2020

Alfred D. Hull

Springboard

3/11/2020

Recommender Systems Final Report

[](https://github.com/ahull002/P001.recommender.systems) [](https://www.linkedin.com/in/alfredhull/) [](https://www.springboard.com/)

# **Table of Contents**

[**Table of Contents** 1](#_Toc34665517)

[Phase I. Business Understanding 3](#_Toc34665518)

[Background 3](#_Toc34665519)

[Business Case 4](#_Toc34665520)

[Set Objective 4](#_Toc34665521)

[Constraints, Limitations, and Assumptions (CLAs) 4](#_Toc34665522)

[Audience 5](#_Toc34665523)

[Cost and Benefits 5](#_Toc34665524)

[Phase II. Data Understanding 6](#_Toc34665525)

[Initial Data Collection Report 6](#_Toc34665526)

[Data Description Report 6](#_Toc34665527)

[Exploratory Data Analysis (EDA) 7](#_Toc34665528)

[Data Quality Report 9](#_Toc34665529)

[Phase III. Data Preparation 10](#_Toc34665530)

[Rationale for inclusion/exclusion 10](#_Toc34665531)

[Data Cleaning 10](#_Toc34665532)

[Phase IV. Data Modeling 12](#_Toc34665533)

[Select Modeling Technique 12](#_Toc34665534)

[Generate Test Design 12](#_Toc34665535)

[Data Sets to be Modeled 12](#_Toc34665536)

[Variance Inflation Factor (VIF) 12](#_Toc34665537)

[Build Model 13](#_Toc34665538)

[**Logistic Regression** 13](#_Toc34665539)

[Phase V. Model Evaluation 14](#_Toc34665540)

[Interpreting The Results 14](#_Toc34665541)

[Odds Ratio 15](#_Toc34665542)

[Conclusion 16](#_Toc34665543)

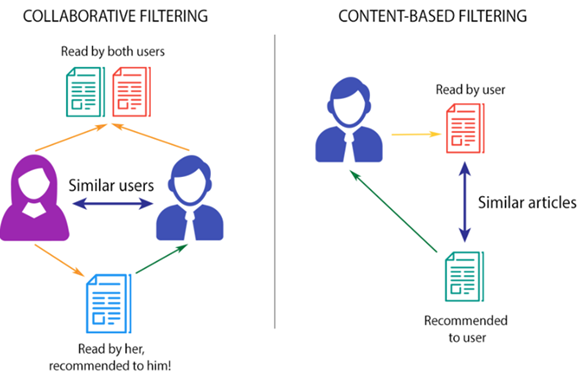
[Looking Forward 16](#_Toc34665544)

[Increase the Number of Observations 16](#_Toc34665545)

# Phase I. Business Understanding

## Background

As of 2019, we have found ourselves at a new precipice in entering the Fourth Industrial Revolution: The Real-Time Enterprise! Just like the intersections of the past: First in 1784, Water and steam; Second in 1870, First conveyor belt; and the Third in 1969, Electronics and information technology, this pause will bring with it both challenges and new opportunities. Data stewards in the discipline have realized associative costs for data storage has expounded this issue while driving talent gaps. As organizations continue to ingest voluminous amounts of data, they must become more tactical with the data they are creating and using to sustain or improve Return on Investment (ROI). More so, the approach of advancing insight through analyses of mountains of data must be clear and consistent so that results are impactful, insightful, but more so reproducible.

Given the current shortfall of talent proportionate to the opportunities seen in mining volumes of data and the time-consuming aspect in discovering the best possible options, it is ever more critical that machine intelligence and automation play a progressive role within organizations. An extremely effective way of optimizing selection in a sea of choice is by deploying recommender systems. Recommender Systems are capable of predicting the future inclination of a set of items for an individual, team, or even entire organization, then optimizing selection based on commonalities of historical transactions. One fundamental value proposition of a recommender system in modern society is that people have too many options to choose from due to the prevalence of the Internet and a decreasing cost of data storage. Imagine going into your local library and searching for a cookbook. The librarian tells you that there is an entire wing of the library dedicated to cookbooks: have fun. While the information collected in the library may sound impressive, the hunt for what you may be looking for will undoubtedly be exhausting. If there was a way to provide insight about your needs given your comfort level of sharing your data, learning algorithms could help match your ideas with others who are similar then prescribe analogs that may better suit your needs.

This project progresses logically through a recommender analysis utilizing the Cross-Industry Standard Process for Data Mining (CRISP-DM). The aim is enabling data stewards and stakeholders to clearly understand what, when, where, why, and how the Data Science work proceeds. The phases include Business understanding, Data Understanding, Data Preparation, Modeling & Application Development, Evaluation, Model Deployment.

