**Predictive Modeling Distribution Assets**

February 7, 2018

**TABLE OF CONTENTS**

[**1.** **Introduction** 3](#_Toc505346014)

[**2.** **Objective** 3](#_Toc505346015)

[**3.** **Input Data and Assumptions** 3](#_Toc505346016)

[**4.** **Predictive Models** 5](#_Toc505346017)

[**5.** **Applications and Tools Used** 6](#_Toc505346018)

[**6.** **Analytical Methods Used for Predictive Analysis** 7](#_Toc505346019)

[**7.** **Measurements** 9](#_Toc505346020)

[**8.** **Predictive Analysis Results** 9](#_Toc505346021)

[**9.** **Advantages of Predictive Model vs. Age-Based Model** 11](#_Toc505346022)

[**10. Implementation** 11](#_Toc505346023)

[**11. Future Steps** 12](#_Toc505346024)

[**Appendix A: Variables & Data Sources** 13](#_Toc505346025)

[**Appendix B: Data Inclusion Specifics for Benefit Analysis** 18](#_Toc505346026)

[**Appendix C: Predictive Machine Learning Algorithms** 20](#_Toc505346027)

# **Introduction**

Southern California Edison (SCE) has developed predictive models for selected major distribution system assets. The predictive models were developed for transformers, poles, overhead conductors, underground primary cables and overhead and underground switches. The need for developing predictive models was identified in 2016 as a part of an overall goal to improve distribution system by replacing equipment before it fails in-service. The process of proactively replacing equipment that is most likely to fail in-service is referred to as predictive maintenance.

# **Objective**

The objective of predictive analysis is to improve system performance by applying analytics to better identify and predict equipment failures at an individual asset-basis and partner with execution organizations in creating processes to cost-effectively and strategically implement these benefits.

# **Input Data and Assumptions**

An extensive series of input variables is used to develop predictive models. Five input data categories were identified in order to investigate and quantify effects of variables on asset performance over its lifetime. The range of variables included electric and non-electric variables, operational, weather, environmental and geographical data. Descriptions of these six categories are provided below.

1: Electrical Data

The modeling of each asset type requires different sets of electrical input data. Typical electrical characteristics included in the predictive models are circuit nominal voltage level (kV), circuit loading (kVA), number of customers per circuit, underground or overhead application, conductor size, type, material – copper or aluminum, insulation type, cable underground installation type, etc. The short-circuit duty was calculated at each segment, structure-with-equipment to structure-with-equipment, on the underground distribution circuit. The calculation was performed using the breaker duty at the substation, cable impedance, and length of the segment between two structures with equipment.

2: Non-electrical Data

Examples of non-electrical data included are component installation date or years in-service (asset age), region and district data, and end-type user (commercial, small and large business, industrial).

3: Historical Performance Data

The effect of historical asset performance was captured by taking into consideration historical outages in the past 5-10 years. The date of operation and the number of CB/switch operations and other operational characteristics were also considered, if suitable and available for the operation of that asset category.

4: Weather Data

SCE obtained hourly weather data from over 25 identified weather stations across the SCE territory as well as data provided by our third party consultant ADS. Weather conditions across the SCE territory were considered for by taking into account a weighted average of ambient temperature, hourly peak temperature, dew point temperature, wet-bulb temperature, relative humidity, directional wind speed, maximum wind speed, solar radiation data and station pressure.

5: Environmental Data

Groundwater and water table elevation data was gathered from an open data source supplied by the California Department of Water Resources.

6: Geographical Data

Asset sensitivity to location is included by modeling longitude, latitude, elevation and average grade or slope of a terrain.

Data Assumptions and Limitations

The accuracy of the predicted results is very dependent on the accuracy of the data used to build the predictive models. Some assets have limited data and some assets, such as underground primary cables, required extensive resources to identify accurate failure data. For example, FIM maps were visually examined for identifying segment-level failures for UG Cable for a limited time-period, which is time consuming, and therefore can only be done on a limited basis.

Data integrity issues were identified in the process of developing predictive models. Data “amputation” was used to populate missing data. The assumptions used for the data amputation utilized SCE’s Distribution Design Standard (DDS), engineering judgement, manufacturer data and acceptable engineering practices.

