Capstone project

Data Science Nanodegree

Bart Leplae 03-Feb-2023

# Section 1: Project Definition

## High-level overview

In the domain of application support, users raise incident tickets to seek resolution for the issues they experience. These tickets are assigned to the support teams that are servicing the respective applications. When these incident tickets are resolved, users are sent a survey to obtain their feedback with the intent to identify improvement opportunities.

On average, 11% of the survey requests are responded to. The organization wants to understand the causes of customer dissatisfaction and benchmark support teams and applications against each other.

## Problem Statement

1. Can 'dissatisfied' responses be correlated with specific ticket attributes?

2. Can the ratio of 'dissatisfied' responses for a subset of tickets (such as application) be predicted (modelled) based on the ticket attributes?

3. What is the predicted satisfaction ratio for tickets that don't have a survey response?

## Description of Input Data

116000 incident tickets have been registered over a period of 1 year. These incident tickets are limited to those reported by users and exclude those that are auto generated through monitoring.

* 11% of the incidents have a survey response (of which 10% have a dissatisfied score)
* 89% incidents don’t have a survey response

Satisfaction rating values are provided through the attribute: survey\_response\_value

* survey\_response\_value = 1 (dissatisfied) - … - 5 (very satisfied)
* survey\_response\_value 1 and 2 are regarded as ‘dissatisfied’.
* survey\_response\_value 0 corresponds with incident tickets without a survey response.

Attributes (detailed in the EDA section): close\_code, breached\_reason\_code, contact\_type, self\_service, incident\_reopened\_flag reopened, sla\_result, sla\_priority, am\_ttr, incident\_has\_ka\_related\_flag has\_knowledge\_article, reassignment\_count, appl\_tier, caller\_vip, caller\_employee\_type, ci\_name, assignment\_group\_company, assignment\_group\_name, kcs\_solution

,

## Strategy for solving the problem

1. Perform a chi2 test to rank the available incident attributes as per the strength of their correlation with incident dissatisfaction responses.
2. Manually evaluate the available values for each of the categorical attributes and apply transformations as needed.
3. Manually evaluate the attributes after transformation to determine which ones to use for modelling. Exclude company, support group and application for modelling but rather use these for subsequent evaluation once a prediction model has been constructed.
4. Build a dissatisfaction prediction model based on the identified attributes.
5. Identify the primary attributes that have been retained by the model.
6. Evaluate the impact of the primary attributes on subgroups of interest: secondary attributes + the attributes that had been selected for subsequent evaluation (such as applications).
7. Apply the model on the incident tickets without user satisfaction feedback.
8. Compare the predicted satisfaction and the impact of primary attributes on the tickets with and without user satisfaction.

## Discussion of the expected solution

We expect the solution to provide:

1. A list of primary attributes that drive customer dissatisfaction
2. The extent to which the customer satisfaction percentage can be improved by influencing each of these primary attributes

## Metrics with justification

We are looking for the percentual impact of each of the primary attributes on customer satisfaction for:

1. Secondary attributes (those that have a correlation with customer satisfaction but had not been retained by the model)
2. Companies, Support Groups, applications

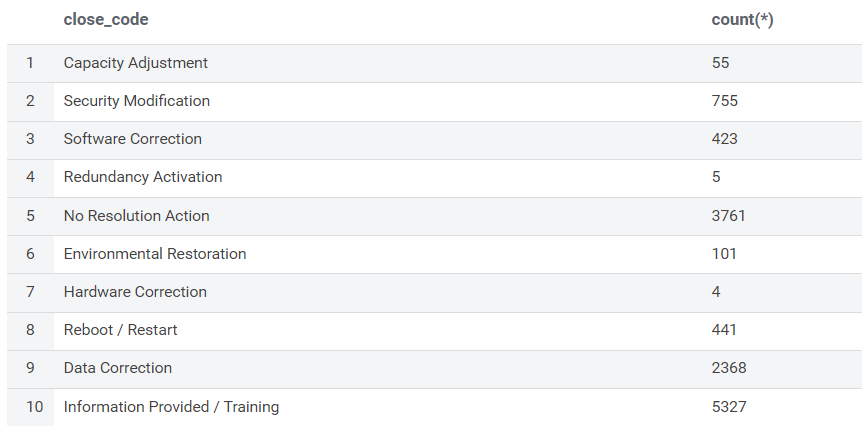
These metrics should provide insight as why customer satisfaction is impacted so that this can trigger further service improvements.

# Section 2: Analysis

## Exploratory Data Analysis (EDA)

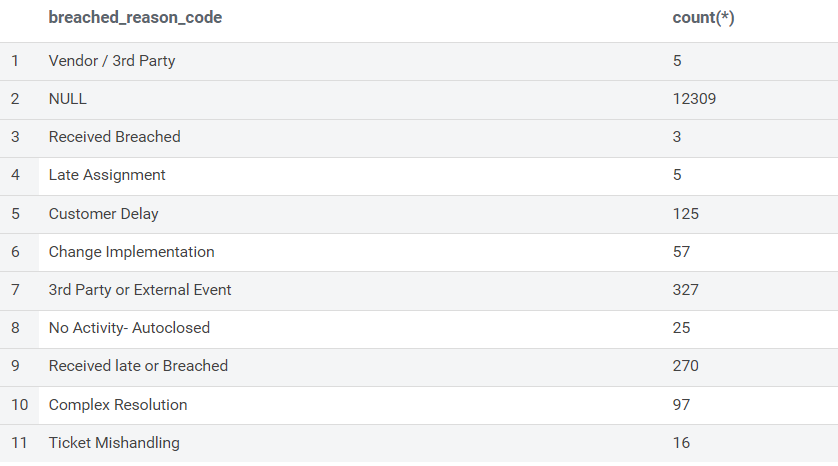
Available attributes (counts are for tickets with a survey response):





**close\_code**: has 10 values, some of which are rarely used

**breached\_reason\_code**: only available when service level has been breached so not well suited for modelling



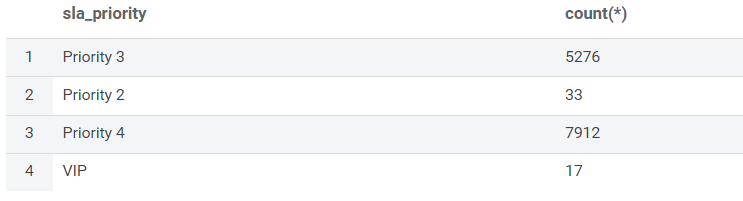


**self-service**: is 1 when contact\_type = self-service

**reopened**: 1 when user has reopened the incident ticket

**sla\_result**: only Achieved and Breached are of practical use



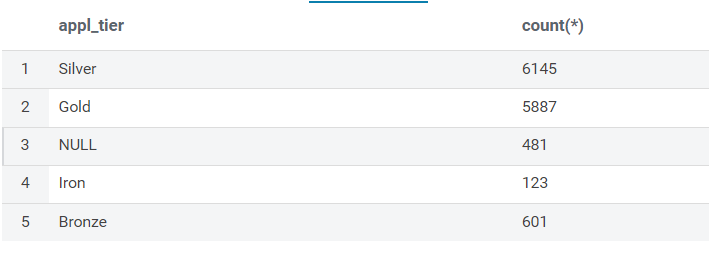


Priority 2 and VIP are infrequent, so not suited for modelling

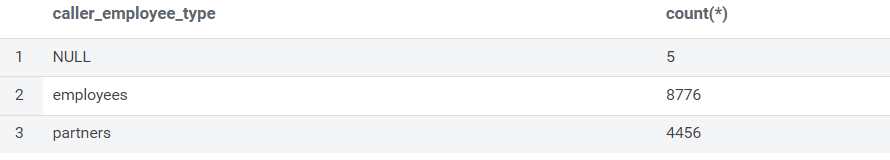
**am\_ttr**: is expressed in seconds, so not very intuitive

**has\_knowledge\_article**: 1 if knowledge article (prescribed solution) has been attached to the incident

**reassignment\_count:** number of times the incident ticket has been reassigned between support teams or staff members. Reassignment counts > 4 seldom occur.