## Set Objective

In this project, I am going to use MovieLes Dataset. This accessible dataset was put together by the Grouplens research group at the University of Minnesota. It comprises 1, 10, and 20 million ratings. And can be found at <https://grouplens.org/datasets/movielens/>.

## Constraints, Limitations, and Assumptions (CLAs)

Different types of Recommender Systems:

The most common sorts of recommendation systems are ***content-based*** and ***collaborative filtering***. In collaborative filtering, the behavior of a cluster of users form recommendations to other users. The actual recommendations predicate the preference of other users. For example, recommending a movie to a consumer because their friend also liked the movie. There are two sorts of collaborative models ***Memory-based*** methods and ***Model-based*** methods. The advantage of *memory-based* techniques is that they're simple to implement, and therefore the resulting recommendations are often easy to evaluate.

Memory-Based methods:

1. **User-based collaborative filtering**: suggests products to a consumer based on the very fact that other users like the products with a high degree of commonalities.
2. **Item-based collaborative filtering**: These systems identify similar items that supported users’ previous ratings. For example, Consumer A, B, and C gave a thumbs-up rating to books X and Y. When the next user D buys book Y, they get a recommendation to purchase book X. This is because the model identifies book X and Y as being similarly supported by the ratings of users A, B, and C.

Model-based methods

Are developed using data mining, machine learning algorithms to predict users’ ratings of unrated items. In this approach, techniques such as dimensionality reduction improve accuracy. Examples of such model-based methods include decision trees, rule-based models, Bayesian methods, and latent factor models.

Content-based methods

Use metadata like genre, producer, actor, musician to recommend items say movies or music. For example, you were supporting Infinity War that featured Robert Downey Junior because someone watched and liked Iron Man. Similarly, you'll get music recommendations from individual artists because you liked their music. Content-based systems support the thought that if you wanted a particular item, you're presumably to love something that's almost like it.

Audience

Maintenance Managers, Operations Managers, Knowledge Management Teams, Human Resource Departments in organizations, Amazon, General Motors, Ford Motors, Apple, and Boeing, and the Department of Defense, etc. can use such a model to predict the likelihood of a customer deciding to allocate resources better. They can then proactively inform their stakeholders well in advance of potential disruptions in their respective operations and or enhance customer service or reduce attrition.

## Cost and Benefits

Every maintenance hour cut in human labor will save the company and an average of 75.69 dollars, which includes fringe benefits. The company's current budget for the machine maintenance (one machine) in this analysis consists of a staff of 3 maintainers overseeing one device with 51 telemetry sensors each week working on average 50 hours a week at an operating expense of $11,353.50 a week or about $590,382. The enterprise’s other 113 machines cost $67M, and the cost savings in deploying machine intelligence will significantly support them.

# Phase II. Data Understanding

## Initial Data Collection Report

The following Data Collection Report is a simple listing of the data sources acquired along with their locations, the procedures used to procure them, and any difficulties encountered. This section will aid both with future replication of this project and with the execution of similar future projects.

**Data Features:**

1. **Data Source:** Proprietary but transformed for public use.
2. **Location**: <https://github.com/ahull002/springboard_datascience/tree/master/SB_Capstone_1_preditive_maintenance>
3. **Sensor logs**: The raw access logs contain information on a select set of sensors which feedback signal data across the enterprise.
4. **Method**: ***With full respect of the organization's security guidelines, the information utilized in the following overview*** was pulled, validated, cleared, obfuscated, and parsed to a .csv file for purposes of this project.
5. **Obstacles**: The data collection process involves a series of steps to ensure the safety and security of any analysis taking place.

At this moment, the organization has no plans to procure external databases or invest in dispatching teams to each site as its engineers, analysts, and managers are busy managing the data they currently have. At some point, however, they might want to consider an extended deployment of data mining results, in which case purchasing additional IT infrastructure to capture sensor data not centrally registered may be quite useful.

## Data Description Report

This report describes the data’s type, quantity, characteristics, and any other surface physiognomies discovered.

**Findings:**

1. Data Types (fields): float64(52), int64(1), object(2), and memory usage: 92.5+ MB
2. Data Shape: This data set consists of 220,320 observations/records, by 55 fields/columns.
3. Data Characteristics: Based on the datatypes and shape consisting of independent and dependent labels, I would suggest prescribing supervised machine learning. Conversely, logistic regression analysis as the sensor reading labels, and its observations could be passed for independent variables while the machine status label and its records could pass for the dependent/target variables: 'NORMAL'=NORMAL(0), 'BROKEN'=FAIL(1), 'RECOVERING'=NORMAL(0).