# **Predictive Models**

Predictive models are the foundation of the predictive maintenance effort. The predictive models estimate the historical impact of all variables to predict an asset’s future performance. The viability of the predictive analysis results depend on the accuracy of the predictive models. Eight predictive models for four types of assets were developed to perform predictive analysis as shown in Table 2.

**Table 2: Predictive Models**

| **Predictive Models by Asset Category** | **Predictive Models by Asset Type** |
| --- | --- |
| Underground primary cable failure models | Underground primary mainline cable & unfused radial cable model |
| Transformer models | Transformer model |
| Overhead conductors model | Overhead conductor model |
| Switch models | BURD switch model |
| Oil & Gas switch model |
| Overhead switch model |
| Automated switch model |
| Padmount switch model |

Although a pole predictive model was created, the model is excluded from this report due to insignificant SAIDI improvements. The majority of the pole failures are due to the “Third Party Category” and a “Car Hit Pole” sub-cause which was excluded from the predictive maintenance analysis as these conditions could not be predicted. The main reason is that the pole inspection cycle and the pole replacement program is cost-efficient and minimal benefits were identified through predictive maintenance analysis.

In order to create a final output, showing probability of failure for each individual asset, multiple sets of data were used for each asset category to create the final output – this is due to creating and applying a predictive model, as explained in more detail in later Sections of this report. The first two datasets used are the training and testing datasets, which are used to create the predictive model. The model is then applied to active assets in order to predict probabilities of failure for each individual asset.

# **Applications and Tools Used**

The list of applications used to obtain the input data utilized for building and validation of the predictive models is shown in Table 3. Python and R programming languages were used to perform predictive analysis.

**Table 3: Application/Tool/Software Utilized to Build Predictive Models**

| **Application/Tool/Software Utilized to Build and Validate Predictive Models** | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| **Application/Tool/Software Name** | **Application/Tool/Software Description** | **Information Used for Predictive Models** | **Transformers** | **Switches** | **Overhead Conductors** | **Underground Primary Cables** |
| eMAP | Circuit maps. |  |  |  |  | x |
| Facilities Inventory Mapping (FIM) | Map based inventory of SCE assets. | Age, insulation type, installation type, size, length, number of phases and voltage rating for each underground cable/segment. |  |  |  | x |
| Outage Database & Metrics (ODRM) | Database containing essential information about area, circuit, substation and transmission outages. | ODRM is used for extracting the SAIDI, SAIFI, CMI, CI, outage start and end time. | x | x | x | x |
| Outage Management System (OMS) | Allows real-time monitoring of distribution circuits. | OMS is used for extracting the number of downstream customers, downstream kVA/% loading, radial vs mainline cable. | x | x | x | x |
| Repair Orders (RO) | Detail incident description. | Material list, incident site picture, incident date, “from” and “to” structure, load off/on, crew type |  |  |  | x |
| SAP | Tool used to establish SCE’s assets | Component Data | x | x | x | x |
| cGIS (Geospatial Information System) | GIS data is a form of geospatial data. | Location, flood zone, elevation, | x | x | x | x |
| Weather | Weather stations (from meteorology group) | Temperature, Solar irradiation, Humidity, Wind speed, Wind direction, Rain. | x | x | x | x |
| Environmental Data | California Department of Water Resources. | Soil chemistry, soil composition, water table and groundwater data. |  | x |  |  |
| Teradata | AMI data | Loading/usage | x | x | x | x |
| Python | Python is a widely used high-level programming language with a multitude of computing, modeling, and visualization packages. |  |  | x |  |  |
| R | R is an open source programming language and software environment for statistical computing and graphics. |  | x | x | x | x |

Predictive models were developed using historical asset data sets and performance, using a rolling 4-year historical dataset. For example, the 2012-2015 data set for transformers and switches was used to predict 2016 failures. Those 2016 predictions were then tested by identifying how many failures actually occurred within 2016, as compared to the assets that had the highest probabilities of failure, as further described in Section 8. The model then used 2013-2016 data to predict failures in 2017.