**Caller\_vip**: 1 if caller has been identified as a VIP person, seldom used so not well suited for modelling.



NULL should be disregarded.

**Survey\_response\_value**: between 1 and 5 (very satisfied)

**CI\_name**: synonym for application

**Assignment\_group\_company**: company that is contracted to provide the service

**Assignment\_group**: group that resolved the incident

**KCS\_solution**: identifier of the knowledge article

# Section 3: Methodology

## Data Preprocessing

Following data transformations have been applied:

1. Convert Float to discrete number of integer values to enable modelling:

* days\_to\_resolve: am\_ttr (seconds) \*24\*3600 rounded to whole number of days
* days\_to\_resolve above 15 days truncated to 15 days (given the low volumes)

1. Convert categorical to binary to enable modelling:

* user\_dissatisfied = 1 if survey\_response\_value < 3 else 0
* sla\_breached = 1 if sla\_result = ‘sla\_breached’ else 0
* caller\_is\_employee = 1 if caller\_employee\_type = “caller\_is\_employee” else 0
* priority\_is\_4 = 1 if sla\_priority = "Priority 4" else 0 (given low occurrences of VIP and Priority2)

1. replace with anonymous values to respect privacy:

* assignment\_group\_company
* assignment\_group\_name
* ci\_name

1. reassignment\_count above 4: truncate to 4 given low occurrences and high variability



1. Close\_code: Capacity Adjustment', 'Hardware Correction', 'Redundancy Activation']: combine with ‘Environmental Restoration’ given the low number of occurrences:



close\_code: create dummy variables for each of the remaining values

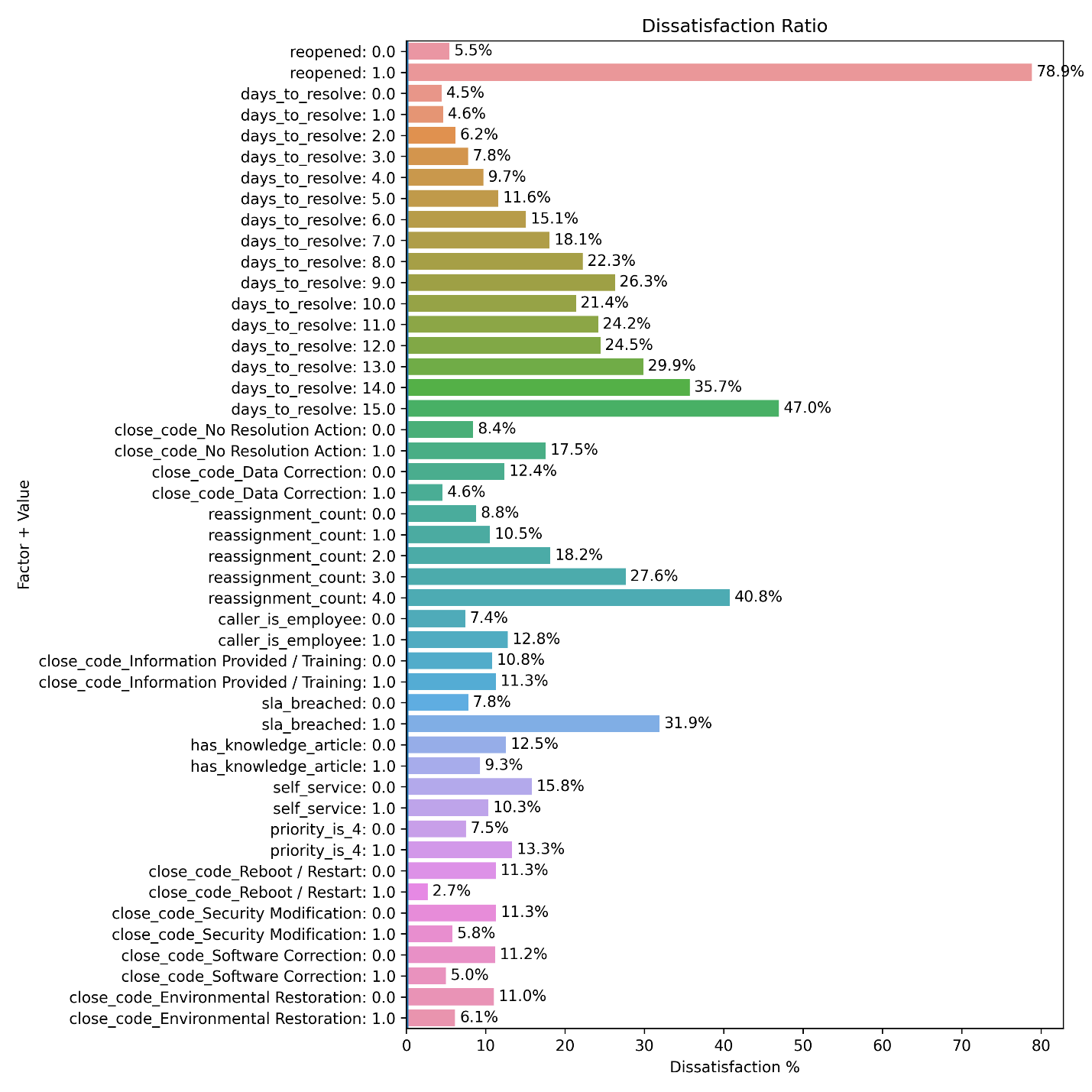
1. Eliminate contact\_type as a separate factor given the similarity with ‘self\_service’:



1. Eliminate breached\_reason\_code given the similarity with ‘sla\_breached’ and limited differentiation of the dissatisfaction ratio:



Data preprocessing results in the following factor-value combinations to be used as input to the prediction model:



Observations:

* The strongest dissatisfaction is expressed by users who had to reopen their incident tickets or for whom the ticket was not resolved (close\_code\_No Resolution action).
* The dissatisfaction ratio increases with the number of days needed to resolve the incidents.
* The dissatisfaction ratio increases with number of ticket reassignments
* Employees are less satisfied as compared to external staff

## Implementation

The model of choice is [**sklearn.tree.DecisionTreeClassifier**](https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html)

* X = the factors resulting from data pre-processing (see above)
* y = user dissatisfied (1 or 0)

By default, the model will seek to predict the incident tickets for which a user will provide ‘dissatisfied’ survey response. Given the strength of the correlation, this results in ‘reopened’ to become the only factor of choice by the model.

