Preprocessing Pipeline Summary Model-Ready Dataset

Date: 2025-03-31

Core Pipeline Outputs

- Cleaned, flattened runner-level dataset created via flatten_day_batch_pkl.py
- Dtype optimization: category, float64, int64, datetime64[ns], and bool applied
- Categorical fields label-encoded into *_idx columns for model ingestion
- comment and spotlight embedded using SentenceTransformer
- 34 domain-specific binary NLP flags extracted via regex
- All engineered race-level features retained (z-scores, ranks, etc.)

Data Validation and Handling

- GoingStick dropped (0 populated); has_GoingStick retained as jurisdictional flag
- Trainer/jockey stats verified: missing values only occur when runs == 0
- form field retained (held out of model pipeline), fuzzy without class context
- No invalid missing values found in trainer/jockey stats
- Embeddings verified: comment and spotlight vectorized as np.array per row

Saved Artifacts

- Embedding indices saved to: data/processed/embedding_encoders.pkl
- Text embeddings saved to .npz with versioned filename
- Model-ready DataFrame saved to .pkl and .csv with date-based naming

Data Types Summary

- Total columns: 102
- category (8): used for embedding fields like country, going, type
- float64 (26): all numerical, including ratings and trainer/jockey data
- int64 (54): binary flags, embedding indices, and numeric fields
- datetime64[ns] (1): race_datetime parsed and usable
- bool (1): non_runner_flag clean binary signal
- object (12): text fields, vectors, IDs, and held-out form string

Next Steps

- Optional: Parse last_run and off_time to datetime for temporal modeling
- Prepare batching logic for transformer ingestion
- Test model-ready file via data loader / batch constructor
- Begin transformer model experimentation