Test Data Directory

This directory contains sample datasets for testing and validating the SEC 8-K Predictor system.

Files Description

📊 sample_predictions.csv

- **Purpose**: Sample prediction results with features and outcomes
- **Records**: 100 synthetic prediction examples
- **Columns**:
- `ticker`: Stock symbol
- `filing_date`: Date of SEC filing
- `category`: SEC 8-K item category
- `content_summary`: Brief description of filing content
- `sentiment_score`: LLM-extracted sentiment (0-1)
- `urgency_score`: LLM-extracted urgency (0-1)
- `financial_impact_score`: LLM-extracted financial impact (0-1)
- `predicted_direction`: Model prediction (positive/negative)
- `predicted_probability`: Confidence in direction prediction
- `predicted_return_5d`: Predicted 5-day return
- `predicted_return_9d`: Predicted 9-day return
- `actual_return_5d`: Simulated actual 5-day return
- `actual_return_9d`: Simulated actual 9-day return
- `prediction_accuracy`: Binary accuracy indicator
- `model_confidence`: Overall model confidence

📈 model_performance_metrics.csv

- **Purpose**: Model evaluation metrics by category and type
- **Metrics**: Accuracy, precision, recall, F1-score, R^2, MAE, RMSE
- **Categories**: Different SEC 8-K item categories (2.01, 2.02, 2.03, 8.01)
- **Model Types**: Classification and regression models

🎯 feature_importance.csv

- **Purpose**: Feature importance rankings for different models

- **Features**: Sentiment scores, TF-IDF features, financial impact scores
- **Rankings**: Ordered by importance score within each model/category

📊 backtesting_results.csv

- **Purpose**: Historical performance of trading strategies
- **Strategies**:
- Buy positive predictions
- Buy high confidence predictions
- Market benchmark
- **Metrics**: Total return, Sharpe ratio, max drawdown, win rate
- **Periods**: Quarterly and full-year results for 2023

🎯 prediction_confidence_analysis.csv

- **Purpose**: Analysis of prediction accuracy by confidence level
- **Buckets**: Very high, high, medium, low, very low confidence
- **Metrics**: Accuracy rate, average returns, success rate

📋 category_performance.csv

- **Purpose**: Performance analysis by SEC 8-K category
- **Categories**: All major SEC 8-K item categories
- **Metrics**: Filing count, accuracy, returns, best model type

Usage

These datasets can be used for:

1. **Testing Prediction Pipeline**: Validate end-to-end prediction functionality
2. **Model Evaluation**: Assess model performance across different scenarios
3. **Visualization Development**: Create charts and dashboards
4. **Backtesting Validation**: Test trading strategy implementations
5. **Feature Analysis**: Understand feature importance and relationships

Data Generation

The test data is generated using realistic statistical distributions and correlations to simulate actual SEC 8-K prediction scenarios. The data includes:

- **Realistic Score Distributions**: Sentiment, urgency, and financial impact scores follow normal distributions
- **Correlated Predictions**: Returns are correlated with feature scores
- **Market Noise**: Random variations to simulate real market conditions
- **Category Variations**: Different performance characteristics by filing category

Integration with Tests

This data is automatically used by the test suite in `tests/test_predictions.py` to validate:

- Prediction pipeline functionality
- Data format consistency
- Model output validation
- Performance metric calculations

Note

This is synthetic test data for development and testing purposes. For production use, replace with actual historical SEC filing and market data.