Predicting Incident Management Service Level Agreement (SLA) Failures

Capstone Written Report

Carolyn M. Hennings

Western Governors University

[chenn15@wgu.edu](mailto:chenn15@wgu.edu)

March 11, 2020

Abstract

IT Service Management practices aim at optimizing the efficiency and effectiveness of IT services provided to users. Incidents represent service disruptions. Service Level Agreements (SLA) establish thresholds for resolution of incidents within specified timeframes based on impact and urgency designations. Decreasing SLA breaches increases the availability of IT services and represents an important consideration for IT service providers.

This study explores indicators of incidents SLA breaches with respect to factors gathered during the early stages of an IT incident’s lifecycle. The study built a Logistic Regression model using Python and a number of tools from the SciKit-Learn library. Some supplementary analysis leveraged the R language. This paper describes the data collection, the data extraction and preparation, and the analysis steps performed throughout the study followed by a summary of findings, implications, and recommendations.

Table of Contents

[1 Research Question 1](#_Toc34671068)

[1.1 Justification for the Question 1](#_Toc34671069)

[1.2 Context 1](#_Toc34671070)

[1.3 Hypothesis Discussion 3](#_Toc34671071)

[2 Data Collection 4](#_Toc34671072)

[2.1 Collected Data 4](#_Toc34671073)

[2.2 Methodology 4](#_Toc34671074)

[2.2.1 Search 4](#_Toc34671075)

[2.2.2 Screen 5](#_Toc34671076)

[2.2.3 Select 5](#_Toc34671077)

[2.3 Methodology Advantages and Disadvantages 5](#_Toc34671078)

[2.4 Challenges 5](#_Toc34671079)

[3 Data Extraction and Preparation 6](#_Toc34671080)

[3.1 Approach 6](#_Toc34671081)

[3.1.1 Acquire 6](#_Toc34671082)

[3.1.2 Clean 6](#_Toc34671083)

[3.1.3 Engineer 6](#_Toc34671084)

[3.1.4 Filter and Bin 7](#_Toc34671085)

[3.2 Techniques and Tools 9](#_Toc34671086)

[4 Analysis 11](#_Toc34671087)

[4.1 Process Overview 11](#_Toc34671088)

[4.2 Calculations and Results 11](#_Toc34671089)

[4.2.1 Investigate Variable Encoding 11](#_Toc34671090)

[4.2.2 Explore Feature Selection 12](#_Toc34671091)

[4.2.3 Analyze Features 12](#_Toc34671092)

[4.2.4 Optimize Model 15](#_Toc34671093)

[4.3 Techniques and Tools Used 16](#_Toc34671094)

[5 Data Summary and Implications 19](#_Toc34671095)

[5.1 Discussion of Results 19](#_Toc34671096)

[5.2 Recommendations 20](#_Toc34671097)

[6 References 22](#_Toc34671098)

[7 Appendix Describing the Source Data Set 24](#_Toc34671099)

[8 Appendix of Executed Python Notebooks and Scripts 26](#_Toc34671100)

[8.1 Notebook: 01. Exploratory Data Analysis and Preliminary Cleaning 8.1-1](#_Toc34671101)

[8.2 Notebook: 02. Cleaning the Source Data Set 8.2-1](#_Toc34671102)

[8.3 Notebook: 03. Creating the Target Variable (SLAFail) 8.3-1](#_Toc34671103)

[8.4 Notebook: 04. Final Data Preparation 8.4-1](#_Toc34671104)

[8.5 Notebook: 05.01.c Bare Bones Analysis Using a Weight of Evidence Encoder 8.5-1](#_Toc34671105)

[8.6 Notebook: 05.02 Feature Selection KBest with ANOVA F-value Score Function 8.6-1](#_Toc34671106)

[8.7 Notebook: 06.01.b Optimize the Logistic Regression Model 8.7-1](#_Toc34671107)

Table of Exhibits

[Exhibit 1. Method for Data Collection – Advantages and Disadvantages 5](#_Toc34671108)

[Exhibit 2. Business Rule for SLAFail Target Variable 7](#_Toc34671109)

[Exhibit 3. Correlation Heatmap 8](#_Toc34671110)

[Exhibit 4. Final Set of Dependent Variables 8](#_Toc34671111)

[Exhibit 5. Data Extraction and Preparation Techniques – Advantages and Disadvantages 9](#_Toc34671112)

[Exhibit 6. Data Extraction and Preparation Tools – Advantages and Disadvantages 10](#_Toc34671113)

[Exhibit 7. Comparison of Model Metrics among Encoders 12](#_Toc34671114)

[Exhibit 8. Comparison of Model Metrics among Feature Selection Methods 12](#_Toc34671115)

[Exhibit 9. PCA Explained and Cumulative Variance 13](#_Toc34671116)

[Exhibit 10. Contribution to Variance in First Two Principal Components 13](#_Toc34671117)

[Exhibit 11. Hierarchical Cluster Analysis Dendrogram 14](#_Toc34671118)

[Exhibit 12. Hierarchical Cluster Analysis Stability Chart 14](#_Toc34671119)

[Exhibit 13. Identified Clusters 15](#_Toc34671120)

[Exhibit 14. Optimization Results 15](#_Toc34671121)

[Exhibit 15. ROC Curve for Optimized Model 16](#_Toc34671122)

[Exhibit 16. Techniques and Tools for Analysis – Advantages and Disadvantages 17](#_Toc34671123)

[Exhibit 17. Logistic Regression Model Coefficients 19](#_Toc34671124)

# Research Question

What factors predict Incident Management SLA compliance?

## Justification for the Question

In the context of Information Technology (IT) organizations providing IT services to customers, Incident Management practices and processes serve as critical customer satisfaction enablers. Incident Management aims to minimize the duration of interruptions in normal service operations while also minimizing the impact of those interruptions (Hanna, 2011, p. 29). Service Level Agreements (SLA) describe services and establish service level targets as negotiated and agreed upon between the service provider and the customer (Hanna, 2011, p. 54). Common SLA elements identify target thresholds for the duration of incidents from initial identification (opened) to restoration of service at normal operational levels (resolved). The swift resolution of business-impacting incidents represents a primary focus for IT service support organizations. The ability to proactively identify incidents at risk of failure to meet an SLA threshold, “SLA-at-Risk”, allows for decisions and actions that reduce the duration and severity of service disruptions.

Service Desk managers, those responsible for the Incident Management process within an IT organization, will benefit from the results of this study. With customization of the data feed to a specific environment, the model will assist in identifying characteristics of incidents at risk of failure to meet an SLA (Higgins, 2016). Within the IT Service Management (ITSM) lifecycle, the results of this study will also interest Problem Management and Continual Service Improvement practitioners.

## Context

IT organizations use ITSM systems to capture information about the execution of Incident Management processes. These systems produce logs containing details about incidents, for example, the steps taken to resolve them, the individuals involved with the incidents, the elements within the IT environment impacted by the incident, and timestamps for actions taken throughout the lifecycle of an incident.