**Table 4: Time Range of Historical Data Utilized for Predictive Modeling**

| **Predictive Model** | **Input Data Time Range** | **To Predict** |
| --- | --- | --- |
| Underground Primary Cable | 2013-2016 | 2017 |
| Transformers | 2013-2016 | 2017 |
| Overhead Conductors | 2014-2016\* | 2017 |
| Switches | 2013-2016 | 2017 |

*\*Due to data discrepancies prior to 2014, only 2014+ data was utilized for the purpose of building the overhead conductor predictive model*

# **Analytical Methods Used for Predictive Analysis**

Types of predictive methods include simulation and machine-learning methods. Typical simulation methods are Markov method and network models. Typical machine-learning methods are binomial regression, decision tree, random forest modeling (RFM) and gradient boosting. The predictive method chosen depends on variable availability, variable types, time constraints, expected accuracy, etc.

Predictive analysis models were developed using the Gradient Boosting (GBM) and Random Forest Model (RFM) machine learning algorithms as shown in Table 5.

**Table 5: Machine Learning Algorithm for Various Predictive Models**

| **Predictive Model** | **Machine Learning Algorithm** |
| --- | --- |
| Transformers | Random Forest |
| Switches | Random Forest, Gradient Boosting |
| Overhead Conductors | Gradient Boosting |
| Underground Primary Cables | Gradient Boosting |

The output of the machine learning algorithms is the relative likelihood of an in-service failure, due to reasons that could be avoided with proactive replacement .

Decision Tree Method

Machine learning algorithms could be described as algorithms that “recognize the patterns” and are using “experience” to make predictions. The term “experience” is used to describe historical component performance. Expression “to recognize the patterns” is used to describe a voting process built in the algorithm. The output of the voting process is prediction.

There are different types of machine learning algorithms based on the output variable, which can be categorical or continuous. An example of categorical variable is asset type. An example of continuous variable is temperature. Classification and clustering machine learning algorithms are used for categorical response, and regression and dimensionality reduction is used for continuous response. The response variable or predictor is an output variable, which is used to make a prediction. It was determined that for the purpose of SCE’s predictive analysis, random forest machine-learning and gradient boosting algorithms perform best.

For all machine learning algorithms created for predictive maintenance, a standard process was followed. The first step of the analysis was to split the historical input datasets into a training dataset and a testing dataset. The input dataset contains historical component failures in the past 4-5 years. The training dataset typically contains 70-80% of the input dataset and the test dataset contains the remaining 20-30% of the historical dataset. The predictive algorithm is created using the training dataset, 70-80% of the historical dataset, and is built by looking at all interactions between different variables to find patterns, specifically to predict equipment failures. The next step is to test the algorithm on the remaining 20-30% of the dataset, or the ‘testing’ dataset. This completes the validation of the model, which is the typical statistical method used to validate predictive models.

Once statistically validated, the model is then validated through comparing predicted failures to actual failures for 2016, since 2016 data was not used in training the model. For this validation, the model is applied to assets that are active starting in 2016 (as of 1/1/2016). Predictions are made on those assets, by assigning probabilities of failure to each asset. In this final step, benefits are quantified by comparing the prediction model results to the age-based model results (or if not age-based, the current method in place). For example, to validate the transformer predictive model, predictions were made using data through end of 2015 for predicting 2016 failures, and then the model was compared to actual 2016 transformer failures. A flow chart is provided below to visually show this validation process for point assets (Switches, Transformers). For linear assets (OH Conductor, UG Cable), the process is adapted, based on the data available.



The methodologies applied to perform predictive modeling are described in more detail in Appendix C.

# **Measurements**

An essential part of the quantitative study is the definition of the criteria in comparison with the performance of the asset/circuit. The SAIDI, probability of failure and ranking on an asset-by-asset basis was used to quantify performance improvement.

The criteria for success and failure was developed for each asset category. The probability of failure criteria defines thresholds established for each asset category to define a success or failure as an output of the machine learning algorithm. True positive, true negative, false positive and false negative result counts were calculated for each asset type.

# **Predictive Analysis Results**

“Matching rates” (see Formula 1 below) of both the predictive model and age-based model for each asset were obtained to identify the benefits of using a predictive model. These matching rates allow comparison between the predictive models against age-based modeling, where age was the only variable used to identify which assets are most likely to fail (ranked from oldest to newest asset). The results of the predictive analysis show that the predictive model predicts two to five times more failures than the comparative age-based models, when looking at matching ratios 1 – 2 years out.