Since the objective is not to predict the user feedback for individual tickets but to predict the ratio of dissatisfied responses for a group of tickets (such as an application), we use the [**predict\_proba**](https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html#sklearn.tree.DecisionTreeClassifier.predict_proba)(X) method instead of [**predict**](https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html#sklearn.tree.DecisionTreeClassifier.predict)(X). This way, we predict the class probabilities of the input samples X. The predicted class probability is the fraction of samples of the same class in a leaf.

## Refinement

Without further measures, the DecisionTreeClassifier largely overfits by creating a very deep tree. We therefore applied cross validation by means of [**sklearn.model\_selection.GridSearchCV**](https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html)

This cross validation method was provided a custom scoring function to ensure that the model predicts well over the entire range of possible dissatisfaction ratios:

* split the tickets in 10 groups based on the predicted dissatisfaction probability
* for each of the 10 groups: calculate the difference between the number of predicted and number of actual dissatisfaction responses
* calculate the average difference across all these 10 groups
* return 1/(the calculated average) to provide a higher score when the differences are lower

Evaluation:

* Simple trees (with limited depth) result in less than 10 groups (number of different percentages that the model can return). These ‘simplistic’ trees return a lower score.
* Complex (overfitting) trees will result in a lower score because of the evaluation against ‘unseen’ ticket data.

## Hyperparameter Tuning

The following hyperparameters have been utilized:

    params = {

        'max\_depth': [5, 6, 7, 8, 9, 10],

        'min\_samples\_leaf': [50, 60, 70, 80, 90, 100, 110, 120, 130],

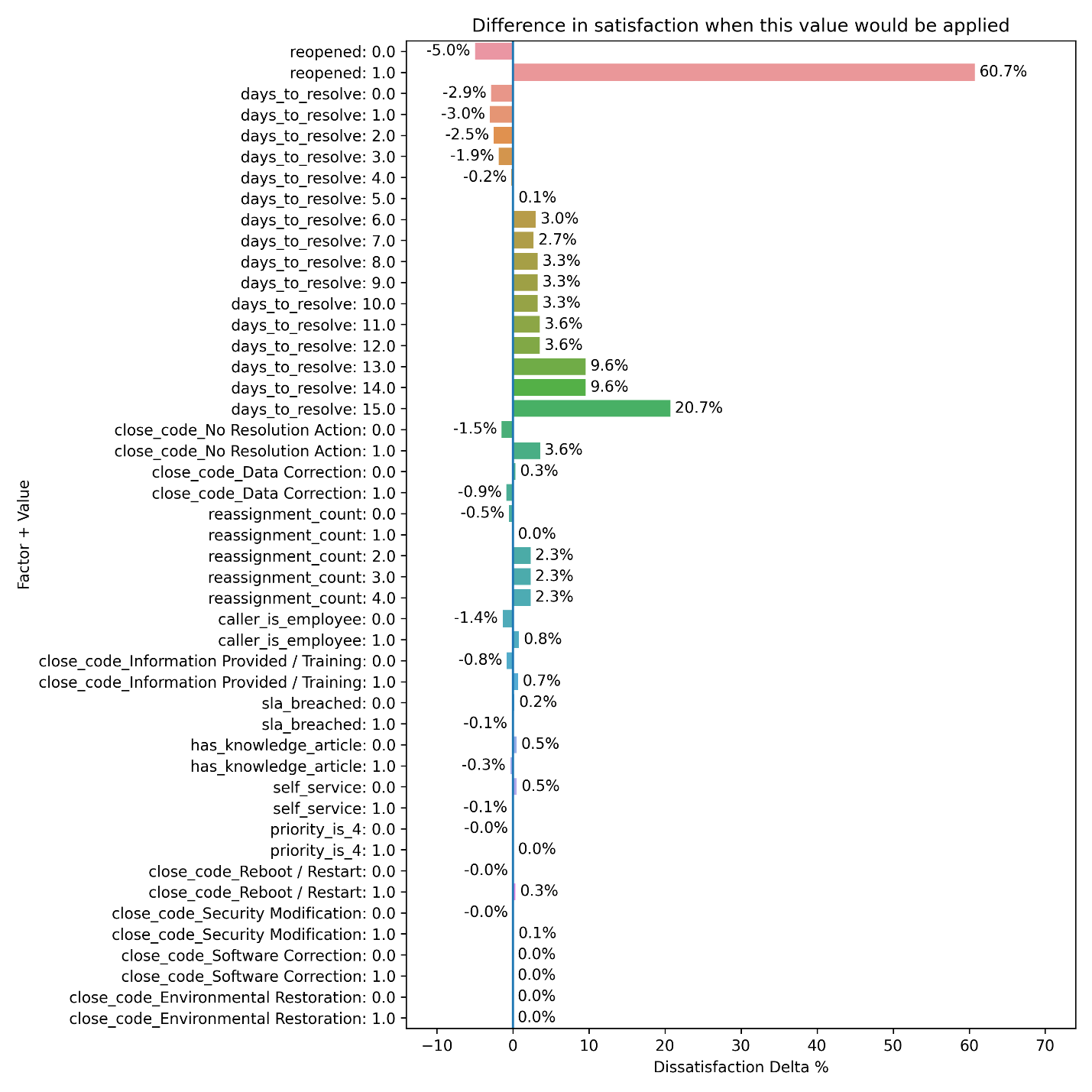
        'criterion': ["gini","entropy"]

    }

# Section 4: Results

## Modeling

The following graph depicts how the overall dissatisfaction ratio would increase or decrease if the given attribute-value combination would be enforced through the prediction model:



The decision tree model reveals 3 primary attributes:

* reopened by user
* resolution timeline
* close code ‘no resolution’.

The other attributes should be regarded as secondary attributes. Some of these will be covered the section ‘comparison tables’ further in this document.