This project investigated an extract of Incident Management data from an ITSM system to determine indicators of failure to meet an SLA threshold and develop a model for predicting those incidents. Insight into SLA-at-Risk conditions notify management of IT environment components requiring attention, similar to a customer churn analysis identifying characteristics for marketing attention.

A literature scan identified related research in the field of process mining, a combination of data mining with process science (Van Der Aalst, 2018, p. 15). Amaral et al. investigated attribute selection methods to build completion time prediction models (Amaral et al., 2019). Hinkka et al. evaluated feature selection algorithms by comparing classification accuracy and response times. The study used the Gradient Boosting Machine classification method and an approximation of the mutual information score across feature selection methods on two different data sets. Sarnovsky and Surma used Random Forests and Gradient Boosting Machine classifiers to identify incident sources and predict impacts (Sarnovsky & Surma, 2018). Malley leveraged split-plot Analysis of Variance (ANOVA) techniques to assess the “extent to which IT staff use of organizational knowledge generated from data warehouse analytical measures reduces the number of IT incidents over a 30-day period”. Buhler et al. used a multinomial logistic regression model to predict impact pattern categorizations (Buhler et al., n.d., p. 13). In summary, these studies focus on questions associated with the efficiency and effectiveness of incident management processes. In contrast, this study focuses on factors contributing to incident causing harm to continual delivery of services.

## Hypothesis Discussion

The hypotheses under study focus on identifying significant factors that indicate the probability of an IT support organizations ability to close an incident within agreed service level thresholds. The study leveraged logistic regression techniques to test the following hypotheses:

H0: Data contains no significant indicators about the final SLA status of an incident (the coefficients of a logistic regression model are zero, )

H1: Data contains significant indicators about the final SLA status of an incident (the coefficients of a logistic regression model are not zero, )

The factors under consideration include data about incidents available in the early stages of an incident’s lifecycle. If any one of the logistic regression model’s coefficients significantly differs from zero, the study will accept the alternative hypothesis (H1), otherwise the study will fail to reject the null hypothesis (H0).

# Data Collection

This section describes the collected data, the methodology used for collection, and associated challenges.

## Collected Data

A study investigating factors contributing to incident management SLA risk requires an extract from an ITSM system used by an IT organization for tracking incidents over a specific period. This study leveraged an existing, publicly-available data set used in the 2014 Business Processing Intelligence Challenge (BPIC) (*10th International Workshop on Business Process Intelligence 2014*, n.d.). While the challenge released four data sets, this project focused only on the Incident Records file (Van Dongen, 2014).The selected data set consists of 46,606 observations having 28 variables extracted from an ITSM system used by a bank located in the Netherlands (*Quick reference BPI Challenge 2014*, n.d.). The terms for use of the data set specify that “The user is allowed to remix, transform or build upon the data, but only for noncommercial purposes” (4TU.Centre for Research Data, 2016).

## Methodology

Data collection followed a three-step approach.

### Search

The first step in data collection for an analysis project involves identifying the location and availability of relevant data sets. Search criteria identified a candidate pool of data sets with relevancy to the topic of Incident Management in an IT Service Management context. Internet-based search tools used by this project to identify publicly available subject data sets included:

* Data.Gov (<https://www.data.gov/>)
* University of California, Irvine Machine Learning Repository (<https://archive.ics.uci.edu/ml/index.php>)
* Google Scholar (<https://scholar.google.com/>)
* Google Dataset Search (<https://datasetsearch.research.google.com/>)
* Kaggle Datasets (<https://www.kaggle.com/datasets>)

### Screen

Search identified two data sets for consideration, as listed below. Both data sets contain similar information.

* 2014 Business Processing Intelligence Challenge (BPIC) Incident Records file (Van Dongen, 2014)
* UCI Machine Learning Repository Incident management process enriched event log Data Set (Amaral et al., 2019)

### Select

This project used the first data set due to the availability of supporting documents such as a detailed, accurate data dictionary and Incident Management process documentation.

## Methodology Advantages and Disadvantages

Exhibit 1. Method for Data Collection – Advantages and Disadvantages

|  |  |  |
| --- | --- | --- |
| Method | Advantages | Disadvantages |
| Internet search for candidate data sets | Broad range of results | Required filtering out cybersecurity and transportation incidents |

## Challenges

Internet searches for publicly available data sets return a large variety of results. Careful formulation and iterative refinement of search queries contributed to subsequently limiting the results to smaller, more relevant results.

# Data Extraction and Preparation

This section describes the process for extracting and preparing the data for analysis followed by a discussion of the tools and techniques leveraged. See the appendices for a summary of the source data set as well as the executed notebooks showing performance of each data extraction and preparation step.

## Approach

Data extraction and preparation involved four steps.

### Acquire

The data set required downloading from the 2014 Business Processing Intelligence Challenge (BPIC) website located at <https://www.win.tue.nl/bpi/doku.php?id=2014:challenge>. Initial exploratory data analysis (EDA) identified a relatively clean data set. Preliminary data cleaning steps included: conversion of strings representing dates to datetime data type, removal of non-incident records, as well as removal of records with a status other than closed.

### Clean

Data profiling during the previous step revealed collinearity among some variables. This step addressed some collinearity through creating aggregated variables and noted other items for later consideration. The project addressed missing values by dropping records representing fewer than 4% of the total, setting values to zero, “Not Applicable”, and “Yes/No” where appropriate (Nisbet et al., 2009, pp. 50–75).

### Engineer

The original data source lacked a binary indicator for the target variable. The project engineered the target variable, SLAFail, by setting the value to 1, according to the business rules described in Exhibit 2. Stanford University IT provides an example of a similar business rule (*Measuring Response and Resolution Times in Remedy | University IT*, n.d.).

Exhibit 2. Business Rule for SLAFail Target Variable

|  |  |
| --- | --- |
| Priority | Time Between Opened and Resolved |
| 1 Very High | Greater than 240 minutes (4 hours) |
| 2 High | Greater than 480 (8 hours) |
| 3 Medium | Greater than 1440 (1 day) |
| 4 Low | Greater than 2880 (2 days) |
| 5 Very Low | Greater than 5760 (4 days) |

The engineered target variable, FailSLA, resulted in 30% of the cases showing a failure to meet the defined SLA (closing the incident within the specified time based on Priority). The study will use proportional stratified random sampling to split the data set for training and testing purpose (Tufféry, 2011, p. 90)

### Filter and Bin

This final data extraction and preparation step addressed the high dimensionality of datetime variables and restricted the data set based on timeframe.

