Study Guide: Developing ML Models with BigQuery ML

Subject: Google Cloud Professional Machine Learning Engineer Exam **Topics:**

- BigQuery ML Model Selection
- Feature Engineering & The TRANSFORM Clause
- Model Evaluation & Interpretability
- Time-Series Analysis (ARIMA_PLUS)
- Unsupervised Learning & Dimensionality Reduction

Summary

This study guide focuses on the practical application of BigQuery ML (BQML) to solve real-world business problems directly within the data warehouse. It covers the selection of appropriate model types (Regression, Classification, Recommendations, and Anomaly Detection), techniques for robust feature engineering using the `TRANSFORM` clause, and the use of specialized BQML functions for evaluation and model explanation.

Key Concepts

- **Model Selection Criteria:** Choosing the right algorithm based on the business objective. Use `LOGISTIC_REG` for interpretable binary outcomes, `MATRIX_FACTORIZATION` for user-item recommendations, and `BOOSTED_TREE_CLASSIFIER` for complex, non-linear relationships.
- **The TRANSFORM Clause:** A critical architectural tool in BQML that allows you to define preprocessing logic (like scaling or encoding) once. This logic is then "packaged" with the model, ensuring that the same transformations are applied automatically during prediction, effectively eliminating **training-serving skew**.
- **Model Interpretability:** Utilizing specialized functions like `ML.EXPLAIN_PREDICT` to move beyond black-box predictions. This helps stakeholders understand which specific features (via Shapley values) influenced a particular outcome.
- **In-Warehouse Advantages:** BQML provides a significant advantage over custom training for tabular data by removing the need for data egress, utilizing standard SQL for faster prototyping, and eliminating the overhead of managing separate infrastructure.
- **Time-Series Forecasting:** Leveraging the `ARIMA_PLUS` model type to automatically handle complex time-series components like trend, seasonality, and holiday effects without manual intervention.

Vocabulary List

- **LOGISTIC_REG:** The primary model type for binary classification where interpretability (viewing feature weights) is a priority.
- **Training-Serving Skew:** A common ML failure where the data distribution or preprocessing logic differs between the training phase and the production inference phase.
- **MATRIX_FACTORIZATION:** The BQML algorithm used to build recommendation engines through collaborative filtering.
- **R-squared (R^2):** A regression metric representing the proportion of variance in the dependent variable explained by the model.
- **ML.ARIMA_EVALUATE:** A specialized function to view diagnostics and metrics (like AIC or seasonality components) for time-series models.
- **ML.PREDICT vs. ML.EXPLAIN_PREDICT:** `ML.PREDICT` provides the raw output; `ML.EXPLAIN_PREDICT` provides the output plus the top feature contributions (Shapley values).
- **L2 Regularization:** A technique to prevent overfitting by penalizing the square of the magnitude of the model weights.
- **AUTOENCODER:** An unsupervised model used primarily for anomaly detection in unstructured or complex data.
- **PCA (Principal Component Analysis):** An algorithm used for dimensionality reduction by transforming features into orthogonal components.
- **ML.STANDARD_SCALER:** A preprocessing function that standardizes features using the formula: $z = \frac{x-\mu}{\sigma}$.

Key Questions

1. **Which BQML clause is the best defense against training-serving skew, and why?**
2. **When should you choose a `BOOSTED_TREE_CLASSIFIER` over a standard `LOGISTIC_REG`?**
3. **If you need to explain to a client why a specific loan was denied, which BQML function would you use?**
4. **What are the four primary components identified by an `ARIMA_PLUS` time-series model?**
5. **What is the formula for calculating 'Precision' and how does it differ from 'Recall' in an evaluation context?**