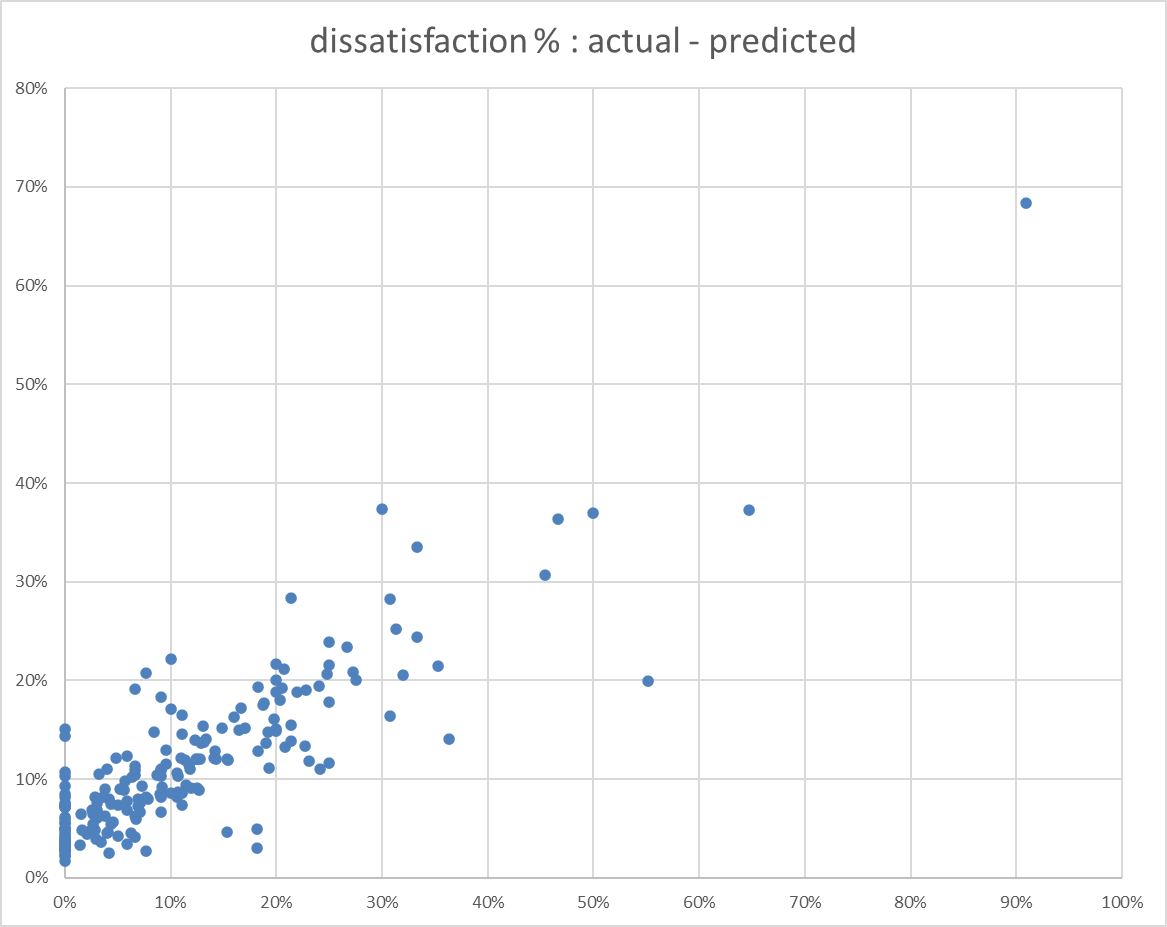
## Results

I started initially with a logistic regression model, but this model suffered from the fact that the ‘reopened’ factor is very dominant. The other factors (‘days to resolve’ and ‘no resolution action’) have a different impact dependent on whether a ticket has been reopened.

Results are presented in the form of a correlation matrix, bar graphs and Excel sheets.

### Correlation:

Actual versus predicted satisfaction ratio for applications with more than 10 tickets:

****

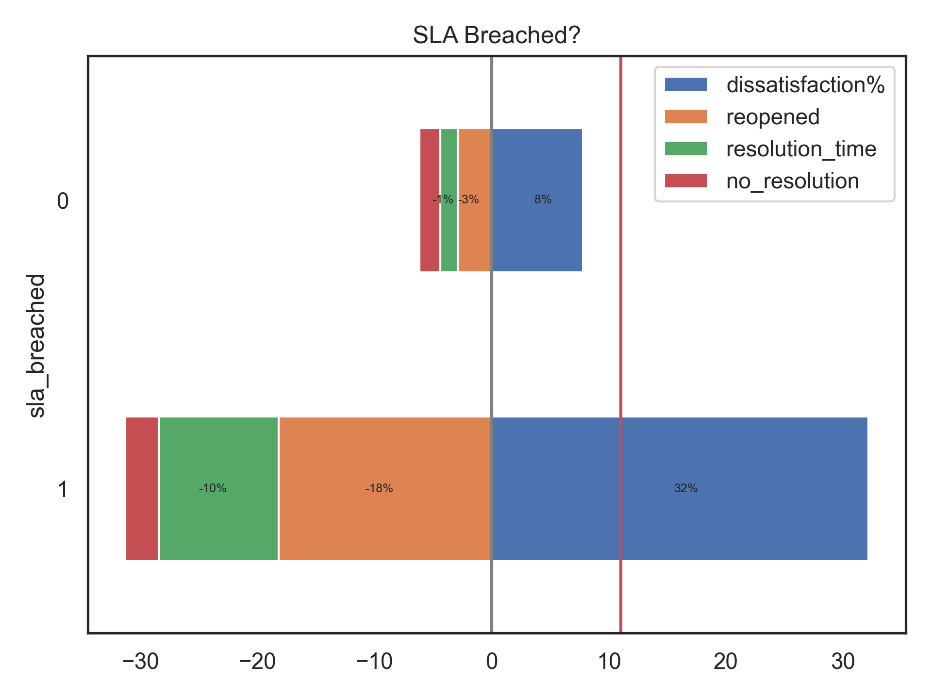




So, the model can identify 87% of the applications that have a dissatisfaction ratio of above 10%.

### bar graphs:

The most relevant bar graphs are covered in section ‘comparison tables’ and have the following format:



The dissatisfaction% is a positive value depicting the actual user responses in each category.

The other percentages are negative values indicating the extent to which the dissatisfaction% would decrease if we would eliminate the factors ‘reopening’, ‘resolution time’ and ‘no\_resolution’.

The red vertical depicts the average dissatisfaction ratio.

### Excel sheets



These Excel sheets are ordered in the order of statistical relevance depicting the applications that need most attention. The negative %-tages indicate the potential opportunity by addressing the primary factor.