To better illustrate how estimated benefits were calculated, we can reference the BURD Switch model. The top 500 BURD Switches that were predicted to fail using an age-based model were compared against the top 500 BURD Switches that were predicted to fail using a predictive model. The age-based model predicted failure of only 18 BURD Switches within the first year, and the probability based model predicted the failure of 42 BURD Switches. The match rate for the age-based and predictive models is calculated as for BURD Switches (BSw):

(Formula 1)

Based on these match rates, we can see that the predictive match rate is better than the age-based model, when comparing the top 500 BURD Switches

A summary of the initial results for each asset are summarized below. It is understood that when implementing predictive technologies to scope work, existing operational constraints must be taken into consideration.

**Table 8: Predictive Model SAIDI Proposed Improvements**

| **Predictive Model SAID Reduction** | **Comparison of Top #** | **Age-Based Match Rate** | **Predictive Model Match Rate** |
| --- | --- | --- | --- |
| Underground primary cable | 1,500  (segments) | 1.5% | 10.5% |
| Transformers | 1,000 | 5% | 22% |
| Overhead Conductors | TBD | TBD | TBD |
| Switches - BURD | 500 | 4% | 8% |
| Switches – OH | 500 | 5% | 11% |
| Switches – Mainline Oil/Gas | 200 | 8% | 16% |
| Switches – Automated | 100 | 1% | 8% |
| Switches – Padmount | 200 | TBD | 3% |

# **Advantages of Predictive Model vs. Age-Based Model**

The age-based model is a powerful framework for analysis applied successfully for many decades. However, the age-based method does not include the effects of the available electrical and non-electrical asset characteristics, weather, environment and asset historical performance that is available to us. Asset age is a dominant variable in predictive modeling, however, it is also able to take into consideration all of the other relevant electrical and non-electrical asset characteristics.

To illustrate the need for the broad vision we can consider the following questions: Would it be possible to assign a health index to cables without knowing the environment that the cable was exposed to? Will a cable that was overloaded multiple times during excessive summer heat be in the same condition as a cable that was not overloaded, if both are the same age?

Machine learning algorithms have made possible the ability to asses and evaluate the effects of multiple variables on individual assets, recognizing patterns that previous models have not been able to recognize, and assign a relative probability of failure on asset-basis.

# **Implementation**

The results of predictive analysis show that evaluation and assessment of historical asset performance and analytical methods, such as machine learning, could be used to predict the future asset performance and improve system . These results may help to improve the accuracy and efficiency of the existing infrastructure replacement process and improve decision-making processes of all business lines, including planning, operations and maintenance.

Another potential application of the results of the predictive analysis is a review of the distribution system design standard. For example, knowing historical performance of the underground cables and knowing the probability of failure per cable segment may help to determine the most optimal underground cable segment length based on the least probability of failure.

# **Future Steps**

One of the most important factors to improve predictive models and system is to continuously maintain and improve asset data procurement. The quality of the input data is directly related to the results of the predictive analysis. Additional data that could be incorporated into future predictive models include:

1. impact of actual time-varying load throughout the year
2. user information including seasonal kW peak and energy load profile
3. additional environmental input data such as elevation gradient
4. additional weather data such as difference between the daily maximum and daily minimum temperature
5. infrared inspection data
6. improved accuracy of the connectivity model

If any future predictive modeling changes are deemed to be necessary by interested parties, we will incorporate these changes into the predictive models.