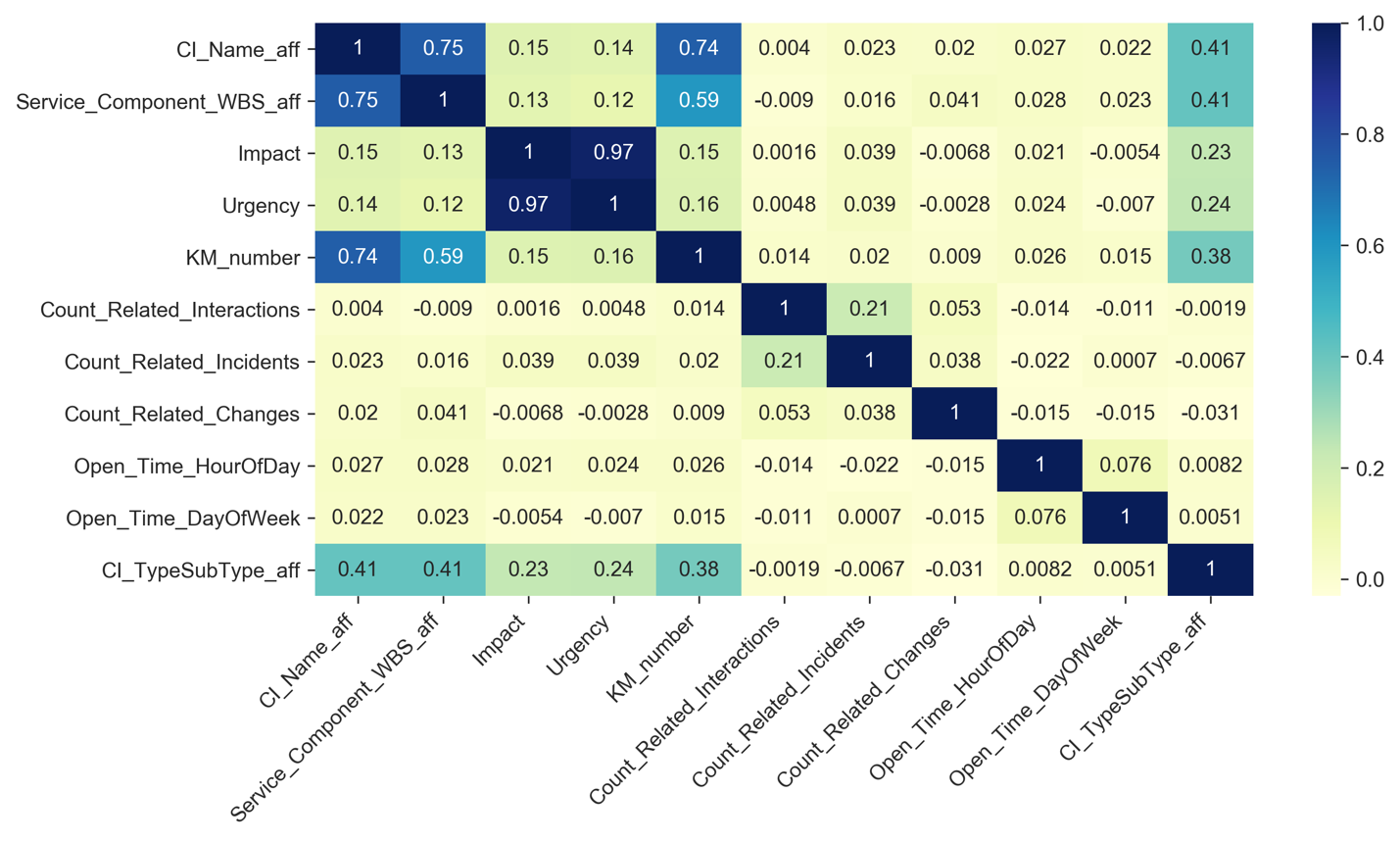
The datetime variables required binning (discretization) to reduce dimensionality (Tufféry, 2011, pp. 31–32). For each datetime variable (Open\_Time, Resolved\_Time, and Closed\_Time), this step created two corresponding binned variables. One for the hour of the day and the other for the day of the week.

This step reduced the data set to include only those incidents within a six-month time window (1 October 2013 through 31 March 2014). This ensures that the records cover the entire lifecycle of the incident, i.e. both open and close dates exist within the window (Buﬀett et al., 2014, p. 4).

Since the study proposes to predict SLA-at-Risk (SLAFail), this filtering step limited the data set used for model development to only those variables available upon creation of an incident record. Given the business case of an Incident Manager needing to identify the SLA-at-Risk incidents shortly after identification, little value would result from including data only available at later stages of an incident’s lifecycle in the model.

Collinearity negatively affects logistic regression models (Tufféry, 2011, pp. 86–87). A heatmap based on Pearson’s Correlation Coefficient showed significant correlation (> 0.70) among variables as shown in Exhibit 3.

Exhibit 3. Correlation Heatmap



The data set used for further analysis dropped the Impact and CI\_Name\_aff variables. Exhibit 4 lists the final data set of dependent variables.

Exhibit 4. Final Set of Dependent Variables

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | | Description | Type |
| KM\_number | Knowledge management article containing default attributes and questions for service desk analyst use | | Categorical |
| Urgency | Indicates incident resolution urgency | | Categorical |
| Count\_Related\_Interactions | Number of updates or changes to the incident record | | Continuous |
| Count\_Related\_Incidents | Number of similar or related incidents (child records) | | Continuous |
| Count\_Related\_Changes | Number of Change Management records associated with the incident | | Continuous |
| Open\_Time\_HourOfDay | Date and time of incident creation | | Categorical |
| Open\_Time\_DayOfWeek | Date and time of incident creation | | Categorical |
| CI\_TypeSubType\_aff | Concatenation of the top-level and second-level categories for the affected CI | | Categorical |
| Service\_Component\_WBS\_aff | Service component identifier for the affected CI | | Categorical |

## Techniques and Tools

Data extraction and preparation steps leveraged techniques described in Exhibit 5.

Exhibit 5. Data Extraction and Preparation Techniques – Advantages and Disadvantages

|  |  |
| --- | --- |
| Techniques | Justification |
| Convert data types | Need to manipulate data stored as strings or numbers for recognition as type required during analysis   |  |  | | --- | --- | | Advantages | Disadvantages | | Increases data consistency and integrity | Additional time required | |
| Aggregate variables | Need to consolidate multiple categorical variables into a single variable   |  |  | | --- | --- | | Advantages | Disadvantages | | Reduces collinearity | Additional time required | |
| Engineer target variable | Need to create missing variable based on existing data   |  |  | | --- | --- | | Advantages | Disadvantages | | Establishes required data point | Introduces business rule assumptions  Additional time required | |
| Bin datetime variables | Need to consolidate multiple continuous values into discrete number of levels   |  |  | | --- | --- | | Advantages | Disadvantages | | Reduce variable dimensionality | Additional time required | |
| Filter dataset by timeframe | Need to limit data set based on subject matter expertise   |  |  | | --- | --- | | Advantages | Disadvantages | | Reduce data | Introduces business case assumptions  Additional time required | |

Tools leveraged during this stage of the study included Python, pandas, NumPy, and Pandas Profiling for data mining and statistical analysis. The mature, stable, well-documented nature of Python, the extensive availability of examples and tutorials, and availability of the selected libraries contributed to selecting Python for this project. Additionally, this project used Jupyter Notebooks to integrate discussions and narrative with the data analytics code. “For data scientists, Jupyter has emerged as a de facto standard, says Lorena Barba, a mechanical and aeronautical engineer at George Washington University in Washington DC” (Perkel, 2018). The availability of this computational notebook capability contributed to the selection of Python. Exhibit 6 describes the tools used for data extraction and preparation along with an advantage/disadvantage summary.