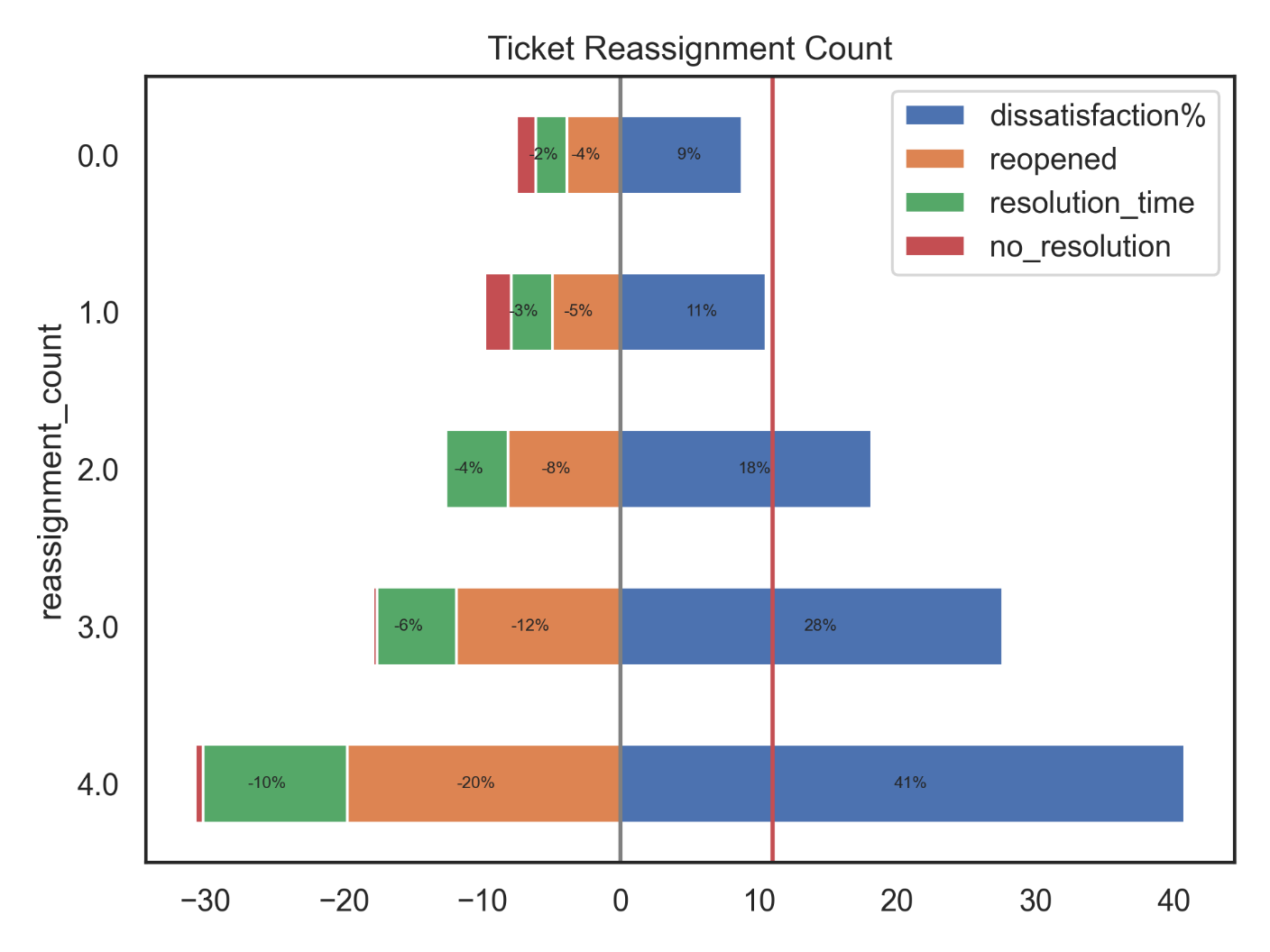
## Comparison TableS

### Ticket reassignment

Tickets are reassigned when passed between different team members or groups.

The dissatisfaction ratio increases with the number of ticket reassignments.

The negative values depict the effect of the primary factors.

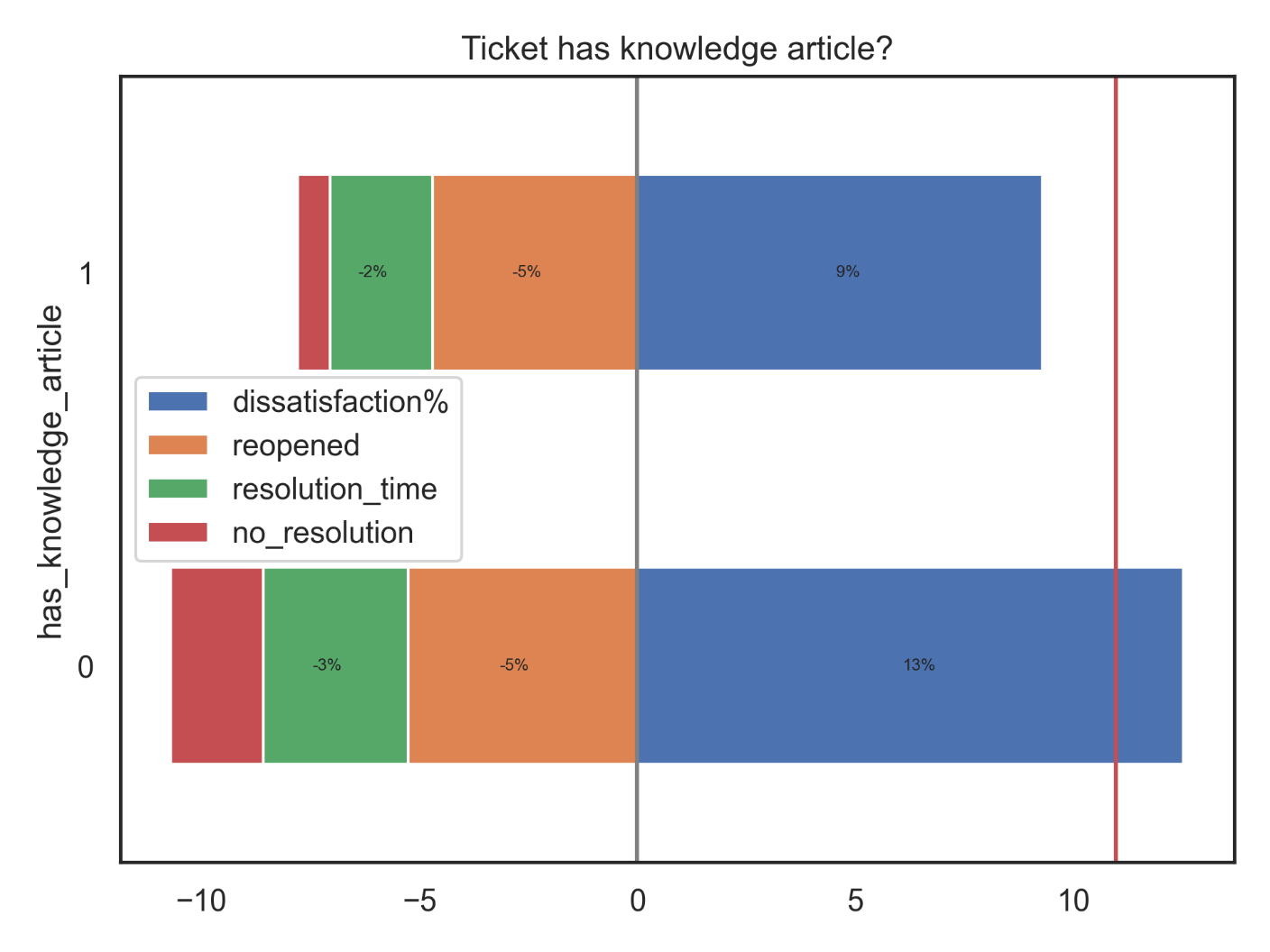


Observations:

* Higher reassignment counts result into longer resolution timelines
* Reopened tickets lead to more reassignments

