What is BigQuery ML?

BigQuery ML lets you create and execute machine learning models in BigQuery using standard SQL queries. BigQuery ML democratizes machine learning by letting SQL practitioners build models using existing SQL tools and skills. BigQuery ML increases development speed by eliminating the need to move data.

BigQuery ML functionality is available by using:

- The Google Cloud console
- The `bq` command-line tool
- The BigQuery REST API
- An external tool such as a Jupyter notebook or business intelligence platform

Machine learning on large datasets requires extensive programming and knowledge of ML frameworks. These requirements restrict solution development to a very small set of people within each company, and they exclude data analysts who understand the data but have limited machine learning knowledge and programming expertise.

BigQuery ML empowers data analysts to use machine learning through existing SQL tools and skills. Analysts can use BigQuery ML to build and evaluate ML models in BigQuery. Analysts don't need to export small amounts of data to spreadsheets or other applications or wait for limited resources from a data science team.

Supported models in BigQuery ML

A model in BigQuery ML represents what an ML system has learned from the training data.

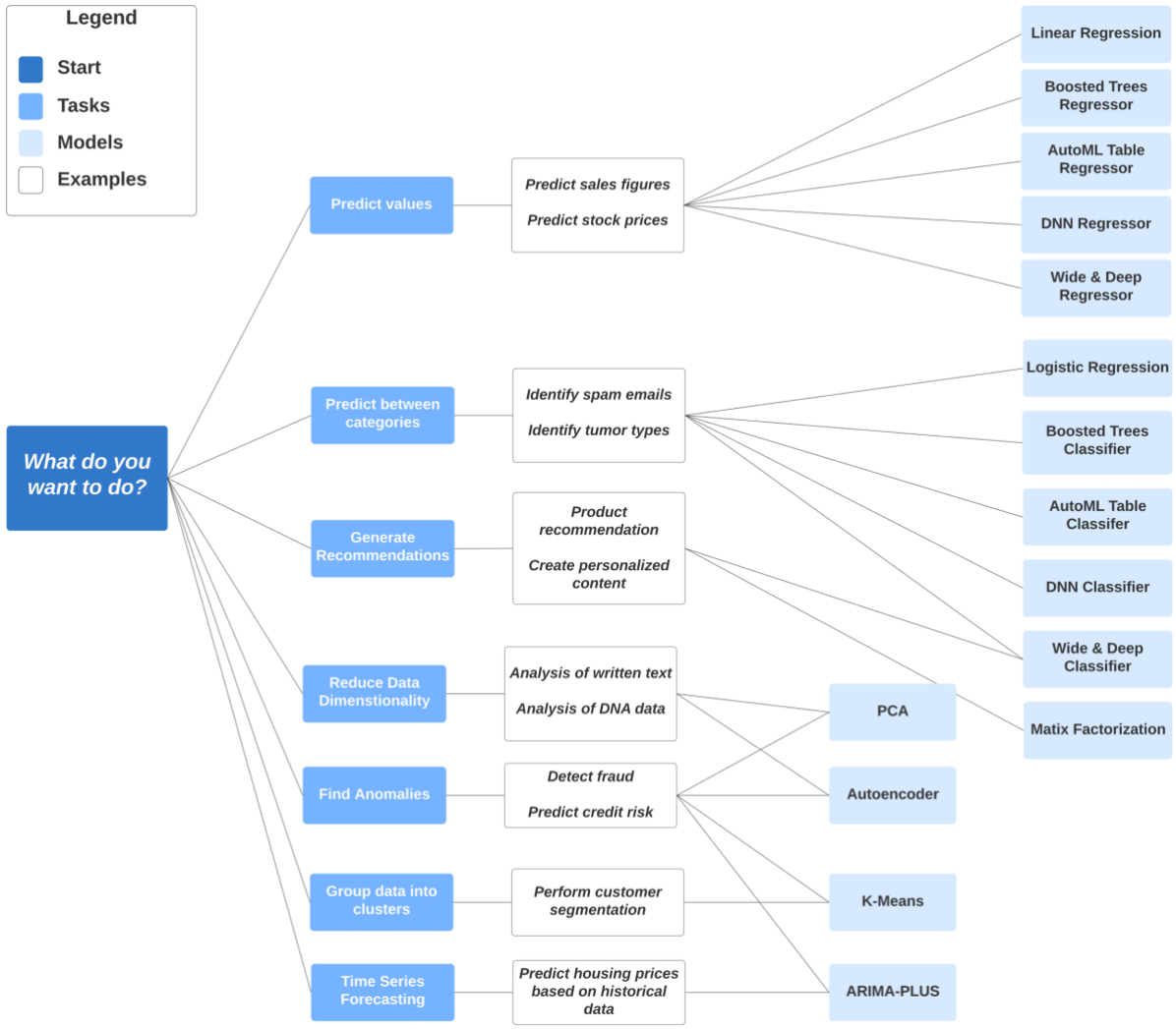
BigQuery ML supports the following types of models:

- Linear regression for forecasting; for example, the sales of an item on a given day. Labels are real-valued (they cannot be +/- infinity or NaN).
- Binary logistic regression for classification; for example, determining whether a customer will make a purchase. Labels must only have two possible values.
- Multiclass logistic regression for classification. These models can be used to predict multiple possible values such as whether an input is "low-value," "medium-value," or "high-value." Labels can have up to 50 unique values. In BigQuery ML, multiclass logistic regression training uses a multinomial classifier with a cross-entropy loss function.
- K-means clustering for data segmentation; for example, identifying customer segments. K-means is an unsupervised learning technique, so model training does not require labels nor split data for training or evaluation.

- Matrix Factorization for creating product recommendation systems. You can create product recommendations using historical customer behavior, transactions, and product ratings and then use those recommendations for personalized customer experiences.
- Time series for performing time-series forecasts. You can use this feature to create millions of time series models and use them for forecasting. The model automatically handles anomalies, seasonality, and holidays.
- Boosted Tree for creating XGBoost based classification and regression models.
- Deep Neural Network (DNN) for creating TensorFlow-based Deep Neural Networks for classification and regression models.
- AutoML Tables to create best-in-class models without feature engineering or model selection. AutoML Tables searches through a variety of model architectures to decide the best model.
- TensorFlow model importing. This feature lets you create BigQuery ML models from previously trained TensorFlow models, then perform prediction in BigQuery ML.
- Autoencoder for creating Tensorflow-based BigQuery ML models with the support of sparse data representations. The models can be used in BigQuery ML for tasks such as unsupervised anomaly detection and non-linear dimensionality reduction.

In BigQuery ML, you can use a model with data from multiple BigQuery datasets for training and for prediction.

Model selection guide



Advantages of BigQuery ML

BigQuery ML has the following advantages over other approaches to using ML with a cloud-based data warehouse:

- BigQuery ML democratizes the use of ML by empowering data analysts, the primary data warehouse users, to build and run models using existing business intelligence tools and spreadsheets. Predictive analytics can guide business decision-making across the organization.
- There is no need to program an ML solution using Python or Java. Models are trained and accessed in BigQuery using SQL—a language data analysts know.
- BigQuery ML increases the speed of model development and innovation by removing the need to export data from the data warehouse. Instead, BigQuery ML brings ML to the data. The need to export and reformat data has the following disadvantages:

- Increases complexity because multiple tools are required.
- Reduces speed because moving and formatting large amounts data for Python-based ML frameworks takes longer than model training in BigQuery.
- Requires multiple steps to export data from the warehouse, restricting the ability to experiment on your data.
- Can be prevented by legal restrictions such as HIPAA guidelines.