## Exploratory Data Analysis (EDA)

EDA describes the results of exploring the data involved in this project. EDA includes the first findings and or initial hypothesis and their impact on the remainder of this project.

**Initial Hypothesis:**

***First Test***

**H0** *Sensor readings cannot help proactively to predict system failure and or catastrophic changes to this machine.*

**Ha** *Sensor readings can help proactively to predict system failure and or catastrophic changes to this machine.*

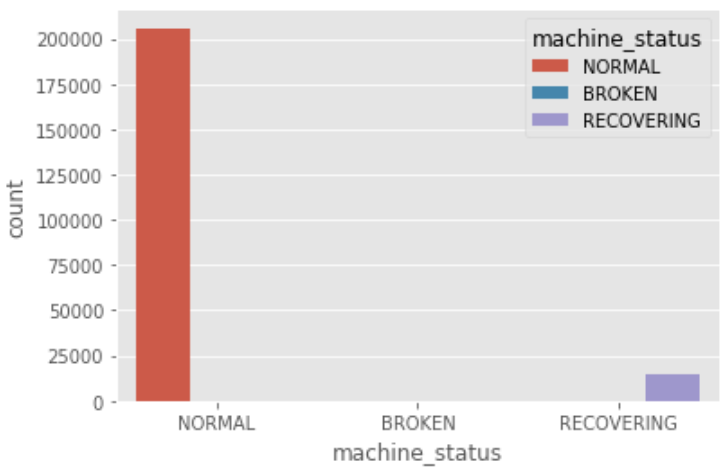
***Second Test***

**H0** *There is no significant evidence to suggest that failure occurs in some cyclical time frame based on machine run time.*

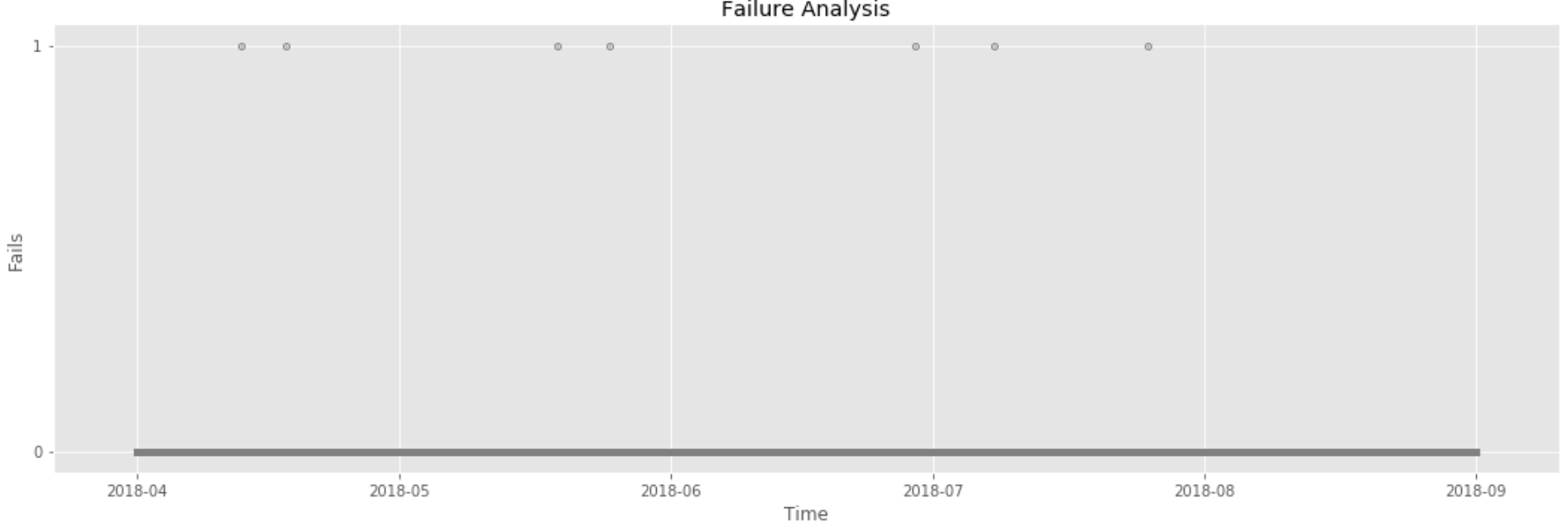
**Ha** *There is significant evidence to suggest that failure occurs in some cyclical time frame based on machine run time.*