Exhibit 6. Data Extraction and Preparation Tools – Advantages and Disadvantages

|  |  |
| --- | --- |
| Tools | Description |
| Python (*Welcome to Python.org*, n.d.) | Programming language   |  |  | | --- | --- | | Advantages | Disadvantages | | Free, open source reduced costs | Learning curve added time | |
| pandas (*pandas—Python Data Analysis Library*, n.d.) | Data analysis and manipulation tool   |  |  | | --- | --- | | Advantages | Disadvantages | | Free, open source reduced costs | Learning curve added time | |
| NumPy (*NumPy—NumPy*, n.d.) | Scientific computing package for Python   |  |  | | --- | --- | | Advantages | Disadvantages | | Free, open source reduced costs | Learning curve added time | |
| Pandas Profiling (*Pandas\_profiling API documentation*, n.d.) | Generates data set profiles identifying data types, unique values, missing values, quantile statistics, histograms, correlations   |  |  | | --- | --- | | Advantages | Disadvantages | | Free, open source reduced costs  Ease of use reduced time | Version compatibility issues added time | |
| Jupyter Notebooks (*Project Jupyter*, n.d.) | Web-based application that integrates code, equations, visualizations, and narrative text   |  |  | | --- | --- | | Advantages | Disadvantages | | Free, open source reduced costs  Ease of use reduced time | Configuration issues added time | |

# Analysis

With the goal of identifying factors that predict SLA-at-Risk incidents, the study focused on developing a logistic regression model supported by factor analysis techniques.

## Process Overview

Analysis developed a logistic regression model that predicts the status of SLAFail. Given the binary nature of the dependent, target variable and mixed nature of the independent variables, Tufféry recommends applying logistic regression techniques as an appropriate predictive method (Tufféry, 2011, p. 170). Analysis investigated three feature-focused aspects: variable encoding, feature selection, and feature analysis. Optimization followed an iterative approach to refining the model. The standard procedure for each of the above steps included splitting the source data set into a training data set and a testing data set. The training data set provided the input for generating the model and the testing data set contributed to the generated evaluation metrics. Throughout the analysis process, decisions stemmed from review of classification accuracy rates and the Area Under the Curve (AUC) score obtained from the Receiver Operating Characteristic (ROC) curve diagnostic (Tufféry, 2011, pp. 454, 458).

## Calculations and Results

### Investigate Variable Encoding

The source data set contains both categorical and numeric data. This step investigated a variety of encoding techniques and tools. Exhibit 7 presents model evaluation metrics used to select an encoder for use in subsequent steps. The methodology executed the same steps while varying only the encoder used on categorical variables. This analyst chose to move forward with the Weight of Evidence (WOE) encoder given the greatest AUC. See the appendices for the WOE calculations performed and the results obtained.

Exhibit 7. Comparison of Model Metrics among Encoders

|  |  |  |  |
| --- | --- | --- | --- |
| Name | Cross Validation Accuracy | F1 Score | AUC |
| MEstimate Encoder | 0.73086 | 0.65857 | 0.77668 |
| Target Encoder | 0.73394 | 0.65680 | 0.77678 |
| WOE Encoder | 0.73776 | 0.66146 | **0.78063** |
| Helmert Encoder | 0.73678 | 0.66134 | 0.76650 |

### Explore Feature Selection

The study evaluated the effectiveness of a variety of automated feature selection techniques available from SciKit-Learn (*1.13. Feature selection—Scikit-learn 0.22.1 documentation*, n.d.). Holding all other parameters equal, analysis used the feature selection methods listed in Exhibit 8 and elected to move forward with the KBest F-Classification (ANOVA F-value) method due to the highest AUC. See the appendices for the KBest F-Classification (ANOVA F-value) calculations performed and the results obtained.

Exhibit 8. Comparison of Model Metrics among Feature Selection Methods

|  |  |  |  |
| --- | --- | --- | --- |
| Name | Cross Validation Accuracy | F1 Score | AUC |
| KBest F-Classification  (ANOVA F-value) | 0.737026 | 0.657922 | **0.780698** |
| KBest Mutual Information | 0.737107 | 0.654011 | 0.777989 |
| Recursive Feature Extraction | 0.735119 | 0.656594 | 0.779613 |

### Analyze Features

Considering the identification of discriminating variables as the primary purpose of the study, the project leveraged a variety of feature analysis techniques. This section describes the techniques used and presents the findings. Techniques used included Principal Component Analysis (PCA) and Hierarchical Cluster Analysis.

#### Principal Component Analysis (PCA)

While investigating feature variability, PCA results showed the need for eight principle components to explain at least 95% of the variance (Exhibit 9). Further analysis of the variance within the first two principal components revealed the following (Exhibit 10):

* Three variables account for the greatest variance within PC 01:

Service\_Component\_WBS\_aff

KM\_Number

CI\_TypeSubType\_aff

* Two variables account for the greatest variance within PC 02:

Count\_Related\_Incidents

Count\_Related\_Interactions

Exhibit 9. PCA Explained and Cumulative Variance

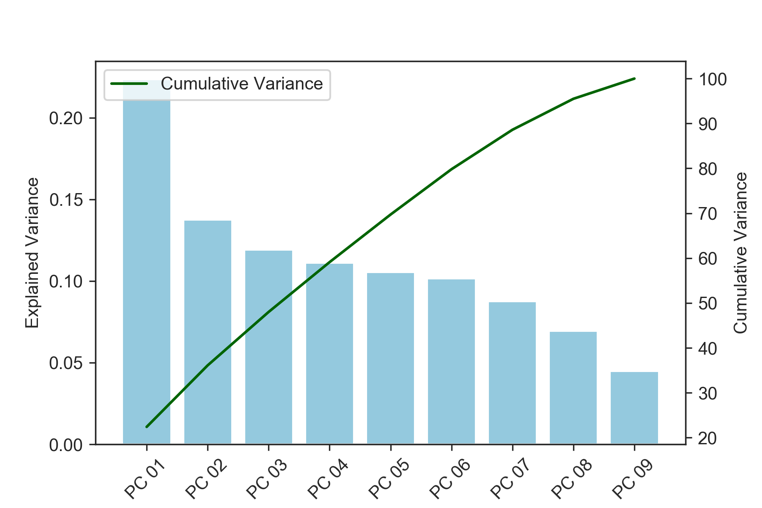
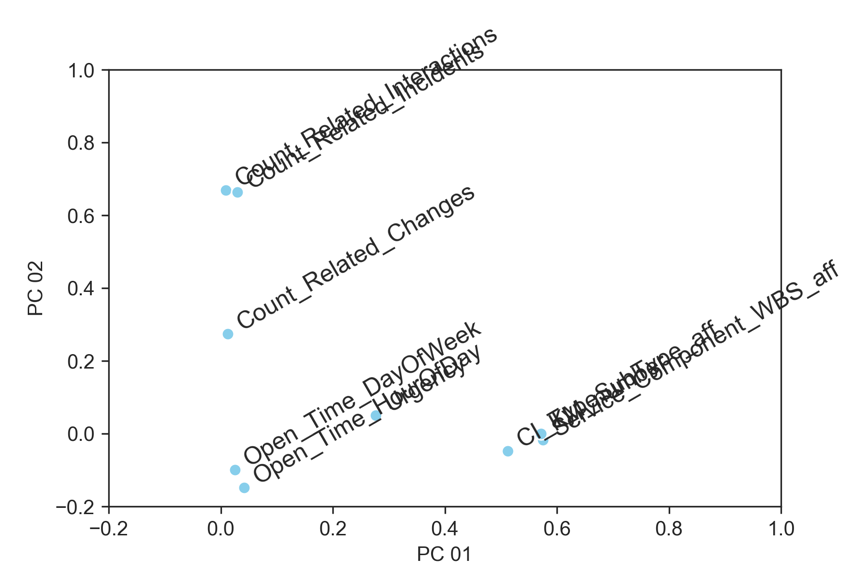


