Advanced Lead Intelligence Platform - Project Report

Overview

This project implements an intelligent lead generation and scoring system using machine learning techniques to identify, score, and segment potential business leads from multiple data sources.

Approach & Architecture

Data Sources Integration

- **LinkedIn Companies**: Simulated company data with employee counts, follower metrics, and industry information
- **Google Search**: Web scraping simulation for company discovery with SEO metrics
- **Crunchbase Startups**: Funding and growth stage information for emerging companies

Machine Learning Models

1. Lead Quality Prediction (RandomForestClassifier)

- **Model**: Scikit-learn RandomForestClassifier with 100 estimators
- **Features**: Company size, industry encoding, social media engagement, contact completeness
- **Target**: Multi-factor quality score based on company size, engagement metrics, and lead scores
- **Preprocessing**: StandardScaler for feature normalization, LabelEncoder for categorical variables

2. Lead Segmentation (K-Means Clustering)

- **Model**: K-Means clustering with 4 segments
- **Features**: Normalized lead scores, company size, engagement metrics, growth potential
- **Segments**: High Value, Growth Potential, Standard, Low Priority
- **Evaluation**: Silhouette score for cluster quality assessment

3. Fuzzy Deduplication (Sentence Transformers + Cosine Similarity)

- **Primary**: SentenceTransformer 'all-MiniLM-L6-v2' for semantic embeddings
- **Fallback**: TF-IDF vectorization when transformer unavailable
- **Similarity Threshold**: 0.8 for duplicate detection
- **Features**: Company name, domain, and location signatures

Data Processing Pipeline

1. **Data Collection**: Multi-source lead generation with configurable filters
2. **Feature Engineering**: Numerical encoding, log transformations, text length metrics
3. **Quality Scoring**: AI-driven lead quality prediction with feature importance analysis
4. **Deduplication**: Semantic similarity-based duplicate removal
5. **Segmentation**: Automated lead categorization for targeted marketing
6. **Contact Enrichment**: Email pattern generation and data completeness scoring

Performance Evaluation

- **Classification Accuracy**: Measured using train-test split (80/20)
- **Clustering Quality**: Silhouette score for segment validation
- **Deduplication Effectiveness**: Similarity threshold optimization
- **Feature Importance**: Random Forest feature ranking for interpretability

Technical Implementation

- **Framework**: Streamlit for interactive web interface
- **ML Libraries**: Scikit-learn, SentenceTransformers, Pandas, NumPy
- **Visualization**: Plotly for interactive charts and dashboards
- **Export Options**: CSV, JSON, and Excel formats with multi-sheet support

Key Features

- Real-time lead scoring with explainable AI
- Interactive filtering and segmentation
- Automated contact information enrichment
- Comprehensive analytics dashboard with conversion metrics
- Smart deduplication preventing data redundancy

Results & Insights

The system successfully demonstrates enterprise-level lead intelligence capabilities with:

- Scalable multi-source data integration
- Accurate lead quality prediction with interpretable features
- Effective customer segmentation for targeted campaigns
- Robust duplicate detection maintaining data quality

This implementation provides a foundation for automated lead generation workflows while maintaining flexibility for various business use cases and data sources.