**Data Visualization and Interpretation:**

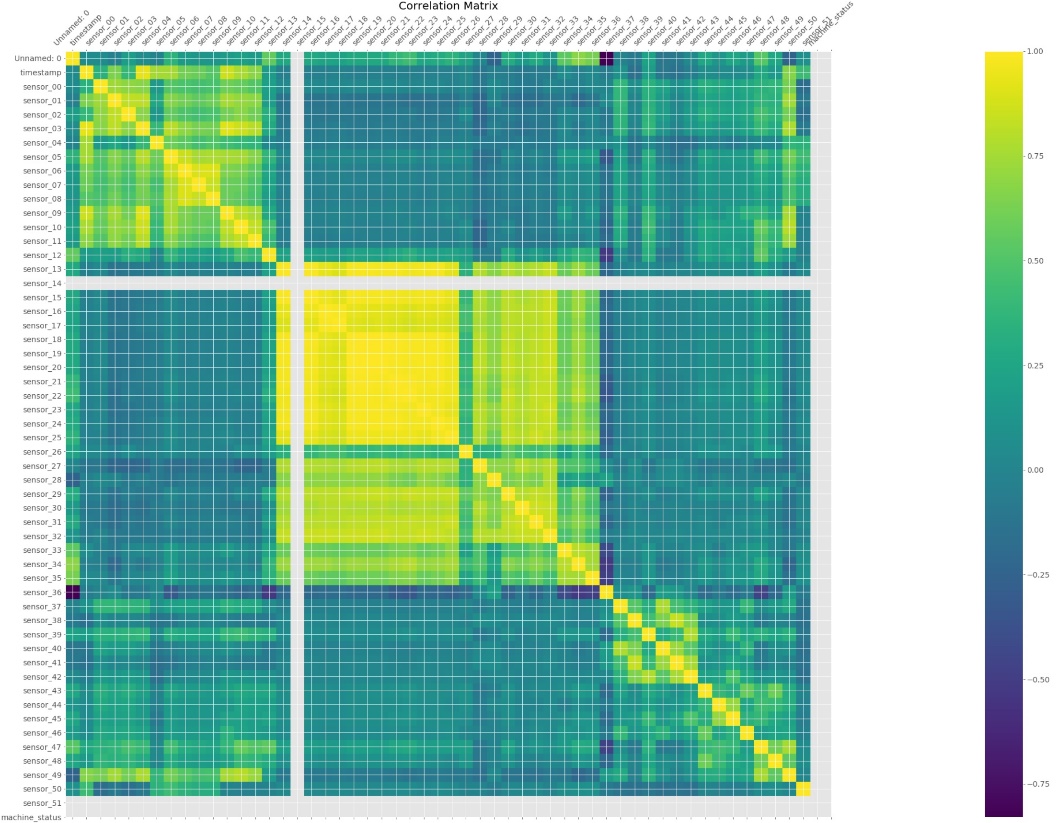
Variable Distribution

There is a class bias, meaning a condition observed when the proportion of events is much smaller than the proportion of non-events. So we must sample the observations in nearly equal proportions to get more efficient models. The following chart displays that the dataset is extremely unbalanced with readings showing: NORMAL = 205,836, RECOVERING = 14,477, and BROKEN = 7. The unbalanced characteristic of this dataset will lead to skewness and or bias in forwarding predictive models if not addressed. To address the unbalanced characteristic here, we may need to apply: upsampling of the dependent variable (broken) to filter out the noise associated with the high distribution of observations for normal and or recovering.