### Knowledge articles

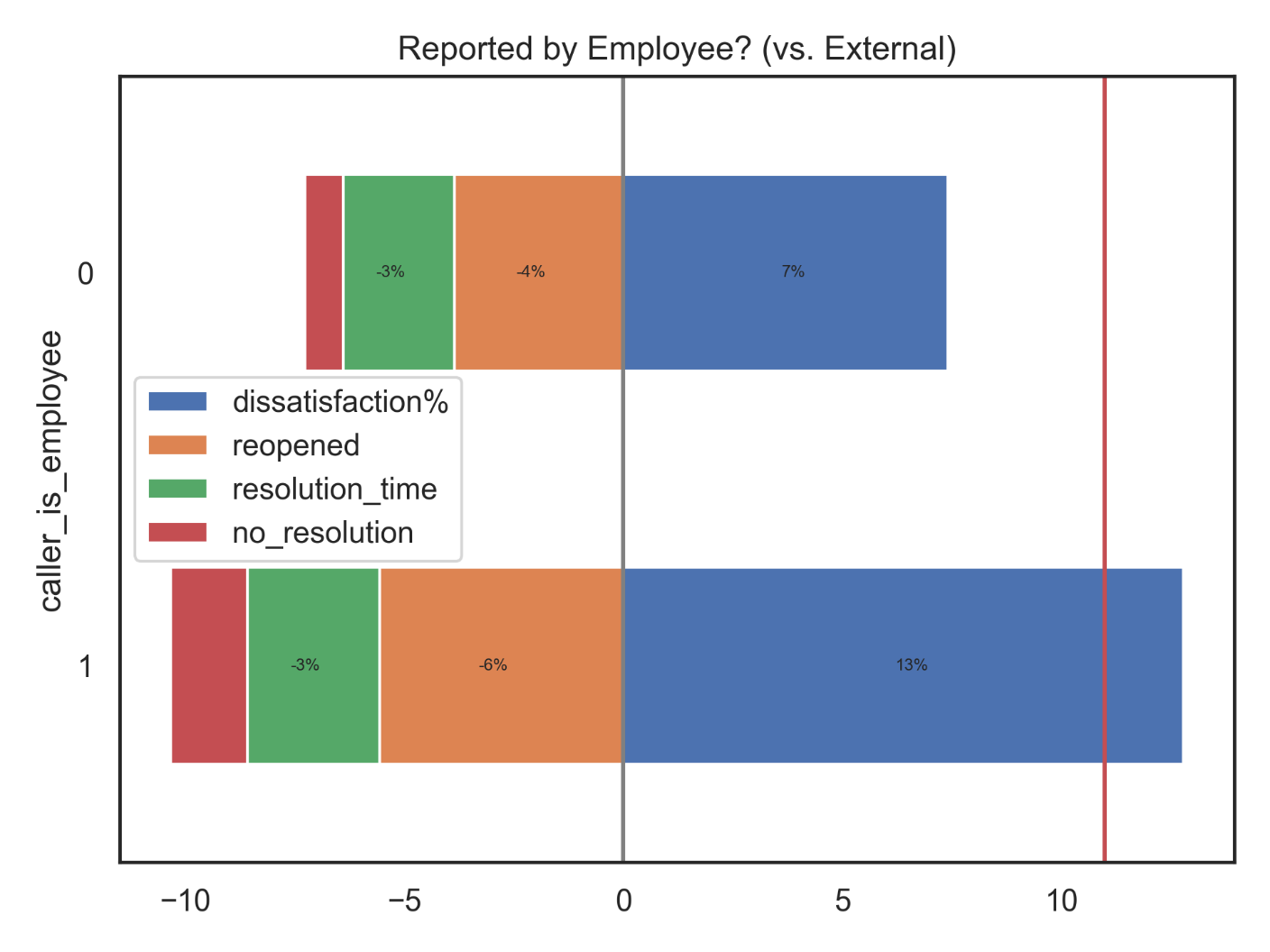
Knowledge articles describe how recurrent tickets can be resolved. These articles are attached to the tickets that made use of them.



Observations:

* The availability of a knowledge article reduces the resolution timeline
* Knowledge articles lead to fewer tickets with ‘no resolution’
* Slight effect: Knowledge articles lead to fewer reopened tickets

### Employee versus external



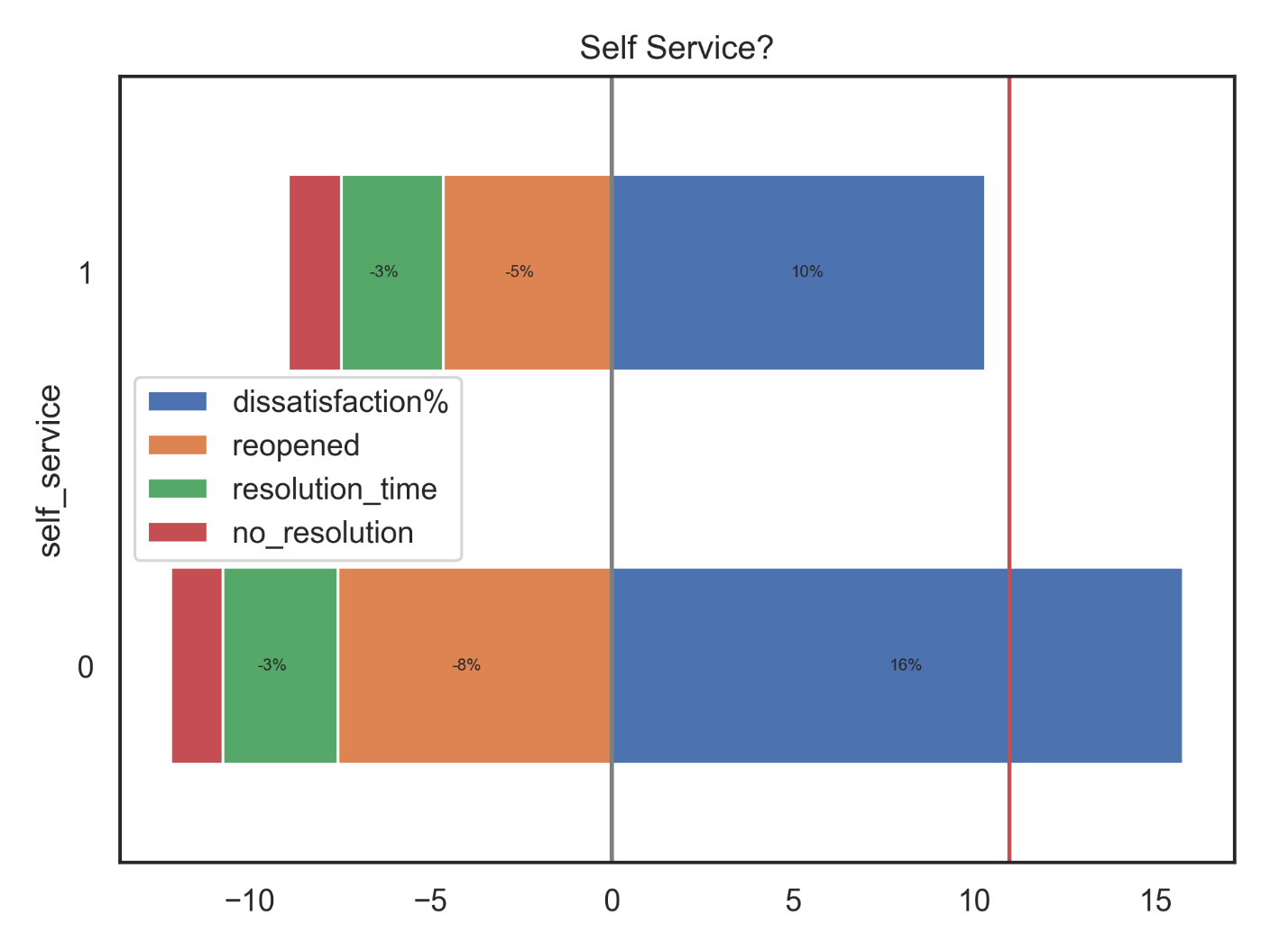
Observations:

* Employees are keener to express their dissatisfaction in their satisfaction survey response
* Employees reopen their incidents more frequently

### Self-service tickets

Self-service tickets are entered by the users themselves.

