

UNIVERSIDAD INTERNACIONAL DE LA RIOJA

MASTERS THESIS

Building a predictive model for CSR Backlog at Ericsson

Author:

Erika Meyer KVALEM SOTO

Supervisor:

Luis Miguel GARAY
GALLASATEGUI

A thesis submitted in fulfilment of the requirements

for the

Msc. BIG DATA AND VISUAL ANALYTICS

July 2020

unir LA UNIVERSIDAD
EN INTERNET



Declaration of Authorship

I, Erika Meyer KVALEM SOTO, declare that this thesis titled, 'Building a predictive model for CSR Backlog at Ericsson' and the work presented in it is my own. I confirm that this work submitted for assessment is my own and is expressed in my own words. Any uses made within it of the works of other authors in any form (e.g., ideas, equations, figures, text, tables, programs) are properly acknowledged at any point of their use. A list of the references employed is included.

Acknowledgements

To my grandmother Aogot Kvalem.

To all the covid-19 victims and health personnel.

I want to truthfully thank my family and my close friends who have always been supporting me and encouraging me to keep moving forward.

Big thanks to Gonçalo San Payo and Filip Gvardjan for their technical support on the machine learning issues. Thanks to the whole team of Automation and Applications at Ericsson led by Tamara Gomez. Thanks to my colleagues Kelmer Klimovas, Rafael Jose Pachon, Ivan Lara Javier Gismero for their unconditional help.

Abstract

The data handled by telecommunications companies nowadays is massive and of increasing complexity. Traditionally, network incidence management has been done in a reactive way, however this approach falls short when dealing with such large data load. For this purpose, a predictive procedure is needed. This work focuses on developing a predictive classification model for the duration of network incidence management. This process is currently being performed by telecommunication experts at Ericsson. This project deals with network incidences, called Customer Service Request (CSR) at Network Managed Services Delivery (NMSD) Support and Repair organization. The scope is narrowed only for the Customer Unit (CU) Iberia composed of Spain and Portugal. The predictive model will work on forecasting a classification of time ranges, for the time without answer of a CSRs. This is an internal Key Performance Indicator (KPI) known as Backlog. Being able to foresee this value, will allow to perform a pre-emptive network maintenance by automatically and statistically detecting early anomalies. Thus, optimizing resource allocation and budget planning by driving internal actions. Moreover, this prediction is useful information to provide to the customer. In this way the client will know approximately how long it will take until the incidence is solved. Providing the client with this knowledge will increase customer satisfaction and the quality of the service. The prediction of Backlog affects other KPIs such as Turn-Around-Time (TaT). TaT measures the total time since the CSR was opened until a Formal Answer(RST) is delivered. Backlog and TaT are KPIs closely monitored by Support and Repair organization to ensure CSR response time. Making predictions about them will allow to fulfill the high market expectations and remain competitive.

Contents

Declaration of Authorship	i
Acknowledgements	ii
Abstract	iii
Contents	iv
List of Figures	vi
List of Tables	vii
Abbreviations	viii
1 Introduction	1
1.1 Summary	1
1.2 Motivation	2
1.3 Objectives	3
2 Background	4
2.1 Predictive analysis	4
2.2 Regression techniques	4
2.2.1 Linear Regression	5
2.3 Machine learning	6
2.4 Classification	6
2.4.1 Logistic regression	7
2.4.2 Support Vector Machine	7
2.4.3 K-Nearest Neighbor	7
2.4.4 Ensemble methods	8
2.4.5 Random forest	8
2.5 Related Work	9
2.5.1 Predictive Analysis of Customer Service Requests (CSR) Volume	9
2.5.2 A Recommender System Architecture for Predictive Telecom Network Management	10
2.5.3 Feature engineering and analysis	11
2.5.4 Algorithm selection	12

2.6	Standard Process	13
3	Proposed methodology	16
3.1	Gathering	16
3.2	Execution	17
3.2.1	Data acquisition	17
3.2.2	Exploratory data analysis (EDA)	18
3.2.2.1	BigML	20
3.2.3	Extraction, transformation and loading (ETL)	24
3.2.3.1	Python	24
3.2.4	Feature Engineering	25
3.2.5	Building the model and Performance	26
3.2.5.1	Algorithms	26
3.2.5.2	Initial performance	29
3.2.5.3	Adjustment	30
3.2.5.4	Adjusted performance	34
3.2.6	Final Results	35
4	Results	38
4.1	75% accuracy threshold	38
4.2	Best scores	40
5	Discussion	41
5.1	Limitations for improvement	42
5.1.1	Limited amount of data.	42
5.1.2	Not optimized data entry.	43
5.1.3	Improve EDA and ETL processes	43
5.1.4	Exploration of other algorithms and strategies.	43
6	Conclusion	44
7	Future Work	46
A	Python code for ETL	48
B	Comprehensive final results	53
	Bibliography	55