Time-Series Analysis



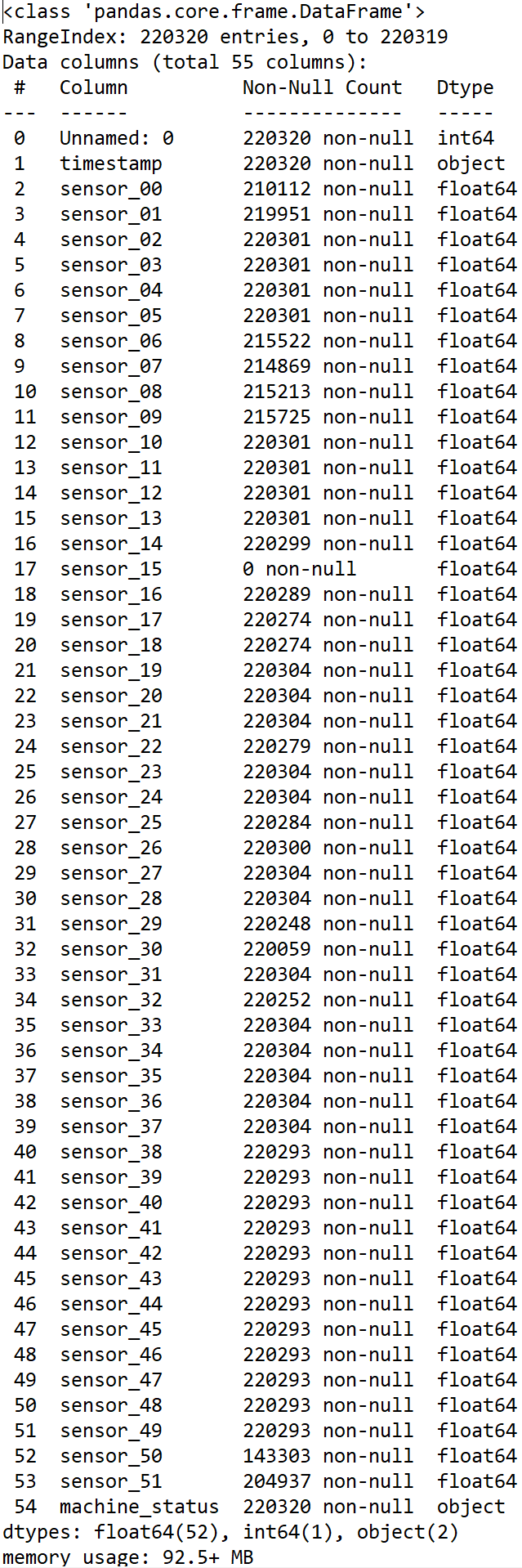
The above time series chart displays that the timing of the month does not appear to be a factor with sensor failures. For instance: In April, they appear in the middle of the month, March the middle to end of the month, June only once, July beginning and end of the month, and August none. We may need to go deeper in examining machine utilization rates and or timing in between failures to engineer a new feature: Meantime between failures in future analysis.

Feature Correlation Analysis

At a very high level, we can see there are various ranges of correlation in the data. We must be careful to ensure that these correlated values together do not outweigh other variables, thus skewing the results of our predictive model. Additionally, we can see that there may also be some data quality issues and or problems with sensor readings. For example, sensor\_15 appears to be not producing results at all in the time frame of this supplied data source. There is also an unnamed feature aside from our sensors that we must ensure is understood before moving forward to data preparation.

## Data Quality Report

This report lists the results of the data quality verification along with solutions vetted by subject matter experts to validate any concerns by assessing the following questions: Is the data complete and accurate; does it encompass errors and how popular are they?



As we can see with the information printout from pandas, the dataset is unbalanced and will require tidying and wrangling in the Data Prep Phase. For instance:

1. There were 220,320 observations recorded during the five months.
2. The first column index [0] is not named: and appears to be a duplicated index.
3. The timestamp is an object data type rather than a date-time value
4. Sensor\_15 has no benefits listed.
5. The overall field data types show mixes between three types: float64(52), int64(1), and object(2).

After speaking with the subject matter expert, she suggested that replacing any null values (miss readings) seen in the data with the mean of that respective sensors run. Additionally, she recommended that machine statuses should are as follows; Normal is Normal or 0, Broken is Failed, or 1, and Recovering is Normal or 0.

# Phase III. Data Preparation

## Rationale for inclusion/exclusion

Steps:

1. Dropped 'sensor\_15': This field was deemed to be a misreading and could skew data once the sensor has been turned back on. Being that any future predictive model would be re-trained/tuned repeatedly, it does not make sense now to add unnecessary noise.
2. Updated machine\_status (dependent variable) datatype from object to numeric: Normal = 0, Recovering = 1, Broken = 1
3. Dropped 'timestamp': As we are only concerned with predicting machine failure based on sensor readings. Time-series data will most likely not have any impact on accepting the null as the upsampling process will remove different chunks of time. Due to this and not having enough data, correctly, the machine runs time data I will prematurely pause analysis for the second test:

***Second Test***

**H0** *There is no significant evidence to suggest that failure occurs in some cyclical time frame based on machine run time.*