Exhibit 10. Contribution to Variance in First Two Principal Components



#### Hierarchical Cluster Analysis

Hierarchical cluster analysis contributed to the identification of variables having similar characteristics ((Tufféry, 2011, p. 236)). The first step created a dendrogram showing the sequence of partitions created by the agglomerative hierarchical clustering algorithm as shown in Exhibit 11. The second step evaluated the stability of the partitions created by the algorithm with the resulting chart shown in Exhibit 12.

Exhibit 11. Hierarchical Cluster Analysis Dendrogram

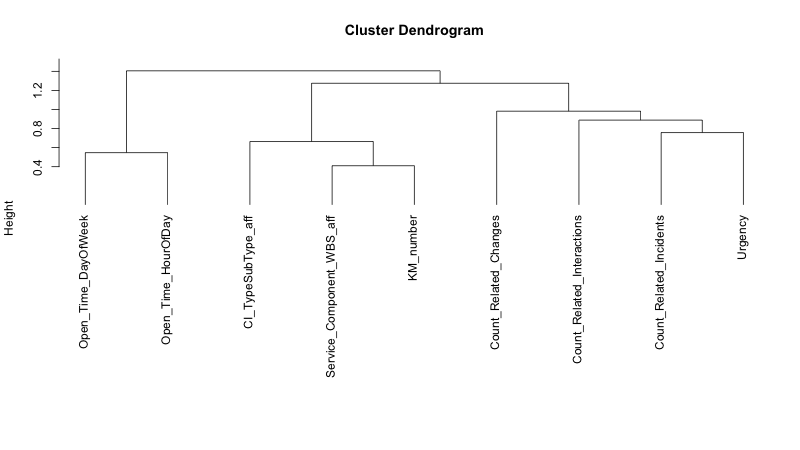
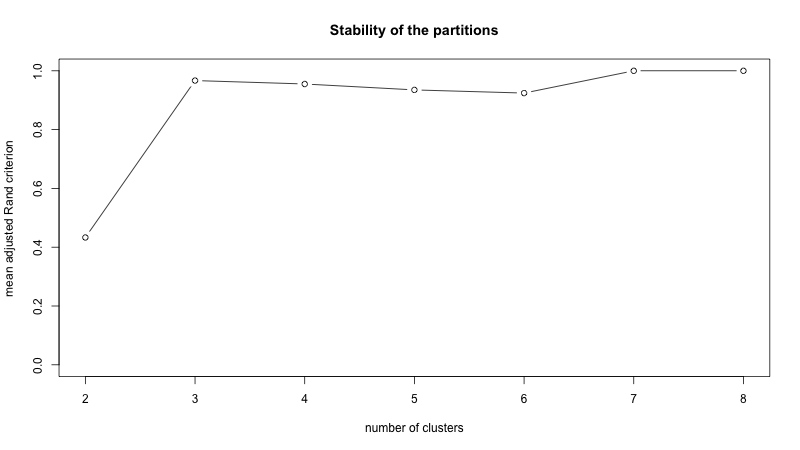


Exhibit 12. Hierarchical Cluster Analysis Stability Chart



At three clusters, the chart indicates the greatest gain in stability prior to leveling off. Exhibit 13 lists the clusters identified as a result of this analysis.

Exhibit 13. Identified Clusters

|  |  |
| --- | --- |
| Cluster | Variables |
| 1 | KM\_number  CI\_TypeSubType\_aff  Service\_Component\_WBS\_aff |
| 2 | Urgency  Count\_Related\_Interactions  Count\_Related\_Incidents  Count\_Related\_Changes |
| 3 | Open\_Time\_HourOfDay  Open\_Time\_DayOfWeek |

