Superhero Public Destruction Prediction

Technical Assessment Report for King Price Insurance

Armand de Wet

February 15, 2026

Contents

1 Introduction

1.1 Task Background

This report details the development of a predictive engine designed to estimate *Annual Public Destruction Events* for prospective superhero clients. The objective was to transform unstructured bureau data into a high-fidelity risk rating engine to assist the Ratings Team in automated underwriting.

1.2 Methodological Approach

The project followed a production-oriented data science lifecycle:

- **Data Ingestion**: SQL-based extraction from a DuckDB analytical database.
- **Feature Engineering**: Robust XML parsing (lxml) to handle schema inconsistencies and feature derivation (Age, Power Encoding).
- **Modeling**: Benchmarking across linear, non-linear, and ensemble learners with 5-fold cross-validation.
- **System Architecture**: Deployment of a modular FastAPI backend and a Streamlit-based "Broker Dashboard."

1.3 Executive Summary of Results

The final solution utilizes a **Random Forest Regressor**, which achieved a Mean RMSE of **5.2311**. This model was chosen for its robust handling of non-linear interactions between fiscal stability (Credit Score) and physical risk factors.

2 Exploratory Data Analysis & Engineering

2.1 Robust XML Parsing (The "Kratos" Case)

The dataset contained a malformed and truncated XML record (notably the "Kratos" record). Rather than discarding these observations, a robust `BeautifulSoup` parser with an `lxml-xml` engine was implemented. This allowed for partial feature extraction and ensured maximum data utility—a critical requirement for high-value insurance leads.

2.2 Feature Engineering Strategy

To transform the raw bureau data into a format suitable for machine learning, several high-impact engineering steps were performed:

- **Temporal Normalization**: Date of Birth values were parsed and converted into a numerical *Age* feature, calculated relative to the reference date of 2026-02-14 to ensure consistency across the batch.

- **Multi-Label Categorical Encoding**: Because superheroes often possess multiple powers, categorical attributes were transformed using the `MultiLabelBinarizer()`. This approach effectively one-hot encoded the set of powers for each hero, allowing the model to weigh the risk of specific power combinations (e.g., Teleportation + Combat) individually. This was also done for the bank predictor, but it was seen that it had a minimal correlation with the target variable.
- **Data Integrity**: A comprehensive data quality check confirmed that no critical predictor values, including *Age*, were missing. Consequently, no imputation was required, preserving the original distribution of the superhero population.

The relationship between these engineered features and the target variable is illustrated in Figure 1.

2.3 Feature Selection

Feature importance was validated using the Correlation Matrix shown in Figure 2. The analysis revealed that *Credit Score*, *Property Count*, *Age*, and the presence of *Teleportation* powers were the most significant predictors of public destruction.

3 Modeling, Selection, and Optimization

3.1 Model Benchmarking

Seven algorithms were evaluated using 5-fold cross-validation to ensure the model generalizes well to unseen applicants.

Model	Mean CV RMSE
Random Forest Regressor	**5.625056**
Ridge Regression	5.829180
Linear Regression	5.835497
Bayesian Linear	5.859369
Support Vector Regression (SVR)	5.926272
XGBoost Regressor	6.146896
Lasso Regression	6.213077

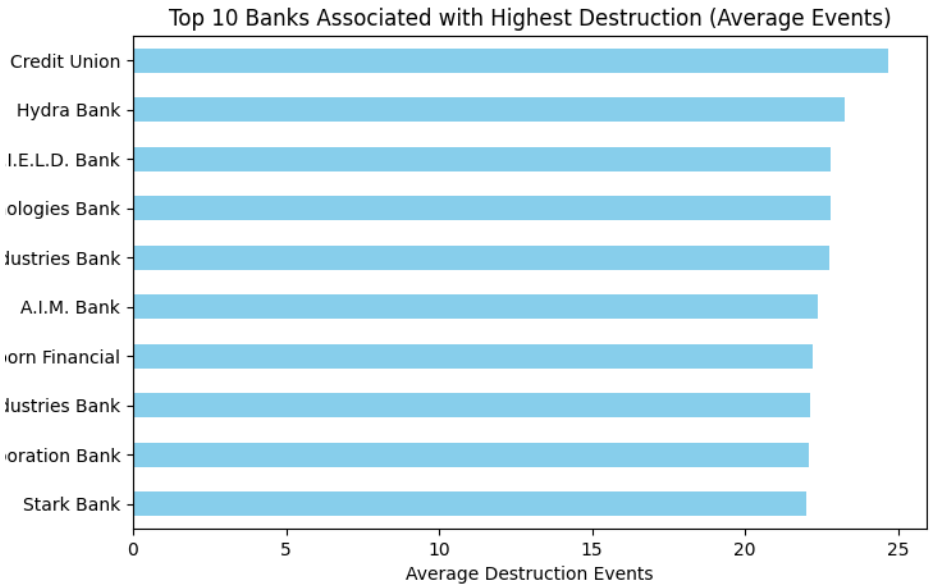
Table 1: Benchmarking results ordered by performance (lower RMSE is better).