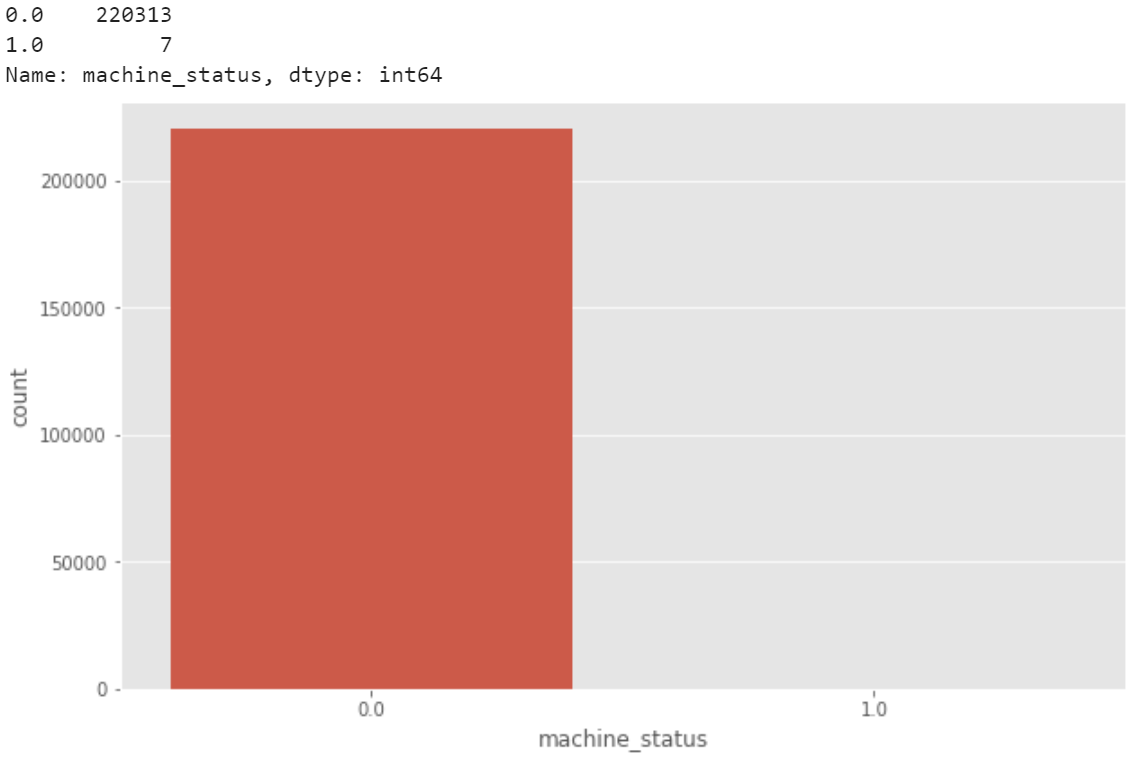
**Ha** *There is significant evidence to suggest that failure occurs in some cyclical time frame based on machine run time.*

## Data Cleaning

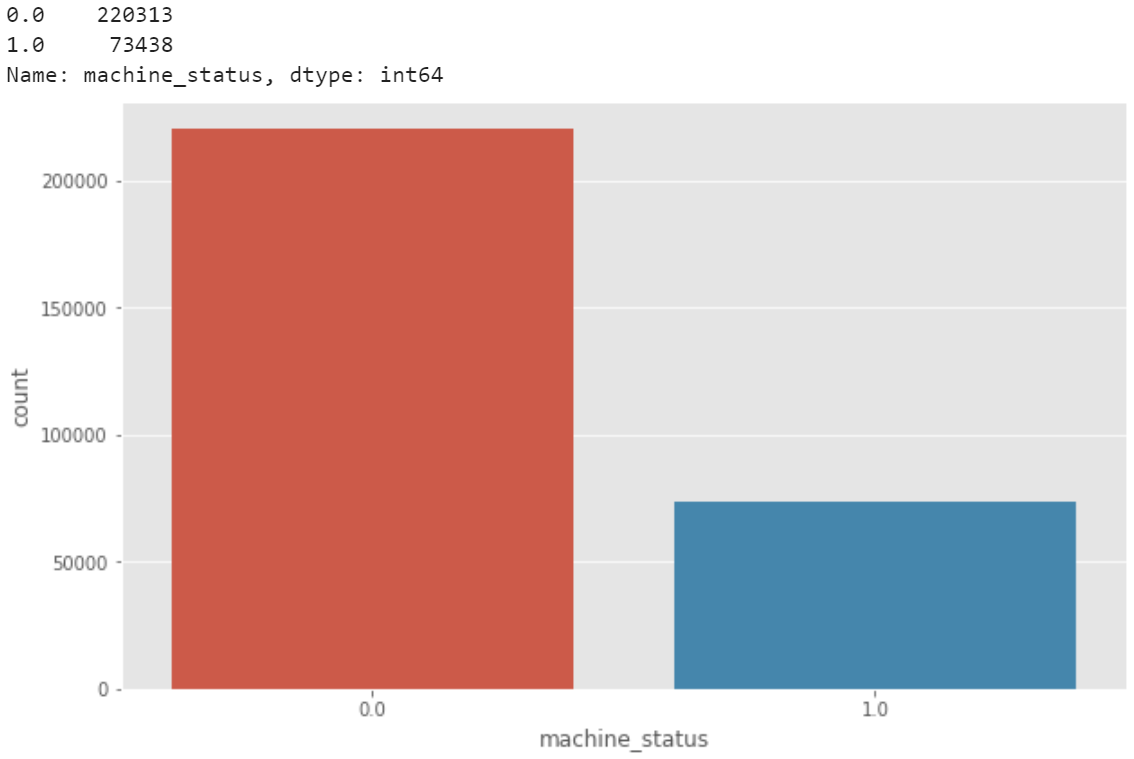
Here I will describe decisions and actions taken to address data quality problems. I will also consider any transformations I made on the data for cleaning purposes and their potential impact on the analysis results.