### Optimize Model

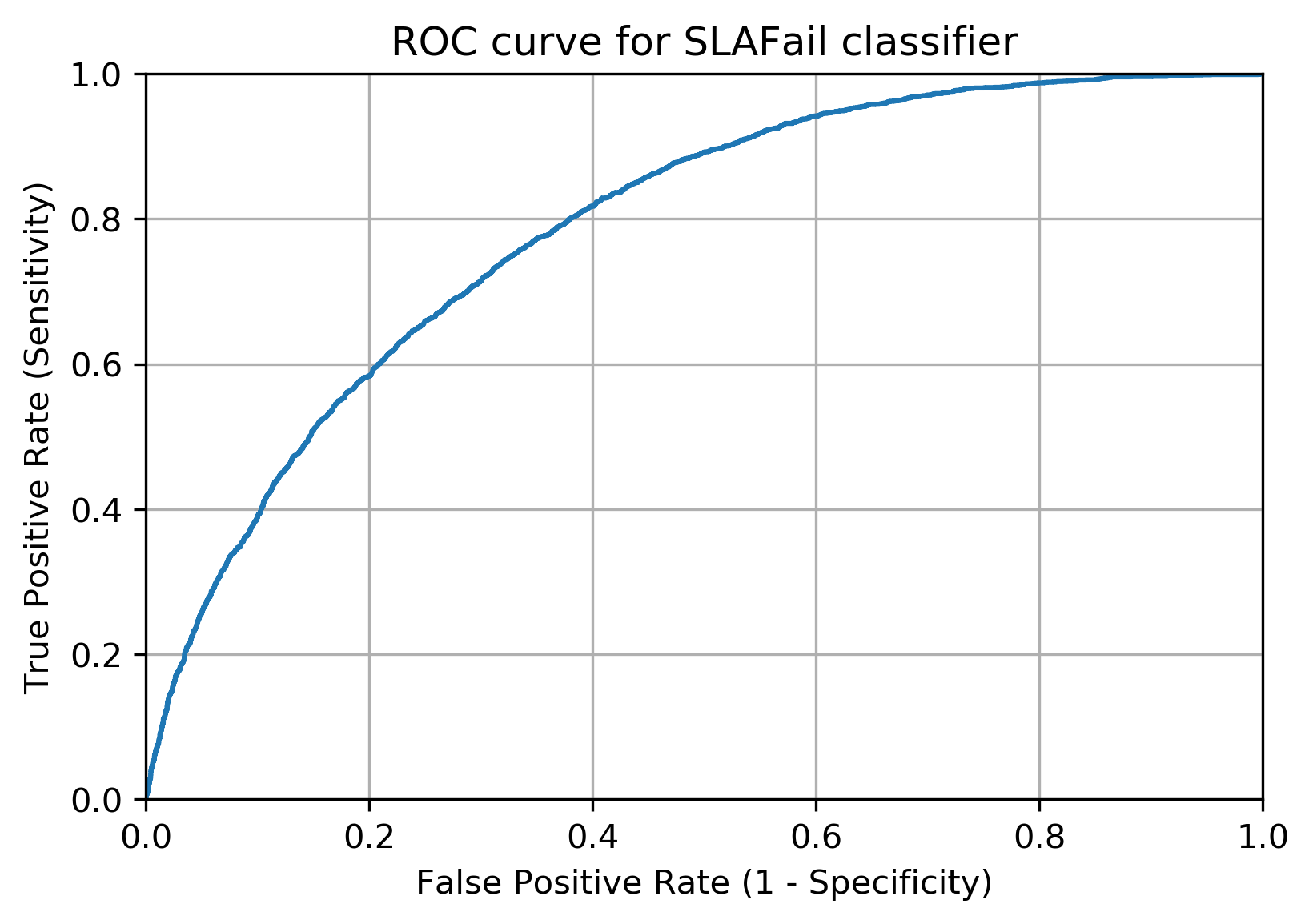
This step leverages the information and insight gained from the previous three steps to develop a logistic regression model that successfully predicts SLA compliance at the early stages of an incident’s lifecycle. With a Null Accuracy of 70%, optimization efforts aim at achieving a greater Classification Accuracy score. Given the inability to obtain more training data or to add features, this analyst’s optimization choices were limited to searching for less complicated/flexible models and looking at more complicate/flexible models (Vanderplas, n.d.). Exhibit 14 presents the results of each iteration performed during the Optimize Model step. See the appendices for an example executed notebook.

Exhibit 14. Optimization Results

|  |  |  |  |
| --- | --- | --- | --- |
| Iteration |  | Classification Accuracy | AUC |
| 1. Use automated feature selection of the k-best features based on F-score | Less complicated/ flexible model | 0.746852 | 0.780629 |
| 1. Use automated feature selection by False Positive Rate with an acceptance threshold of 0.05 for alpha/p-value | Less complicated/ flexible model | 0.746758 | 0.780663 |
| 1. Add interaction terms, use automated feature selection of the k-best features based on F-score | More complicated/ flexible model | **0.751775** | **0.786672** |
| 1. Add interaction terms, use automated feature selection by False Positive Rate with an acceptance threshold of 0.05 for alpha/p-value | More complicated/ flexible model | 0.748178 | 0.782095 |

The third iteration produced slightly higher accuracy and AUC scores than all other attempted methods. Exhibit 15 shows the ROC curve.

Exhibit 15. ROC Curve for Optimized Model



While the third iteration provided slightly better results, the introduction of interaction terms significantly increased the model’s complexity. SciKit Learn’s PolynomialFeatures created the interaction terms while a cross-validation grid search (GridSearchCV) selected the *k* best features based on f-scores. This resulted in the logistic regression model using 45 terms, versus the original nine.

Given the minimal performance improvement of the more complex model, this analyst recommends the second iteration step.

## Techniques and Tools Used

Analysis leveraged a number of techniques supported by a variety of tools. Python and the SciKit-Learn machine learning library served as the primary tool set. The SciKit-Learn library’s pipeline functionality automated repeatable steps. However, it introduced complexity in obtaining results from the sequence of steps, thus adding time and effort to reviewing and reporting on the output. This analyst encountered computing resource constraints when attempting to perform Hierarchical Cluster Analysis with Python and SciKit-Learn. As a result, R with the FactoMiner and CustOfVar libraries supplemented the tool set. As a free, open-source, well-supported, and well-documented tool, R provides advantages similar to Python (Tufféry, 2011, p. 126). The need for an alternative tool required iterative development in one tool and saving the results to a file for use in the other tool. While this could be considered a disadvantage, this analyst asserts that this modularized approach created greater flexibility in available options for downstream analysis. R’s inherent graphical capabilities produce clear and usable plots with minimal configuration while the Python plotting libraries require additional knowledge and time to configure usable plots.

Exhibit 16. Techniques and Tools for Analysis – Advantages and Disadvantages

|  |  |
| --- | --- |
| Techniques and Tools | Description |
| Weight of Evidence Coding  category\_encoders.woe.WOEEncoder | Converts categorical variables to numeric values as needed for logistic regression models using the log(odds) of the event (*SAS Training—Predictive Modeling Using Logistic Regression*, n.d.)   |  |  | | --- | --- | | Advantages | Disadvantages | | Returned highest AUC among tested encoders | Introduces an additional step | |
| ANOVA F-value  sklearn.feature\_selection.SelectKBest  sklearn.feature\_selection.f\_classif  sklearn.feature\_selection.SelectFpr | Measures the degree of linear dependency between random variables (*1.13. Feature selection—Scikit-learn 0.22.2 documentation*, n.d.)   |  |  | | --- | --- | | Advantages | Disadvantages | | Returned highest AUC among tested methods  Valid for positive and negative values (chi-squared required only positive values) | None identified | |
| Principal Component Analysis (PCA)  sklearn.decomposition.PCA | Groups variables based on their correlations (Tufféry, 2011, p. 175)   |  |  | | --- | --- | | Advantages | Disadvantages | | Industry accepted method for investigating variance and reducing dimensionality | Required encoding of categorical variables to numeric values | |
| Hierarchical Cluster Analysis  R, FactoMiner, ClustOfVar | Groups data into disjoint clusters of observations (Hastie et al., n.d., p. 521)   |  |  | | --- | --- | | Advantages | Disadvantages | | Ability to identify similar variables  Highly interpretable visualization with dendrograms | Required use of R as an alternate tool | |
| Logistic Regression  sklearn.linear\_model.LogisticRegression | Statistical analysis technique for binary dependent variables (Tufféry, 2011, pp. 170–171)   |  |  | | --- | --- | | Advantages | Disadvantages | | Appropriate classification method given binary nature of the target variable | Requires careful interpretation and presentation of results | |

# Data Summary and Implications

This section summarizes analysis implications in the context of early identification of incidents likely to cause breaches in Service Level Agreement thresholds.

## Discussion of Results

What factors predict Incident Management SLA compliance?