3.2 Final Model: Optimized Random Forest

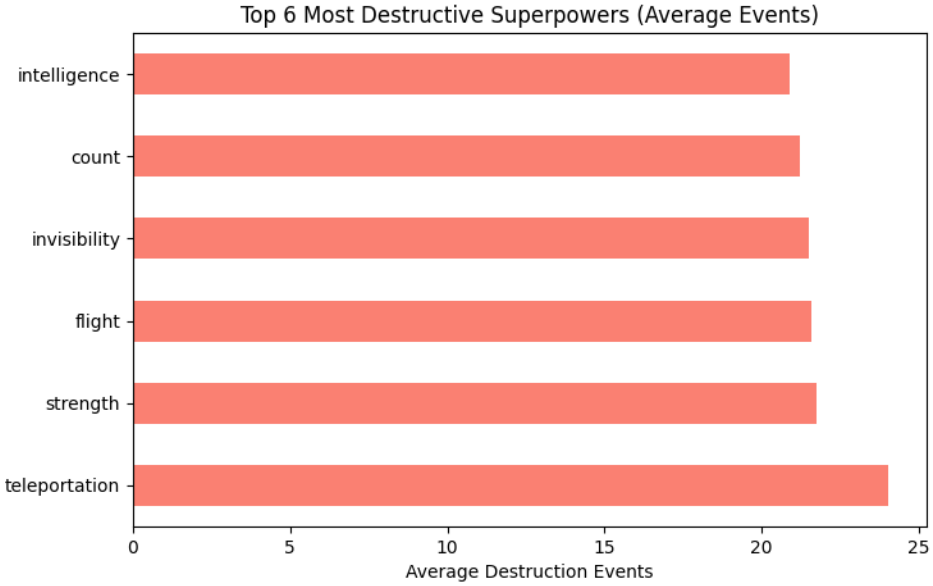
The Random Forest model was selected for its superior resilience to outliers and non-linear data structures. Post-tuning via `GridSearchCV`, the model achieved a finalized **RMSE of 5.2311**.

Optimal Hyperparameters:

- `n_estimators`: 50 — `min_samples_split`: 10
- `max_features`: *sqrt* — `max_depth`: *None*



(a) Top 10 Banks Associated with Highest Destruction (Average Events)



(b) Top 6 Most Destructive Superpowers (Average Events)

Figure 1: Bivariate analysis of average destruction events by Bank and Power.

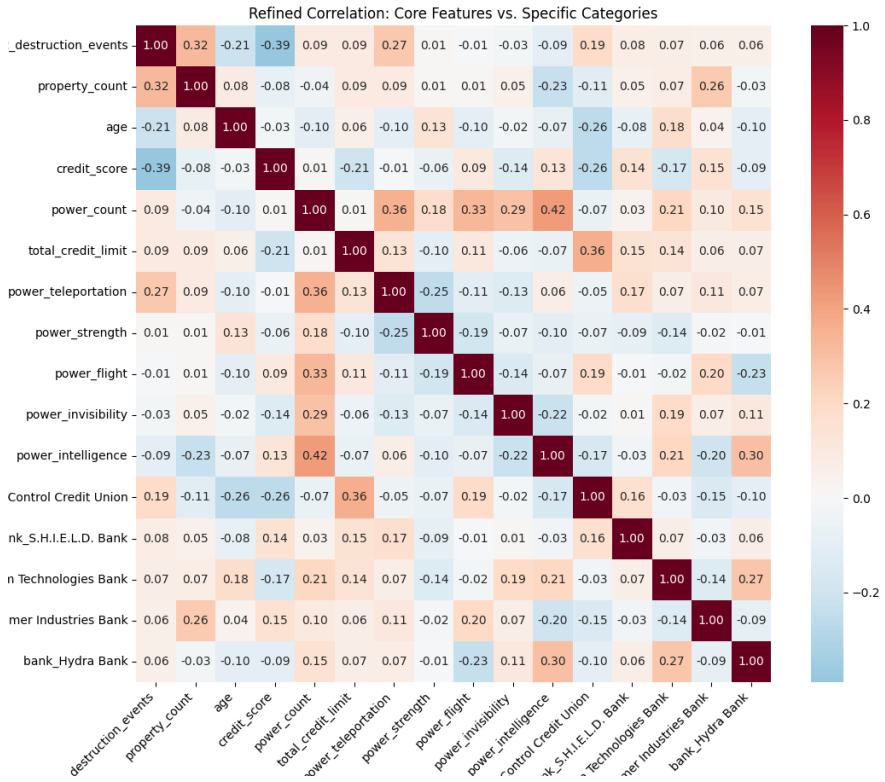


Figure 2: Correlation Matrix of Superhero Attributes and Destruction Events

4 Deployment Architecture

4.1 Modular API Design (FastAPI)

The backend was refactored into a modular, class-based architecture to demonstrate enterprise-grade coding standards. By encapsulating the model and logic within a `SuperheroProcessor` class, the system achieves a clear separation of concerns from the `APIRouter`. This design ensures that the data pipeline is not only scalable but also highly maintainable and ready for integration into a continuous deployment (CI/CD) pipeline.

4.2 Interactive Risk Dashboard (Streamlit)

A Streamlit-based UI was implemented to demonstrate a real-world "Broker Dashboard" application. This interface allows stakeholders to perform manual risk assessments and observe the sensitivity of the model to specific inputs, such as *Credit Score* or *Teleportation* status. By importing the core processing class directly from the API package, the dashboard maintains perfect logic parity with the production endpoint, eliminating the risk of "logic drift."

5 Conclusion

The resulting predictive engine successfully navigates the complexities of unstructured bureau data to provide a robust, interpretable risk assessment solution. Through it's

rigorous and modular software design, the solution moves beyond a simple analytical script into a production-ready system capable of assisting the Ratings Team in identifying high-risk applicants with precision.

5.1 Acknowledgments

I would like to express my sincere gratitude to the King Price AI Team for the opportunity to engage with this technically stimulating and creative task. The process of developing a solution that balances data integrity with real-world deployment challenges was highly rewarding.

5.2 Repository Access

The complete source code, including the modular API, modeling notebooks, the interactive dashboard, report and results, is available for review:
`https://github.com/Armand-deWet/king_price_project`

A compressed archive containing all of the above with the required .env file will also be provided.