# **Appendix A: Variables & Data Sources**

**A1. Electrical Input Variables**

| **Electrical Input Variables** | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| **Input Variable Name** | **Description** | **Source** | **Transformers** | **Switches** | **Overhead Conductors** | **Underground Primary Cables** |
| KVA | Asset nominal kVA rating. | SAP | x |  |  |  |
| PRIMARY\_RATING |  | SAP | x |  |  |  |
| LINE\_VOLTAGE | Phase-to-phase voltage. | SAP | x |  |  |  |
| PHASE\_VOLTAGE | Line to ground voltage. | SAP | x |  |  |  |
| PrimaryVoltage | Transformer high side voltage. | SAP | x |  |  |  |
| SecondaryVoltage | Transformer low side voltage. | SAP | x |  |  |  |
| Short Circuit Duty | Short circuit current at the substation. | SAP | x |  |  |  |
| Asset nominal current |  | IT/GIS Cable Database |  |  | x | x |
| Current density |  | IT/GIS Cable Database |  |  | x | x |
| Subtype | Dry type or oil type transformer. | SAP | x | x |  |  |
| TransformerClass | Class I or class II transformer insulation | SAP | x |  |  |  |
| Usage | (Eric to check) Indoor or outdoor | SAP | x |  |  |  |
| User\_status |  | SAP | x | x |  |  |
| CircuitID | Circuit ID number | SAP | x | x |  |  |
| Voltage | (Eric to check) Primary operating voltage or Design voltage | SAP | x |  |  |  |
| Manufacturer | Manufacturer name (ABB, GE, etc.) | SAP | x | x |  |  |
| ModelNumber | Asset serial number | SAP | x | x |  |  |
| Time Operated |  |  |  | x |  |  |

**A2. Non-electrical Input Variables**

| **Assets Non-Electrical Input Variables** | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| **Input Variable Name** | **Description** | **Source** | **Transformers** | **Switches** | **Overhead Conductors** | **Underground Primary Cables** |
| Age | Asset true age. | SAP | x | x | x | x |
| Start\_Up\_Date | An actual date when the asset was installed in the field at the specific location. | SAP | x | x | x | x |
| Commercial\_Indicator | Commercial customer is an end-type user. | SAP | x | x | x | x |
| RemovalReason | The reason for removing asset from the field (failed, replaced, re-located, spare). | SAP | x | x | x | x |

**A3. Historical Performance Data**

| **Assets Historical Performance Data** | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| **Input Variable Name** | **Description** | **Source** | **Transformers** | **Switches** | **Overhead Conductors** | **Underground Primary Cables** |
| Outage Information | Asset historical performance | ODRM | x | x | x | x |
| Asset operational data | The number of switching device operations | SAP |  | x |  |  |

**A4. Weather Input Variables**

| **Weather Input Variables** | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| **Input Variable Name** | **Description** | **Source** | **Transformers** | **Switches** | **Overhead Conductors** | **Underground Primary Cables** |
| Temperature | Ambient temperature, hourly peak temperature, dew point temperature, wet-bulb temperature | Weather stations\* | x | x | x | x |
| Solar Irradiation | The amount of solar energy that can be “harvested” per ft2 (kWh/m2). | Weather stations\* | x | x | x | x |
| Humidity | Presence of moisture and humidity in the field negatively impacts transformers and cable insulation, transformer oil condition, etc. | Weather stations\* | x | x | x | x |
| Wind | Location average wind speed measured in meters/sec). | Weather stations\* | x | x | x | x |
| Rain | Location average rainfall measured in mm. | Weather stations\* | x | x | x | x |

*\*25 weather stations, across SCE territory – data supplied by internal meteorology group*

**A5. Environmental Input Variables**

| **Environmental Input Variables** | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| **Input Variable Name** | **Description** | **Source** | **Transformers** | **Switches** | **Overhead Conductors** | **Underground Primary Cables** |
| ClimateZone | Assets located is the extreme climate could experience more failures. |  | x |  |  | x |
| Climate Code |  |  | x | x |  | x |
| FloodZone | Floodplain maps identify flood hazard areas. |  | x | x | x | x |
| ρs | Soil Resistivity | Substation Grounding System database input data | x |  |  | x |

**A6. Geographical Performance of the Circuit Input Data**

| **Geographical Input Variables** | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| **Input Variable Name** | **Description** | **Source** | **Transformers** | **Switches** | **Overhead Conductors** | **Underground Primary Cables** |
| OHUG | Overhead or underground asset | SAP, OMS, eMAP, FIM | x | x | x | x |
| Latitude | Geographical coordinates | cGIS | x | x | x | x |
| Longitude | Geographical coordinates | cGIS | x | x | x | x |
| Elevation | Geographical coordinates | cGIS | x | x | x | x |
| HighFireArea | Is asset located in the high-fire hazard area. | SAP, OMS, eMAP, FIM | x | x | x | x |
| FunctionalLocation | Functional location number (FLOC) for poles and underground structures. | SAP | x | x | x | x |
| CorrossionZone | Equipment is installed in the environment prone to electrochemical process. | cGIS | x | x | x | x |
| Region | Metro East, Metro West, North Coast, Orange, Rurals, San Jacinto, San Joaquin. | SAP | x | x | x | x |
| Structure | Vault, splice box, manhole, burd structure, padmount, encloser or CST. | SAP | x | x | x | x |