Other tickets are entered by service desk agents (upon a call or chat).



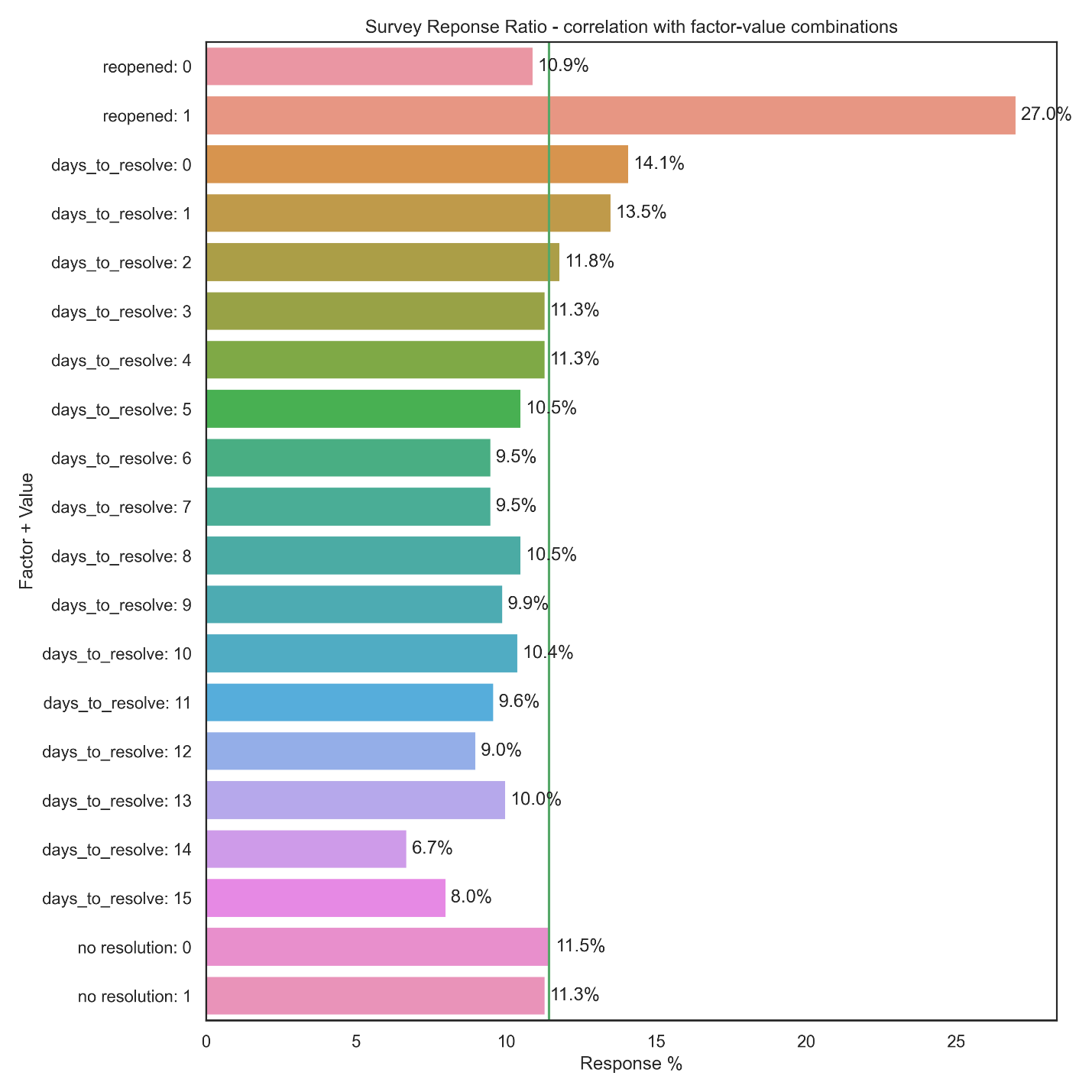
Observations:

* Tickets entered by the service desk are re-opened more frequently (leading to dissatisfaction)
* Tickets entered by the service desk result in a reduced satisfaction

Remaining questions:

* Are the tickets entered by the service desk of higher complexity?
* Are the tickets entered by the service desk missing essential information because of which their resolutions are suboptimal?