Given the study’s research question and hypotheses, the project developed a logistic regression model predicting the final SLA status of an incident based on data available during the early stages of an incidents lifecycle. Recall:

H0: Data contains no significant indicators about the final SLA status of an incident (the coefficients of a logistic regression model are zero, )

H1: Data contains significant indicators about the final SLA status of an incident (the coefficients of a logistic regression model are not zero, )

If any one of the logistic regression model’s coefficients significantly differs from zero, the study will accept the alternative hypothesis (H1), otherwise the study will fail to reject the null hypothesis (H0). Exhibit 17 lists the coefficients resulting from the selected optimization step of model development. With seven of the eight variables showing a significance level (P-value) less than alpha (0.05), the study accepts the alternative hypothesis that data does contain significant indicators of an incident’s final SLA status. Note that automated feature selection removed CI\_TypeSubType\_aff as insignificant prior to model generation.

Exhibit 17. Logistic Regression Model Coefficients

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Coefficient | F score | P-value |
| Service\_Component\_WBS\_aff | 0.1040 | 141.5840 | 0.0000 |
| Urgency | 0.1881 | 6104.4400 | 0.0000 |
| KM\_number | 1.1217 | 177.7560 | 0.0000 |
| Count\_Related\_Interactions | 1.1726 | 679.7630 | 0.0000 |
| Count\_Related\_Incidents | 1.2459 | 902.1090 | 0.0000 |
| Count\_Related\_Changes | 0.0388 | 62.0758 | 0.0000 |
| Open\_Time\_HourOfDay | 0.1824 | 5.6072 | 0.0179 |
| Open\_Time\_DayOfWeek | -0.0358 | 0.3519 | 0.5530 |

While the study provided statistical evidence for the acceptance of the alternative hypothesis, multiple limitations restrict the practical value of the resulting model. The best model identified by the study achieved a classification accuracy score of 78% while the null accuracy (always selecting the majority event) resided at 70%. Given implementation costs, this analyst questions the value gained with only an 8% accuracy increase. Also, given the nature of the source data set containing anonymous information, the study was unable to investigate additional factors that could provide further discrimination and benefit to the model. Categories within Knowledge Management articles represented by KM\_number, as well as data representing configuration item types (Service\_Component\_WBS\_aff), could potentially improve the model.

## Recommendations

Based on these results, this analyst recommends a course of action focused on further investigation of the specific Configuration Items and Knowledge Management articles causing incidents that exceed SLA thresholds. Analysis also showed that a subset of variables provide Incident Management professionals with direction for swift identification of incidents that may breach an SLA threshold. Principal component analysis and the hierarchical cluster analysis both grouped the following variables together:

Service\_Component\_WBS\_aff

KM\_Number

CI\_TypeSubType\_aff

With this understanding, coupled with the identification of Service\_Component\_WBS\_aff and KM\_Number as significant indicators of SLA-at-Risk, Incident Management professionals can prioritize attention on identifying specific items in the IT environment causing incidents that exceed SLA thresholds.

Recommendations for future study include seeking out additional data for use in model development and consideration of alternative classification models. As previously mentioned, additional information about categories of Knowledge Articles and Configuration Items could improve model accuracy. Further investigation of data describing the users impacted by incidents could provide additional discrimination unavailable from the source data set and the necessity of removing one of the two highly correlated variables (Impact and Urgency). Without additional data, two alternatives for further study exist. First, investigating the efficacy of other classification techniques may produce superior results, for example, decision trees, support vector machines (SVM), naive Bayesian classifiers, or *k* nearest neighbors. Second, in-depth factor analysis of individuals could identify the specific Knowledge Articles and Configuration Items with the highest contribution towards SLA breaches.

# References

*1.13. Feature selection—Scikit-learn 0.22.1 documentation*. (n.d.). Retrieved March 4, 2020, from https://scikit-learn.org/stable/modules/feature\_selection.html

*1.13. Feature selection—Scikit-learn 0.22.2 documentation*. (n.d.). Retrieved March 8, 2020, from https://scikit-learn.org/stable/modules/feature\_selection.html#univariate-feature-selection

4TU.Centre for Research Data. (2016, August). *General terms of use for 4TU.Centre for Research Data*. https://data.4tu.nl/repository/resource:terms\_of\_use/PDF

*10th International Workshop on Business Process Intelligence 2014*. (n.d.). https://www.win.tue.nl/bpi/doku.php?id=2014:challenge

Amaral, C., Fantinato, M., Reijers, H., & Peres, S. (2019). *Enhancing Completion Time Prediction Through Attribute Selection*. https://doi.org/10.1007/978-3-030-15154-6\_1

Buﬀett, S., Emond, B., & Goutte, C. (2014). *Using Sequence Classiﬁcation to Label Behavior from Sequential Event Logs*. 27.

Buhler, P., Callaghan, R. O., Aubry, S., DeJoy, D., Kuo, E., Shoup, N., Khosla, I., Ginsburg, M., Hartman, N., & McBride, N. S. (n.d.). *Service Desk and Incident Impact Patterns Following ITIL Change Implementation*. 36.

Hanna, A. (2011). *ITIL(r) glossary and abbreviations*. AXELOS Limited. https://www.axelos.com/corporate/media/files/glossaries/itil\_2011\_glossary\_gb-v1-0.pdf

Hastie, T., Tibshirani, R., & Friedman, J. (n.d.). *The Elements of Statistical Learning*. 764.

Higgins, S. (2016, April 26). *How predictive analytics have turned Incident Management on its head -*. http://www.theitsmreview.com/2016/04/predictive-analytics-turned-incident-management-head/

*Measuring Response and Resolution Times in Remedy | University IT*. (n.d.). Stanford University IT. Retrieved February 19, 2020, from https://uit.stanford.edu/service/helpdesk/support/sla

Nisbet, R., Elder, J. F., & Miner, G. (2009). *Handbook of statistical analysis and data mining applications*. Academic Press/Elsevier.

*NumPy—NumPy*. (n.d.). Retrieved February 24, 2020, from https://numpy.org/

*Pandas\_profiling API documentation*. (n.d.). Retrieved February 24, 2020, from https://pandas-profiling.github.io/pandas-profiling/docs/

*pandas—Python Data Analysis Library*. (n.d.). Retrieved February 24, 2020, from https://pandas.pydata.org/

Perkel, J. M. (2018). Why Jupyter is data scientists’ computational notebook of choice. *Nature*, *563*(7729), 145–146. https://doi.org/10.1038/d41586-018-07196-1

*Project Jupyter*. (n.d.). Retrieved February 24, 2020, from https://www.jupyter.org

*Quick reference BPI Challenge 2014*. (n.d.). Retrieved February 17, 2020, from https://www.win.tue.nl/bpi/lib/exe/fetch.php?media=2014:quick\_reference\_bpi\_challenge\_2014.pdf

Sarnovsky, M., & Surma, J. (2018). PREDICTIVE MODELS FOR SUPPORT OF INCIDENT MANAGEMENT PROCESS IN IT SERVICE MANAGEMENT. *Acta Electrotechnica et Informatica*, *18*(1), 57–62. https://doi.org/10.15546/aeei-2018-0009

*SAS Training—Predictive Modeling Using Logistic Regression*. (n.d.). SAS Virtual Learning Environment. Retrieved February 19, 2020, from https://support.sas.com/edu/schedules.html?id=13086&ctry=US

Tufféry, S. (2011). *Data Mining and Statistics for Decision Making: Tufféry/Data Mining and Statistics for Decision Making*. John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470979174

Van Der Aalst, W. M. P. (2018). *Process Mining: Data Science in Action*. Springer Publishing Company, Incorporated.