# **Appendix B: Data Inclusion Specifics for Benefit Analysis**

All ODRM outage counts excluded 3rd Party and Animal outage causes.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model - Equipment | SAP Object Types | SAP SubTypes | ODRM System Level 3 Equipment | ODRM Assumptions |
| Transformers | ED\_XFMR | BD-BURD,CP-CURRENT/ SELF PROTECT,CV-CONVENTIONAL,DF-DOUGLAS FIR,DY-DRY,FI-FUSED INTERNAL,FOR BURD,LTC,MC-MODIFIED CONVENTIONAL,MP-MODIFIED SELF PROTECT,OT-OTHER,PAD,PAD-PAD MOUNTED,PLAS-PLASTIC,PM-PADMOUNT,RO-REGULATED OUTPUT,SP-SPECIAL,ST-CST,SU-SUBWAY | BURD TRANSFORMERS, PADMOUNT TRANSFORMERS, TRANSFORMERS, CURRENT TRANSFORMER, POTENTIAL TRANSFORMER |  |
| BURD Switches | ED\_SWITCH, ED\_SWGAS | BD-BURD,BD-OIL BURD SWITCH,BDS-GAS BURD SWITCH | BURD SWITCH |  |
| Mainline Oil/Gas Switches – Oil | ED\_SWITCH, ED\_SWGAS | RAC-OIL SWITCH RAC,RAD-OIL SWITCH RAD,RAJ-OIL SWITCH RAJ,RAK-OIL SWITCH RAK,RAL-OIL SWITCH RAL,RAM-OIL SWITCH RAM | OIL SWITCH | Get the line device number from SAP using the ODRM "CKT\_NRST\_STRCT\_NUM". Exclude Padmounted Switches. |
| Mainline Oil/Gas Switches – Gas | ED\_SWITCH, ED\_SWGAS | PMS-PAD MOUNTED SWITCH,PSS-GAS PAD MOUNT,RAG-GAS SWITCH RAG,RMS-GAS SWITCH RAM,RQS-GAS SWITCH RAQ,RCS-GAS SWITCH RAM,RCS-GAS SWITCH RAM | GAS SWITCH | Get the line device number from SAP using the ODRM "CKT\_NRST\_STRCT\_NUM". Do not include RCS or RAR switches |
| OH Switches | ED\_SWITCH, ED\_SWGAS | DS1-DISCONNECT SINGLE PHASE,DS3-DISCONNECT 3 PHASE,GOP-GANG OPERATED ,HORIZONTAL,HORIZONTAL INVERTED,OMN-OMNI SWITCH,VERTICAL,VERTICAL INVERTED | SWITCH/DISCONNECT/AR | Get the line device number from SAP using the ODRM "CKT\_NRST\_STRCT\_NUM". Do not include RCS or RAR switches |
| Automated Switches - RAR | ED\_SWITCH, ED\_REC | AIDR-ASSET INFORMATION DETAILS,GOP-GANG OPERATED,LOW-LOW GROUND,PH-PHASE GROUND,SEN-SENSITIVE GROUND,VAC-VACUUM SWITCH | SWITCH/DISCONNECT/AR | Get the line device number from SAP using the ODRM "CKT\_NRST\_STRCT\_NUM". Include RCS. |
| Automated Switches - RCS | ED\_SWITCH, ED\_SWGAS, ED\_RSA | OT-OTHER,PMH-PADMOUNT HOUSING,RSA-REMOTE SWITCH ACTUATOR,VAC-VACUUM | SWITCH/DISCONNECT/AR | Get the line device number from SAP using the ODRM "CKT\_NRST\_STRCT\_NUM". Include RAR. |
| UG Cable | N/A | N/A | CABLE |  |
| OH Conductor | N/A | N/A | CONDUCTOR/WIRE |  |
| PadMounted Switches | ED\_SWITCH, ED\_SWGAS, ED\_RSA\* | PMH-PADMOUNT HOUSING,PHS-GAS PMH,PMC-PRI METER CABINET,PME-PMH ELBOW CONNECTED, PAD-PAD MOUNTED\* | GAS SWITCH, OIL SWITCH | Get the line device number from SAP using the ODRM "CKT\_NRST\_STRCT\_NUM". Include Padmounted Switches. |