List of Figures

2.1	An OLS regression model describing the best fit (1)	5
2.2	Predicted results for Customer Service Requests (CSR) Volume (2)	10
2.3	E-stream pattern matcher and predictor phase (3)	11
2.4	CSR Handling Process work flow at Ericsson	14
2.5	Applying machine learning to the CSR handling process to make predictions (4).	15
3.1	Duration (Days) distribution for each CSR ID	19
3.2	Distribution of the number of CSRs with same duration	19
3.3	BigML Model Summary Report. Field Importance for only “Finished” CSR and 9 selected features.	21
3.4	BigML Model Summary Report. Field Importance for Only “Finished” CSR + all columns.(80)	22
3.5	BigML Model Evaluation. For only “Finished” CSR and 10 selected features.	23
3.6	BigML Model Evaluation. For only all CSR Status and all columns (80).	23
3.7	Target variable ”Class” biased distribution.	25
3.8	Python analysis. Ranking of most important features using Pearson coefficient, Chi Squared, Recursive Feature Elimination, Logistic Regression Ridge Regularization L1, Logistic Regression Lasso Regularization L2, SelecFromModel Random Forest and SelectFromModel Linear Regression	26
3.9	Best k for KNN to obtain highest accuracy	34
3.10	Description of the process followed to develop the model	35
3.11	Sweetviz Automated EDA report. Example of the correlation matrix and some of the features analyzed	36
4.1	Visualization of the results with accuracy levels over 75%	39
4.2	F1 scores for the classes 1 and 2 obtained using the algorithms with over 75% of accuracy.	39
4.3	Graph of final best results for the classification algorithms.	40
5.1	(a) Downsampling redistribution of classes (b) Downsampling effects on 4 classes (c)Downsampling effects on 3 classes (d) Downsampling effects on 2 classes	42

List of Tables

3.1	Table summary of BigML Model Summary Report. Field Importance for only “Finished” CSR and 10 selected features.	21
3.2	BigML Model Summary Report. Field Importance for Only “Finished” CSR + all columns.(80)	22
3.3	Classification report. Comparing Only “Finished” CSR + All columns VS Only “Finished” CSR + 10 features. Accuracy and f1 score for class 1, 2, ,3 and 4 . .	29
3.4	Classification report. Only “Finished” CSR + All columns + Downsampling. Accuracy and f1 score for class 1, 2, ,3 and 4	30
3.5	Classification report. Only “Finished” CSR + All columns + Downsampling + Tuning + Feature selection Algorithms. Accuracy and f1 score for class 1, 2, ,3 and 4	31
3.6	Classification report. Only “Finished” CSR + All collumns + Downsampling + Tuning + Feature selection Algorithms + 3 classes. Accuracy and f1 score for class 1, 2, and 3.	34
3.7	Time ranges of the different approaches for 2,3, and 4 classes	37
4.1	Final Classification report for 2 classes.	38
4.2	Selected algorithms for each approach and scores.	40
B.1	Final Classification report for 4 classes.	53
B.2	Final Classification report for 3 classes.	54
B.3	Final Classification report for 2 classes.	54

Abbreviations

CSR	Customer Service Request
CU	Customer Unit
KPI	Key Performance Indicator
TaT	Turn-Around-Time
RST	Formal Answer
NMSD	Network Managed Service Delivery
CART	Classification and Regression Trees
PCA	Principal Component Analysis
GS	Global Support
EDA	Exploratory Data Analysis
LS	Local Support
OLS	Ordinary Least Squares
OvR	One Versus Rest
SVM	support Vector Machine
KNN	K-nearest neighbor algorithm
ARIMA	Auto regressive integrated moving average
RMSE	Root Mean Squared Error
OFI	Opportunities for improvement
EDA	Exploratory data analysis
ETL	Extraction, transformation and loading
SMOTE	Synthetic Minority Over-sampling Technique
LDA	Linear discriminant analysis
PCA	Principal Component Analysis