Van Dongen, B. F. (Boudewijn). (2014). *BPI Challenge 2014: Incident details*. Rabobank Nederland. https://doi.org/10.4121/UUID:3CFA2260-F5C5-44BE-AFE1-B70D35288D6D

Vanderplas, J. (n.d.). *Hyperparameters and Model Validation*. Python Data Science Handbook. Retrieved March 7, 2020, from https://jakevdp.github.io/PythonDataScienceHandbook/05.03-hyperparameters-and-model-validation.html

*Welcome to Python.org*. (n.d.). Python.Org. Retrieved February 24, 2020, from https://www.python.org/

# Appendix Describing the Source Data Set

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Type | | Description |
| Alert\_Status | | Categorical, CONSTANT value "closed" | Status for monitoring service levels |
| Category | | Categorical, four levels | Grouping of types of incidents |
| CI\_Name\_aff | | Categorical, HIGH CARDINALITY | Configuration Item (CI) where a disruption is noticed (Affected CI) |
| CI\_Name\_CBy | | Categorical, HIGH CARDINALITY | Configuration Item (CI) which caused the disruption (Caused-By CI) |
| CI\_Subtype\_aff | | Categorical, HIGH CARDINALITY | Second-level category of Affected CI |
| CI\_Subtype\_CBy | | Categorical, HIGH CARDINALITY | Second-level category of Caused-By CI |
| CI\_Type\_aff | | Categorical, 13 levels |  |
| CI\_Type\_CBy | | Categorical, 14 levels | Top-level category of Caused-By CI |
| Close\_Time | | Date Minimum: 2013-10-01 06:45:43 Maximum: 2014-03-31 22:47:32 | Date and time of incident closure |
| Closure\_Code | | Categorical, 15 levels 1.0% missing | Classification of disruption type |
| Count\_Reassignments | | Continuous, ZEROS (58.9%) | Number of times responsibility for the incident changed |
| Count\_Related\_Changes | | Continuous, MISSING (98.8%) | Number of Change Management records associated with the incident |
| Count\_Related\_Incidents | | Continuous, MISSING (97.4%) | Number of similar or related incidents (child records) |
| Count\_Related\_Interactions | | Continuous, SKEWED, MISSING (0.03%) | Number of updates or changes to the incident record |
| Handle\_Time\_Hours | | Continuous | Time required to actively resolve the incident |
| Impact | | Categorical, five levels | Impact of the disruption to the customer |
| Incident\_ID | | Categorical, UNIQUE, HIGH CARDINALITY | Unique identifier for each incident |
| KM\_number | | Categorical, HIGH CARDINALITY | Knowledge management article containing default attributes and questions for service desk analyst use |
| Open\_Time | | Date  Minimum: 2012-02-05 13:32:57 Maximum: 2014-03-31 17:24:49 | Date and time of incident creation |
| Priority | | Categorical, five levels | Priority derived from the Impact and Urgency values |
| Related\_Change | | Categorical, MISSING (98.8%), HIGH CARDINALITY | Change record identifier (if only one change is related to the incident) |
| Related\_Interaction | | Categorical, HIGH CARDINALITY | Interaction record identifier (if only one interact is related to the incident) |
| Reopen\_Time | | Date Minimum: 2013-04-10 09:15:55 Maximum: 2014-03-31 16:21:15 MISSING (95.9%), | If the incident is re-opened shortly after closure based on customer feedback, the date and time the incident was re-opened |
| Resolved\_Time | | Date Minimum: 2013-10-01 06:45:36 Maximum: 2014-03-31 22:47:29 3.8% Missing | Date and time of incident resolution |
| Service\_Component\_WBS\_aff | | Categorical, HIGH CARDINALITY | Service component identifier for the Affected CI |
| ServiceComp\_WBS\_CBy | | Categorical, HIGH CARDINALITY | Service component identifier for the Caused-By CI |
| Status | | Categorical, two levels | Status of the incident |
| Urgency | | Categorical, five levels | Indicates incident resolution urgency |

# Appendix of Executed Python Notebooks and Scripts

The following pages show the content and results generated by a few of the Jupyter Notebooks used throughout this project.

|  |  |  |  |
| --- | --- | --- | --- |
| Report Section | Notebook Title | File Name | Page |
| 3.1.1 | 01. Exploratory Data Analysis and Preliminary Cleaning | 01. EDA\_Detail\_Incident.ipynb |  |
| 3.1.2 | 02. Cleaning the Source Data Set | 02. Cleaning\_Detail\_Incident.ipynb |  |
| 3.1.3 | 03. Creating the Target Variable (SLAFail) | 03. Create\_SLAFail.ipynb |  |
| 3.1.4 | 04. Final Data Preparation | 04. Final Data Prep.ipynb |  |
| 4.2.1 | 05.01.c Bare Bones Analysis Using a Weight of Evidence Encoder | 05.01.c Bare Bones Analysis WOE Encoder.ipynb |  |
| 4.2.2 | 05.02 Feature Selection KBest with ANOVA F-value Score Function | 05.02.a Feature Selection KBest with f\_classif.ipynb |  |
| 4.2.4 | 06.01.b Optimize the Logistic Regression Model | 06.01.b Optimize 2 Select FPR with f\_classif and pval.ipynb |  |

## Notebook: 01. Exploratory Data Analysis and Preliminary Cleaning

Output from executed notebook begins on the next page.

INSERT 01. Exploratory Data Analysis and Preliminary Cleaning

## Notebook: 02. Cleaning the Source Data Set

Output from executed notebook begins on the next page.

INSERT 02. Cleaning the Source Data Set

## Notebook: 03. Creating the Target Variable (SLAFail)

Output from executed notebook begins on the next page.

INSERT Notebook: 03. Creating the Target Variable (SLAFail)

## Notebook: 04. Final Data Preparation

Output from executed notebook begins on the next page.

INSERT 04. Final Data Preparation

## Notebook: 05.01.c Bare Bones Analysis Using a Weight of Evidence Encoder

Output from executed notebook begins on the next page.

Insert 05.01.c Bare Bones Analysis WOE Encoder.ipynb

## Notebook: 05.02 Feature Selection KBest with ANOVA F-value Score Function

Output from executed notebook begins on the next page.

Insert 05 kbest select

## Notebook: 06.01.b Optimize the Logistic Regression Model

Output from executed notebook begins on the next page.

Insert 06.01.b Optimize Select FPR.ipynb