SPARKPREDICT®
by SPARKCOGNITION

Raise Machine Prognostics to a Cognitive Level

Predictive Analytics Product Overview

SparkCognition leverages its proprietary Automated Model Building solution to generate the most globally optimal model for each and every data set. Automated Model Building is capable of performing multiple analytical techniques, including Anomaly Detection, Clustering, Classification and Regression.

SparkCognition's product, SparkPredict® is able to consume unlabeled data (or data without known failures and states, also known as unsupervised learning) and perform an ensembled, automated clustering technique. After specific clusters are identified, users are able to classify them via the user interface. SparkPredict is also capable of handling labeled data (Data with known failures, also known as supervised learning). It does this by leveraging automated classification and regression algorithms to optimize for the fitness of any new data in a streaming format. In the same way as the clustering algorithms, these classifications can be relabeled and modified within the user interface to retrain the system.

All models generated by SparkPredict are dynamic in nature, meaning they retrain themselves given more data to work with. As part of our user interface, a semi-supervised approach is used to allow for user interactions to help rebuild future models. This model refinement process can occur in either a real-time setting or as a batched process. The automated nature of the clustering and classification is a unique differentiator for SparkCognition and works in four high level steps: (1) Automated data cleansing, (2) Automated feature extraction and selection, (3) Automated model building, (4) Explainable A.I.

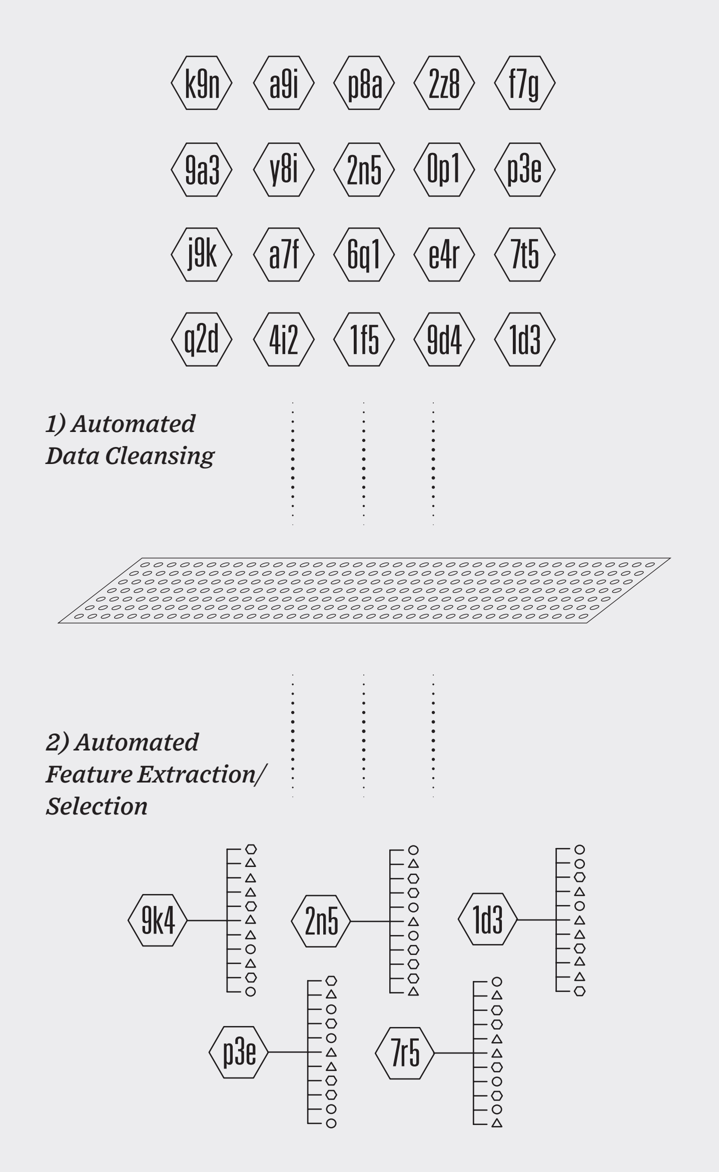
1) Automated Data Cleansing

SparkPredict as part of its data ingestion process runs a set of routines for cleaning and filtering data to address "dirty data" such as bad sensor data. Additionally, user specified limits may be set on a particular data set/variable to ensure data integrity is maintained. The algorithms then used by SparkPredict incorporate a method by which missing data can be interpolated to maintain data quality. Multiple interpolation techniques are used in order to do this dependent upon the quality and type of the data. For example, sparse yet consistently generated variables such as digital tags identifying asset state can be maintained within the data set by carrying the last value forward. Truly missing data can be interpolated using a variety of methods, from more advanced clustering techniques to simple linear fitting. SparkPredict is able to automatically identify the best method for interpolating the data in these instances and implement the appropriate technique. This is a standard methodology of data science and something innately built into the SparkPredict product.

2) Automated Feature Extraction/Selection

SparkPredict leverages a proprietary algorithm, SparkArtemis™, to derive thousands of additional features for every data set. These derived features are first, second, third, and even fourth order linear and nonlinear combinations of the raw input parameters.

SparkPredict utilizes a patented algorithm, SparkPythia™, to perform feature selection. Once hundreds of new features are derived using SparkArtemis, it is requisite to identify the features and patterns most strongly correlated with the problem at hand (e.g. asset deterioration). SparkPythia does this by using a combination of supervised and unsupervised methods to identify exactly which features can be removed from the model building process. Should any weighting or extra consideration of inputs be requisite, the user has the ability to "prioritize" different features during the ingestion period.