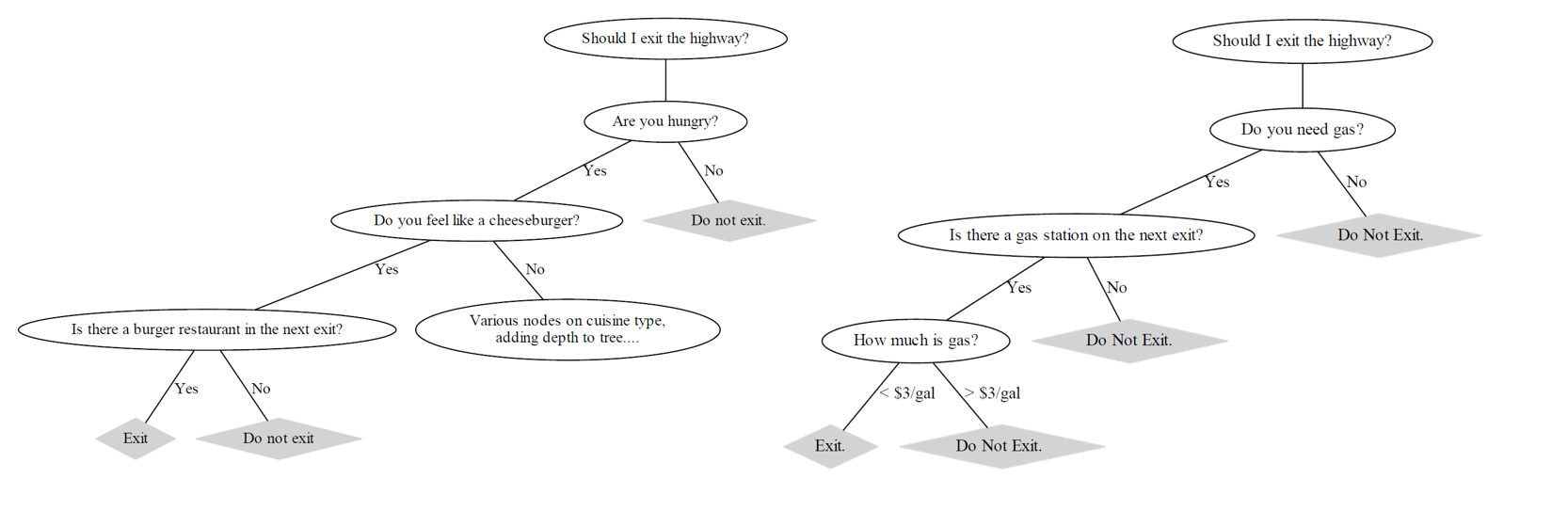
*\*Notes for future changes: PAD-PAD MOUNTED subtypes will be moved to Mainline Oil and Gas Switch model; ED\_RSA will be moved to RCS in future model revisions.*

# **Appendix C: Predictive Machine Learning Algorithms**

Machine learning is the process with which we allow computers to find patterns and relationships in data without these behaviors being explicitly programmed. In essence, it is the utilization of computer algorithms that can learn and make predictions using data. With a basis in statistics, the most basic machine learning models include familiar linear and logistic regressions but have also been extended to the most cutting edge neural networks used by leading tech companies.

**C1. Tree-Based Models – “Ensemble Methods”**

There are many types of machine learning algorithms that can be used to solve different types of problems. To solve predictive maintenance problems, we have relied on assorted models called *tree-based models* or *ensemble methods* because they are relatively simple to use, explain, and understand. They are called *tree-based* because under the hood of these models are things called *decision trees.* We can think of the most basic decision tree as a conditional fork-in-the-road, i.e. if condition A is satisfied, then go down road 1, otherwise go down road 2. This decision point is also called a *node*. Additionally, we can envision more complicated decision trees based on multiple nodes in the decision process. Continuing the road example, say we’re on a highway and we are trying to decide if we want to stop. Of course the basic tree mentioned would merely consist of the question “If I should stop, then take the off-ramp, otherwise, stay on the highway,” but say we wanted to make a more complicated decision. We could ask “Am I hungry?” or “Do I need gas?” Either of these are valid reasons to stop and can each be represented by additional nodes in our ever-growing decision tree, but each would have different conditions to satisfy. Does the next stop have a gas station? Does the next stop have restaurants? As each of these decisions are made, we go further down into the decision tree. Large decision trees with many nodes are sometimes called *deep* or *strong learners*, whereas smaller trees are called *weak learners*. In machine learning, trees are constructed by the computer using data and statistics to create nodes and are utilized in specific algorithms in different ways in order to achieve better accuracy.