**Steps involved:**

1. Logistic regression models require that the independent (x-variables) are of a consistent shape with the dependent (y-variable): tidying.
2. It was critical that management normalized the understanding of the y-variable: 'NORMAL': '0', 'RECOVERING': '0', 'BROKEN': '1', so that the target being predicted is consistent (220,320).



1. Filled nan values with column mean.
2. Up-sampled failed readings 1:4 (25% failure and 75% normal) for a total of 293,751 to reduce the amplification of unbalanced data seen in the lack of failed instances.



# Phase IV. Data Modeling

## Select Modeling Technique

1. Based on the data set, we have it is proper to deploy either a random forest model and or a logistic regression model as the x-variables (independent) are of a consistent shape with the y-variable (dependent).

* **Logistics regression** is applicable when we have one nominal variable with two values (male/female, 0/1, etc.) and one measurement variable. The nominal variable is the dependent variable, and the measurement variable is the independent variable. Additionally, because we are dealing with a nominal variable (1 or 0): 'NORMAL': '0', 'RECOVERING': '0', 'BROKEN': '1', we could use a logistic regression model.
* In machine learning, the **Random Forest Algorithm** also known as the random forest classifier. It is a popular classification algorithm. One of the most exciting things about this algorithm is that it can double both as a classification and a regression algorithm. A random forest algorithm creates multiple decision trees and merges them to obtain a more stable and accurate prediction. In general, the more trees in the forest, the more robust would be the prediction and thus higher accuracy.

## Generate Test Design

The data-frame from Phase III. Data Preparation divided into a Train and Test set (70/30) will be analyzed from this point forward. In this case, this means that 88,123 observations were designated for testing, while the balance: 205,525 views will train the model.

### Data Sets to be Modeled

1. All Features
   1. 49 features of the 51 used to predict machine failure before VIF
2. Variance Inflation Factor (VIF) Features
   1. Five features used to predict machine failure

### Variance Inflation Factor (VIF)

Collinearity is when two or more variables highly correlate with one another; in other words, duplicate information within a dataset. Ideally, features in a dataset display different information. Collinearity can cause inflation of the variance of a regression coefficient, which may cause predictions with significant errors. A VIF value is calculated for each feature/column of data:

After removing all features where the VIF is more significant than five, the dataset is left with only five features: sensor\_13, sensor\_34, sensor\_37, sensor\_38, and sensor\_48.

## Build Model

### Logistic Regression

**logit(p) = log(p/(1-p))= β0 + β1\*sensor\_1 + β2\*sensor\_2 + β3\*sensor\_3 + BnXn**

Logistic regression introduced in the early twentieth century was used across many applications in the biological sciences. Logistic regression is employed when the dependent variable(target) is categorical.

For example,

To predict whether an email may be spam (1) or (0)

Whether the tumor is malignant (1) or not (0)