### Tickets without survey responses

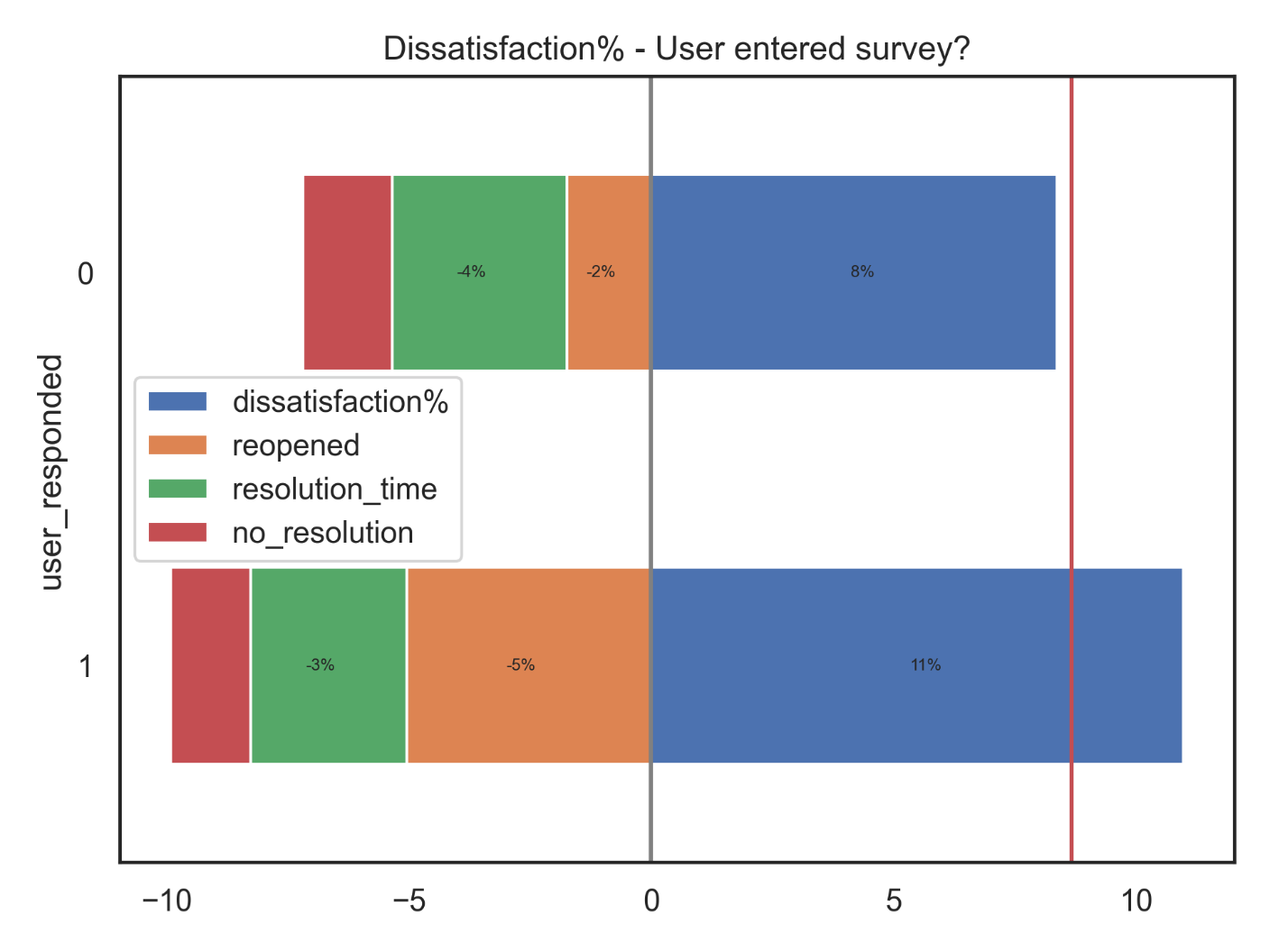


Observations:

* Users who reopened their incident tickets more often respond to the survey to express their dissatisfaction
* Users more often respond to the survey to express their satisfaction when a resolution is provided with 2 days.

### Dissatisfaction% for tickets without survey response

When the model is applied to calculate the expected dissatisfaction% for tickets that have and don’t have a survey response:



Observations:

* Tickets with a survey response have a higher dissatisfaction ratio because users who reopened a ticket tend to enter the survey response to express their dissatisfaction
* Tickets with short resolution times have more survey responses. The resolution time is not a differentiator.
* The overall dissatisfaction ratio (8.6%) is less than what could be expected from the dissatisfaction ratio derived from the survey responses (11.0%).

# Section 5: Conclusion

## Conclusion

Summarize the end-to-end problem solution and discuss one or two particular aspects that you find interesting or difficult to implement.

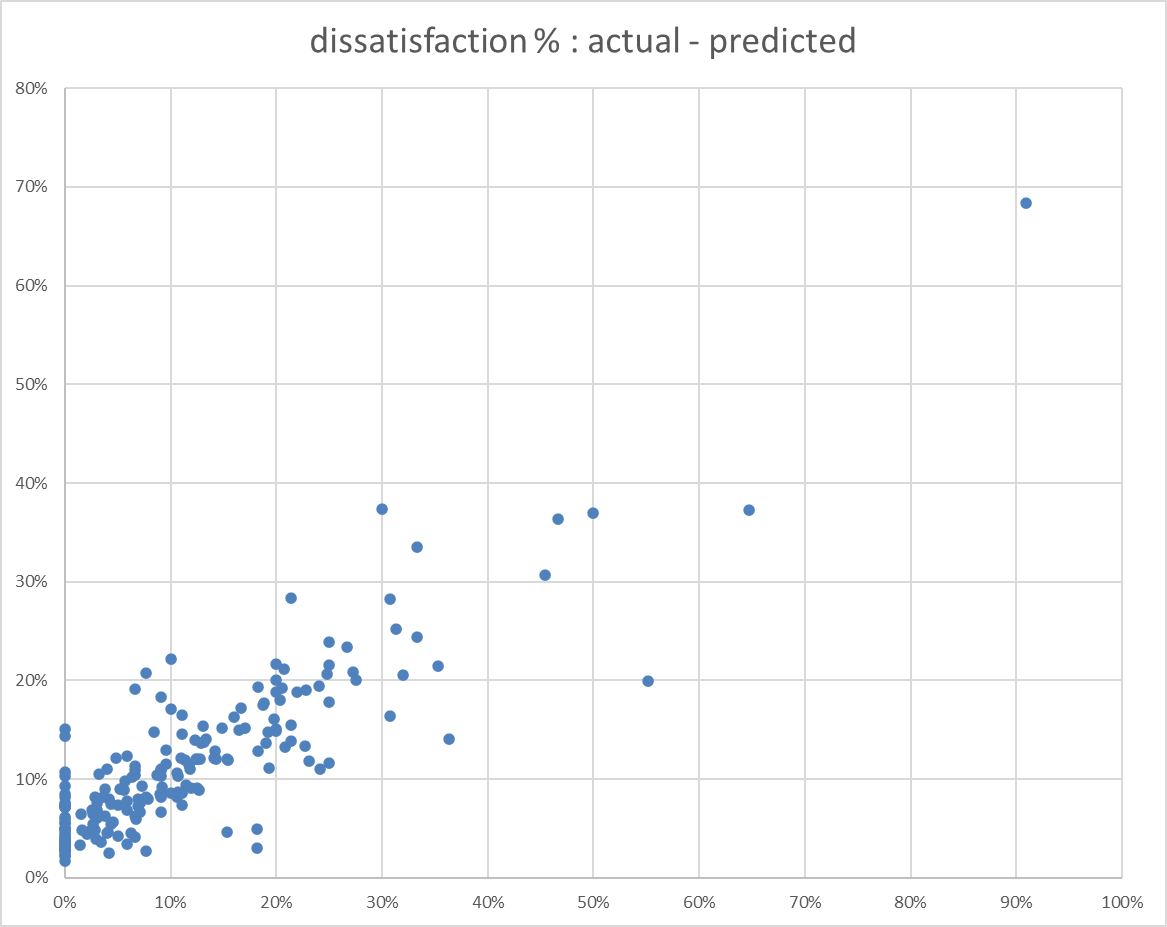
1. **Can 'dissatisfied' responses be correlated with specific ticket attributes?**

‘Dissatisfied’ survey responses are the consequence of not providing an appropriate resolution (user reopened the ticket or ticket closed with ‘no resolution action’) or when the resolution takes more than a week.

While other attributes are correlated with user dissatisfaction, it looks like these can be mostly explained through the above primary factors.

1. **Can the ratio of 'dissatisfied' responses for a subset of tickets (such as application) be predicted (modelled) based on the ticket attributes?**

Actual versus predicted satisfaction ratio for applications with more than 10 tickets:

****

While the model predicts well for dissatisfaction% below 20%, it looks like the model tends to underpredict the dissatisfaction% for values above 20%.

1. **What is the predicted dissatisfaction ratio for tickets that don't have a survey response?**The model predicts a dissatisfaction% of 8% for the tickets that don’t have a survey response. This is better than the 11% dissatisfaction for tickets with a survey response.

## possible Improvements

1. Create a trend report that highlights relevant increases in customer dissatisfaction.
2. Create report that shows additional underpinning attributes for applications with high dissatisfied%
3. Investigate why tickets opened by the service desk (instead of users through self-service) are reopened more frequently.