*Figure c.1. Two examples of decision trees built around reasons to exit a highway during a road trip. Note that nodes can be made on binary logic as well as to divide continuous values.*

**C2. Bagging and Random Forests**

The road trip exit decision example is a relatively easy one to think about but it falls short in one significant way: it treats the problem like an anecdotal one-off decision. The point of these machine learning models is to use historical observations, that is, a large quantity of previously made decisions, to better inform future decisions. In this respect, it would be better to think that we are creating trees based on a survey of decisions made by a multitude of other road trippers in order to aggregate their decisions into a new decision based on current circumstance. This comes with an increasing amount of difficulties the more complicated we make the decision process. Each traveler was subject to different situations. One might have exited for one reason that is irrelevant to another’s decision, and so on. From this, we can see that patterns based on various conditions would emerge. A singular decision tree would not be able to capture all the various conditions that could appear. Were we to utilize a single decision tree, it would work very well for decisions of the same type, but would never work for other types of observations. This is a problem called *overfitting* and is what happens when a model over-prioritizes the effect of one type of observation over others. In our example, this is like the model saying “only people that are hungry are worth considering when deciding whether or not to exit” and would completely misclassify those running out of gas, or those that need to rest, or any other reason to exit the highway. Instead, let’s create subsets of observations by randomly selecting and copying events from the full set of observations, and then let’s build our trees on these subsets. Note: we are sampling the complete set of observations repeatedly without changing its size (sampling with replacement), we are not taking individual samples out which would decrease the size of the full set (sampling without replacement). This is called *bagging*, a portmanteau of *bootstrap aggregation.*

The building of these different trees would be akin to asking “Can you hold off eating until you arrive at an area with lower gas prices and do you have enough gas to do so?” This construction puts more weight on your hunger than the gas requirement. A similar tree built on the same factors could put more importance on gas: “Do you have enough gas to get to a place with lower gas prices and more food options?” Based on this example we see that you can weigh factors in different ways, each of which change the risk you take when deciding to exit or not. Each of these statements are different decision trees constructed with different subsets of the same data.

It is relatively easy to make these trees from random combinations of variables. Let’s say we were to combinatorically build trees in parallel based on different combinations of factors from various subsets, in varying orders. After we did that, we could create a score of how well each of the different trees performed when making a decision. Furthermore, you could go back and aggregate the decisions made by each tree, while taking into account how well the tree performed individually into the aggregation. This is a sort of toy model of an algorithm called a *Random Forest*, aptly named due to its random construction of trees and using all of them to create a stronger decision than one made with a single tree. Random forests are powerful tools because they, by construction, look at a classification problem from all possible angles and decide which angles are the most important. In data science, we call this ordered hierarchy of variables that affect the outcome the model’s *important features*. In practice, these feature importances can provide valuable insights with which to make business strategy decisions.

**C4. Boosted Decision Trees, Gradient Boosting**

As we build trees, each node essentially splits the data in two, and each subsequent node is intended to increase the accuracy of the model. Let’s say that, during the process of building a tree, we stop and evaluate which nodes performed the worst in previous branches. We can then use this evaluation to *boost* the importance of the data that performed the worst previously. In doing this, we are essentially making up for the model’s previous poor performance of certain subsets of classifications at each node. Another way to look at it is that, when we reprioritize the worst performing subset at each node, we are looking to maximize the performance of the next possible step we take. Mathematically speaking, we are looking for the steepest change at each step we make, a process called *gradient descent*. Because of this, this algorithm is called *gradient boosting*. There are a handful of gradient boosting models whose minutiae handle different performance aspects of the model in different ways.