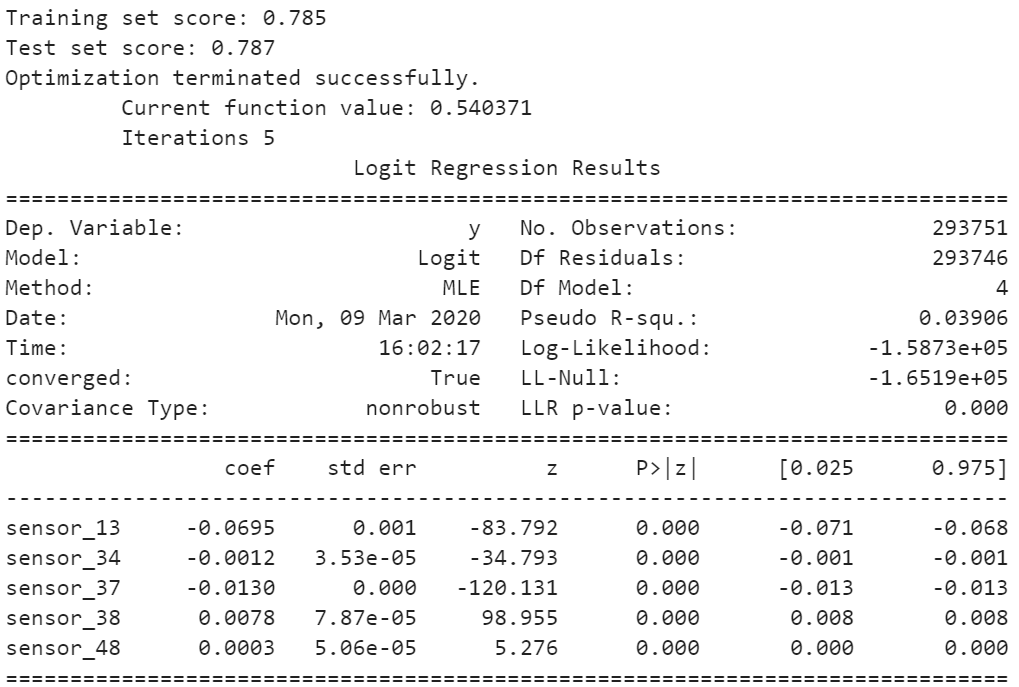
# Phase V. Model Evaluation

Here I will interpret the models according to my domain knowledge, the data mining success criteria, and the anticipated test design. I will also decide on the success of the application of the modeling and discovery methods technically, may contact concerned business analysts and domain experts later to discuss data mining results in the appropriate business perspective. While this task considers models, the evaluation phase (Phase 5) further considers all other effects that occurred in the course of this analysis.

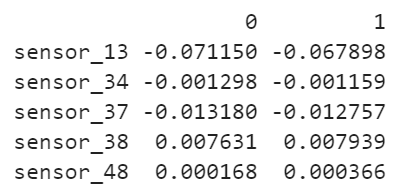
At this stage, it is also important to rank the constructed models while assessing them according to any evaluation criteria. I will take the business objectives' success criteria into account as far as possible here.

## Interpreting The Results

A function was created to initialize, train, and test a Logistic Regression model. Here is an overview of the coefficients of the model, how well those coefficients fit, the overall fit quality, and several other statistical measures.

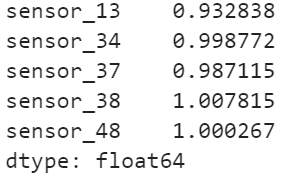


The resulting object also lets you isolate and inspect parts of the model output. The confidence interval gives you an idea of how robust the coefficients of the model are. For instance, sensor\_13, 34, and 37 have an inverse relationship on system failure comparative to sensor\_38, 48.

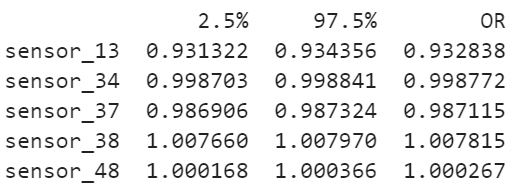


## Odds Ratio

Take the exponential of each of the coefficients to generate the odds ratios. This tells you how a 1 unit increase or decrease in a variable affects the odds of a systemic failure. As you can see in the output below, a 1 unit increase will offset system failure on sensors 13, 34, and 37, while conversely, an improvement on 38 and 48 will add to system failure.



We can also do the same calculations using the coefficients estimated using the confidence interval to get a better picture of how uncertainty in variables can impact the failure rate.



# Conclusion

Primarily the analysis read into the sensors and discovered that virtually five sensors readings point to system failure: sensor\_13, 34, and 37 have an inverse or negative relationship on system failure comparative to sensor\_38, 48.

## Looking Forward

### Increase the Number of Observations

I think it would be genuinely impactful if we were able to look at seasonality over a two-year time horizon. Given the class unbalance seen in the data set, I would recommend pulling in more on this machine to analyze normal to failed rates through automated time series. Examining the state of sensor readings 7 to 10 days before a failed event occurred would be necessary along with the hours ran a feature to see precisely how long the machine was running before failure.

