statement	2
1.3 Research questions	3
1.4 Objectives and output	3
1.5 Research strategy	4
1.6 Thesis outline	5
2 Literature Survey	6
2.1 Web analytics tools	6
2.2 Time series forecasting	7
2.3 Components of a time series	8
2.4 Feature engineering	9
2.5 Forecasting methodologies	10
2.5.1 Classical, statistical methodologies	10
2.5.2 Machine learning methodologies	11
2.5.3 Hybrid methodologies	13
2.6 Evaluation methodologies	14
2.6.1 Performance evaluation	14
2.6.2 Model evaluation	15
2.6.3 Uncertainty estimation	16
2.7 Summary	16
3 Methodology	17

3.1	Setting up the environment	18
3.1.1	Setting up the local environment	18
3.1.2	Setting up the cloud environment	19
3.2	Finding the dataset	19
3.3	Analyze and prepare the dataset	19
3.3.1	Filter the dataset	19
3.3.2	Visualize the dataset	20
3.3.3	Transform the dataset	21
3.4	Training the learning models	22
3.5	Hyperparameter tuning	24
3.6	Testing trained model	25
4	Implementation	27
4.1	Data loader	28
4.2	Model builder	29
4.3	Model evaluator	30
5	Performance evaluation	32
5.1	Workload	32
5.2	Experimental setup	34
5.3	Accuracy	35
5.4	Summary	35
6	Conclusions	36
6.1	Summary	36
6.2	Research limitations	37
6.3	Future work	37
	References	38
A	Data loader source	41
B	Model builder source	42
C	Model evaluator source	45

List of Figures

2.1	Decomposed time series	9
2.2	Window selection in walk forward validation	15
2.3	Uncertainty estimation	16
3.1	High level architecture	18
3.2	Campaign count variation over time	20
3.3	Website visits variation over time	21
3.4	DeepAR input format	21
3.5	Dataset separation as training and test	22
3.6	Selected hyperparameters	24
3.7	Prediction Results	26
4.1	Activities diagram of the process	28
4.2	Data loader pseudocode	29
4.3	Model builder pseudocode	30
4.4	Evaluator pseudocode	31
5.1	Visits variation of the evaluation dataset	33
5.2	Campaigns variation of the evaluation dataset	33
5.3	Selection of visit dataset windows	34
5.4	Selection of campaign dataset windows	34

5.5	Predictions from all sliding datasets	35
5.6	Predictions from three sliding datasets	35

List of Tables

3.1	Transformed dataset	20
3.2	DeepAR Tunable Hyperparameters	23
3.3	Metrics Computed by DeepAR Algorithm	24
3.4	Best Training Job Hyperparameters	25

Abbreviations

ANN	Artificial Neural Network
ARMA	Auto Regressive Moving Average
ARIMA	Auto Regressive Integrated Moving Average
AWS	Amazon Web Services
DeepAR	Deep Auto Regressive
GCP	Google Cloud Platform
GUI	Graphical User Interface
IaaS	Infrastructure as a Service
K-NN	K-Nearest Neighbour
LSTM	Long Short-Term Memory
MAD	Mean Absolute Deviation
MLMVN	Multi Layer Multi Valued Neurons
MSE	Mean Squared Error
PaaS	Platform as a Service
QRF	Quantile Random Forest
RNN	Recurrent Neural Network
S3	Simple Storage Service
SDK	Software Development Kit
TS	Time Series