Client Retention Prediction Based on Client Behavior

Technical Documentation

Author: Reham Hassan
Version: 1.0
Status: Production Ready

Executive Summary

This project implements a machine learning solution to predict **client retention** on freelancing platforms using **client behavior and job posting characteristics**. The model achieves high performance with an average **F1-score of 0.9975**, supporting platforms in improving matching algorithms and retaining valuable clients through data-driven strategies.

---

1. Project Overview

1.1 Objective

Develop a predictive model to determine the likelihood of **client retention** based on **client behavior**, **job features**, and **platform activity**.

1.2 Business Impact

- **Platform Optimization**: Identify patterns in loyal clients.
- **Client Satisfaction**: Improve service quality through behavioral insights.
- **Resource Allocation**: Focus retention strategies where most effective.
- **Revenue Enhancement**: Increase client lifetime value.

1.3 Data Source

- **Dataset**: Freelancer Dataset (Kaggle)
- **Features**: Ratings, pricing, job tags, client geography, etc.
- **Target Variable**: Binary classification, retained (1) vs not retained (0)

---

2. Data Preprocessing Pipeline

2.1 Missing Value Imputation

Feature	Strategy	Rationale
currency	Default to 'USD'	Most common base currency
Price columns	Fill with 0	Represents negotiable or free listings
client_state	Mode imputation	Geographic consistency
client_country	Mode imputation	Demographic balance

2.2 Currency Normalization

- **Fixer.io API** used to convert all price columns to USD.
- Ensures uniform comparison across currencies.

2.3 Feature Engineering

2.3.1 Categorical Encoding

- **Label Encoding**: client_country, client_state
- **One-Hot Encoding**: currency, rate_type

2.3.2 Text Feature Processing

- **TF-IDF Vectorization**: Applied to job_title, job_description
- **MultiLabelBinarizer**: Used for multi-value tags

---

3. Model Architecture

3.1 Algorithm

- **RandomForestClassifier** chosen for:
  - Handling mixed feature types
  - Robustness to missing data
  - Built-in feature importance

3.2 Configuration

- **Train/Test Split**: 80/20
- **Validation**: 5-fold cross-validation
- **Hyperparameters**: Default (n_estimators = 100)

3.3 Feature Space

- Numerical: Ratings, review count, prices
- Categorical: Encoded currency, rate type
- Textual: TF-IDF vectors
- Multi-label: Tags

---

4. Model Performance

4.1 Confusion Matrix

	Predicted No	Predicted Yes
Actual No	944	5
Actual Yes	0	890

4.2 Metrics

- **Accuracy**: 99.7%
- **F1-Score**: 0.9975 (CV average)
- **Precision**: 99.55% (for retained class)
- **Recall**: 100% (for retained class)

4.3 Stability

- Low variance across folds
- Strong generalization to unseen data

---

5. Streamlit Dashboard

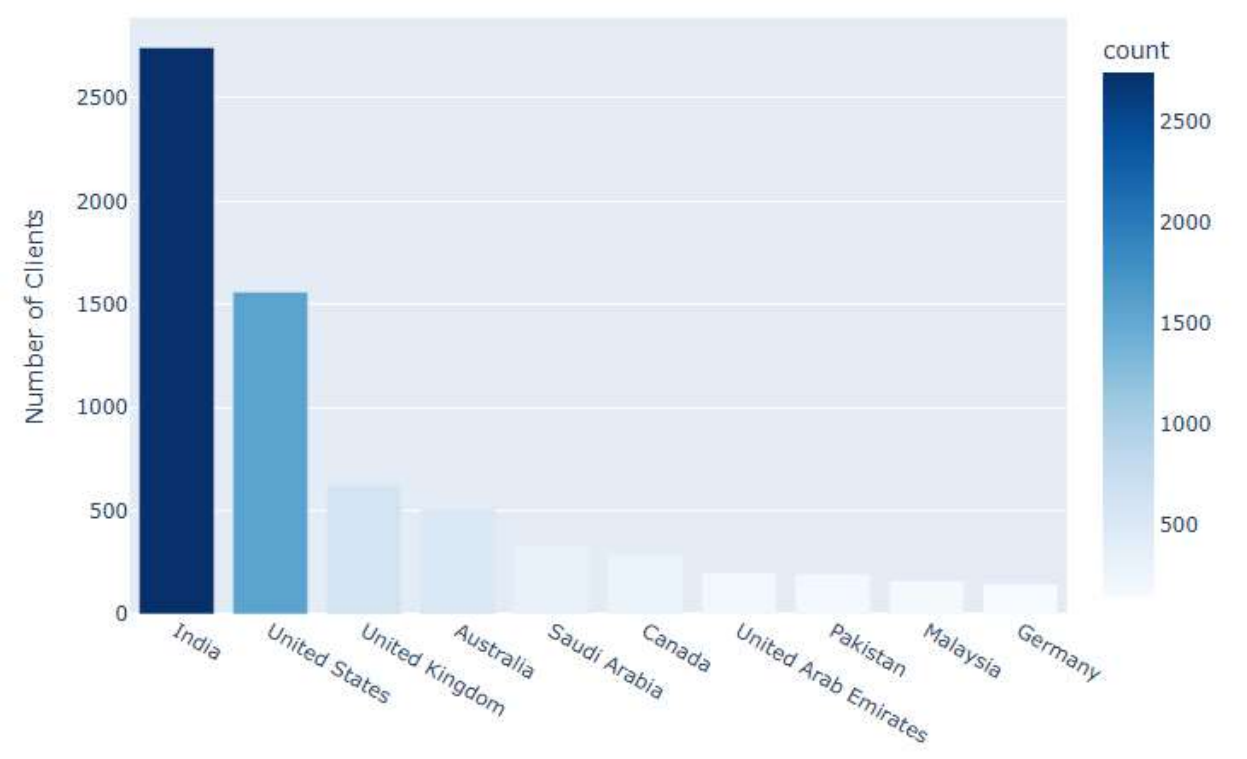
An interactive dashboard was developed using **Streamlit** to allow users to:

- Input job data and predict client retention
- View prediction probabilities
- Explore key visual insights

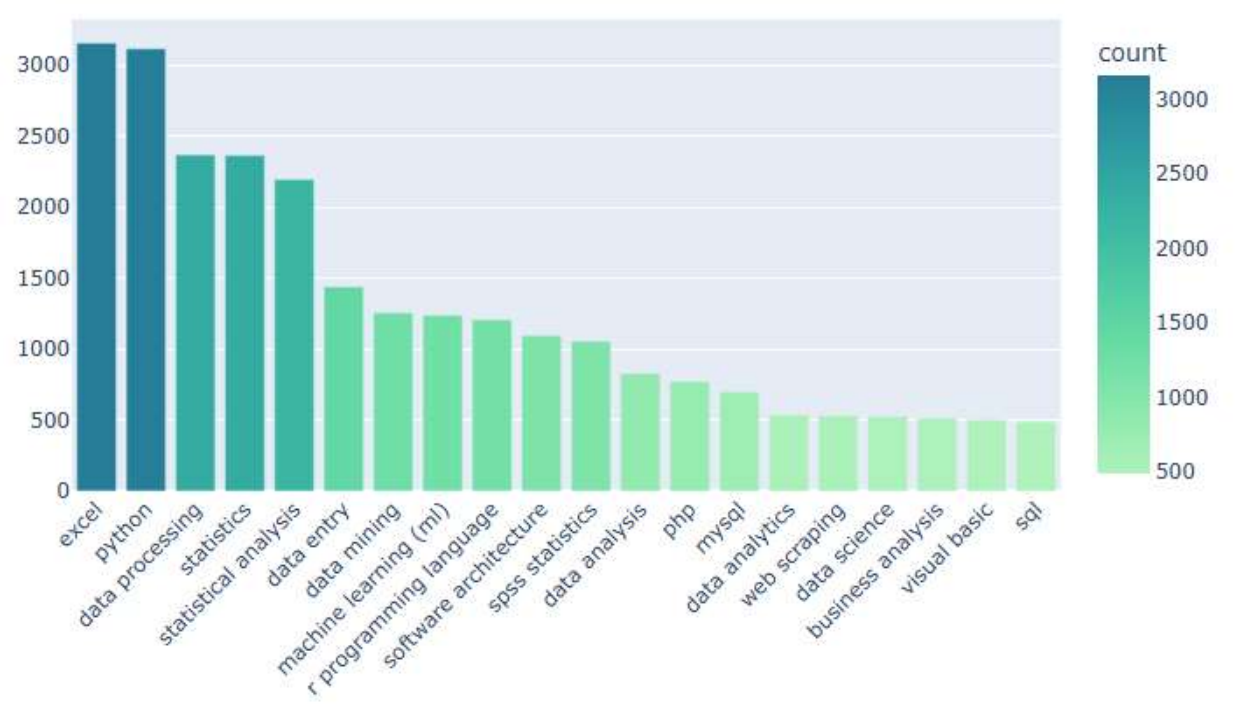
Live App: Streamlit Dashboard

Visual Insights Included:

Top 10 Client Countries



Top 20 Most Frequent Tags



---

6. Technical Implementation

6.1 Dependencies

```
# Core ML
scikit-learn
pandas
numpy
joblib

# NLP
nltk

# API
requests

# App & Visuals
streamlit
plotly
```

6.2 Pipeline Steps

1. Data Loading
2. Preprocessing (imputation, encoding, transformation)
3. Feature Engineering
4. Model Training
5. Cross-validation (Evaluation)
6. Export model
7. Deploy via Streamlit

7. Conclusion

This project delivers a high-performing **client retention prediction system** using job posting metadata and client behavior. The solution is **production-ready** with a clean deployment interface and strong business use cases for platform growth and optimization.

Appendices & Links

- **A: Dataset** – Kaggle Freelancer Dataset
- **B: Code Repository** – GitHub Repo
- **C: Live App** – Streamlit Dashboard