2. Quality and Risk Analysis

This document details the primary risks to the Microenterprise Density Prediction System and the strategies to manage them, ensuring the system remains reliable, accurate, and secure.

Risk 1: Data Integrity and Availability Failures

Description:
The system's predictions are only as good as the data it uses. Key failures include:

- **Incorrect or Missing Data:** Source data from providers like Kaggle could have errors, wrong formats, or large gaps, leading to flawed model training and unreliable predictions.
- **Source Disappearance:** A critical data source could become unavailable or change its structure without warning, breaking the data ingestion process.

Mitigation Strategies:

- **Robust Data Processing:** Implement the "robust interpolation" and "anomaly detection" (like Isolation Forest) mentioned in the report (Table A.2) to automatically handle missing values and strange data points.
- **Data Validation:** Use "Data Contracts" to check that incoming data matches the expected format (column names, types) before processing.
- **Backup Sources:** Identify and integrate alternative data sources for key variables to use if a primary source fails.

Monitoring and Response:

- **Monitoring:** A dashboard will show a "Data Quality Score" for each new data batch. Alerts will trigger for failed data jobs or low-quality scores.
- **Response:** The operations team will be alerted to switch to a backup data source or investigate and fix the data pipeline.

Risk 2: Model Performance Decay (Model Drift)

Description:
The real world changes, so a model trained on old data becomes less accurate over time. This is called "model drift." For example, a post-pandemic economy behaves differently, making old patterns obsolete.

Mitigation Strategies:

- **Adaptive Retraining Loop:** Use the system's "Adaptive Feedback Loop" (Section 3.5, Figure B.2). This system automatically detects when the model's performance is dropping ("statistical drift") and triggers a retraining process with new data.
- **Scheduled Updates:** As stated in Table A.1, perform "automated monthly batch updates" to regularly refresh the model even if no drift is detected.
- **Ensemble Models:** The use of a hybrid model (ARIMA, XGBoost, LSTM) makes the system naturally more stable and resistant to minor data changes.

Monitoring and Response:

- **Monitoring:** Continuously track key performance metrics (RMSE, MAE) using MLflow. Set up automatic drift detection to monitor for changes in data patterns.
- **Response:** If drift is detected, the system automatically queues a new model training job. The data science team reviews the new model's performance before replacing the old one, ensuring a safe update.

Risk 3: Security Breaches and Ethical Concerns

Description:
Although the system uses open data, it must be protected from unauthorized access and ensure its predictions are fair.

- **Security Breach:** Hackers could try to steal data, manipulate the system's predictions, or take the service offline.
- **Model Bias:** If the training data is biased against certain regions or groups, the model's predictions could be unfair and lead to poor policy decisions.

Mitigation Strategies:

- **Secure Access:** Implement "encrypted authentication protocols" (Table A.1) and role-based access to ensure only authorized users can access data and systems.
- **Bias Checking:** Conduct pre-training audits of data for fair representation and post-training tests to ensure the model's error rates are similar across different regions (e.g., urban vs. rural).
- **Ethical Governance:** Follow "Ethical and Data Governance Considerations" (Appendix B.5), using anonymized data and avoiding the collection of unnecessary personal information.

Monitoring and Response:

- **Monitoring:** Use security tools to monitor for unusual activity, like many failed login attempts or a sudden surge in data downloads.

- **Response:** If a security threat is detected, immediately block the source and investigate. If model bias is suspected, retrain the model with techniques to correct the unfairness.