This creates a notice for each machine learning algorithm to treat that feature differently and ensure its use during model generation.

3) Automated Model Building

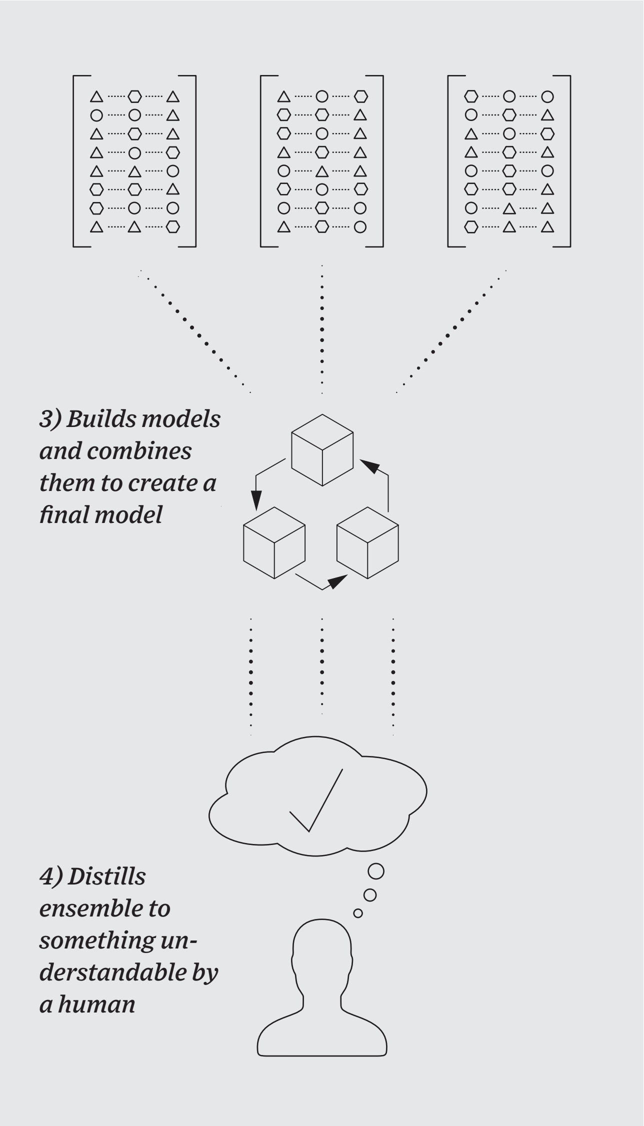
SparkPredict's unique automated model building algorithm tests state of the art tree-based algorithms, different types of regression, deep recurrent neural networks, and more to build initial models. Then, the top performing initial models are combined using genetic algorithms and ensembling techniques to evolve a final model that achieved a global optimal solution.

Outside of this ensembling, it is important to call out a key differentiator used as a part of this approach. SparkPredict uses advanced techniques to "grow" neural networks in such a way that the smallest network is able to adequately solve any problem at hand. It is the model building equivalent of the Bill Gates quote: "I will always choose a lazy person to do a difficult job because a lazy person will find an easy way to do it."

The automated model building solution not only automates the optimization of connection weights, but also the activation functions and even the network topology. This is a paradigm shift from traditional deep learning, which is primarily trained via back-propagation (using a gradient descent optimizer). The result of this method is an optimized network that is not over fit and can be produced with significantly less computational requirements.

4) Explainable A.I.

SparkPredict will not only develop the optimal algorithm for any solution, but it can also distill the complex ensemble back into something that is understandable to a human expert. SparkPredict does this by breaking down the derived features used in the model to identify top contributing original features. This creates the capability to not only identify when a specific event or anomaly might occur but also identify precursor variables that are leading to the event or anomaly. This information can then be displayed as a reinforcer so that a human analyst might agree with the prediction or other algorithmic output.


3) Builds models and combines them to create a final model
4) Distills ensemble to something understandable by a human

Natural Language Processing Capabilities

SparkPredict utilizes proprietary NLP technology to ingest and analyze free form text from a variety of sources including databases, websites, Twitter feeds, product and service manuals, etc. Metrics gathered from analyzing these sources can be directly accessed via "Question and Answer" like interfaces and/or directly processed within machine learning algorithms in order to supplement and augment the results.

SparkPredict's patented algorithms can extract text and pictures from a variety of PDF documents, including those that contain fractured pictures (one pic broken into many) and tables. In addition, SparkPredict utilizes advanced feature extraction from text, term frequency–inverse document frequency (TF-IDF), Vectorization and Part-Of-Speech Tagging (POS Tagging). Using hypothesis validation we can "fact check" axioms and see if different documents provide contradictory evidence allowing us to do complete document similarity analysis. Maintenance personnel are then able to enter fault procedure codes and SparkPredict NLP analysis will provide the appropriate steps to resolve issues in a far more efficient manner.

Technical Capabilities

SparkCognition's product, SparkPredict is built upon the latest openstack technologies, including Hadoop, Hive, Spark, Kafka, and Flume. These technologies allow for high-speed streaming and processing of data because of the linear scalability they offer. SparkCognition has dealt with hundreds of terabytes of sensor data, dozens to hundreds of terabytes of malware data, and gigabytes of individual text corpora. SparkCognition has even trained prediction models using GPU infrastructures in hours. SparkCognition can leverage streaming using Spark clusters, CPU, and GPU setups to hit a variety of performance requirements based on customer's needs.