End-to-End Big Data & AI Platform

Real-Time & Batch Analytics for Customer Churn Prediction

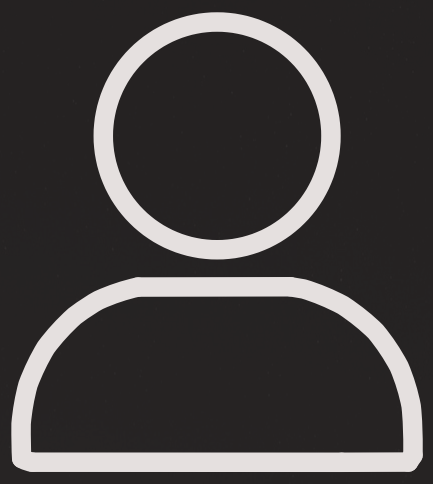
Supervision of : ENG / Ahmed Azab

Meet the Team

Our project is driven by a dedicated and talented team, bringing diverse expertise to deliver an end-to-end Big Data & AI platform.

Abdullah Ibrahim Mahmoud

Team Leader



Mansour Mohamed Mansour

Ezzeldeen Elsayed Mohammed

Ahmed Mohamed Ahmed

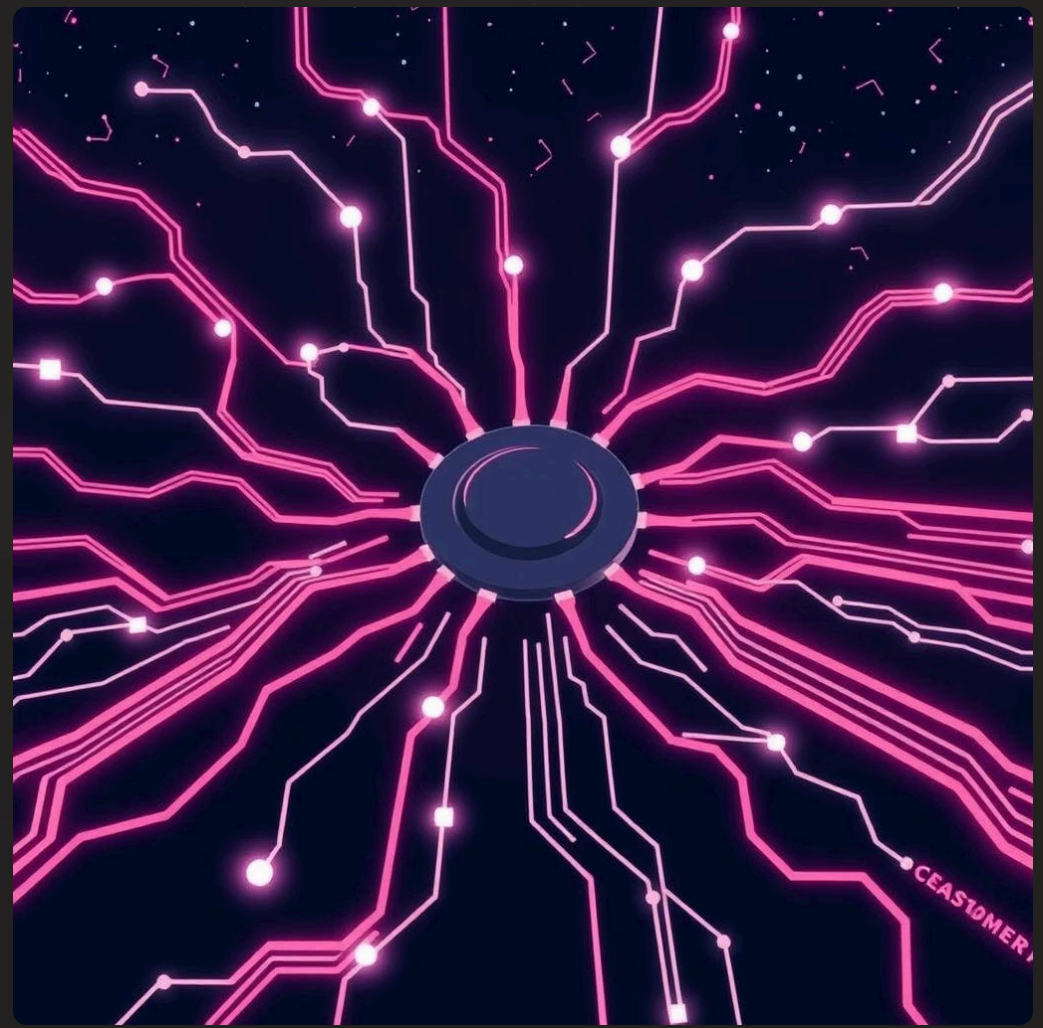
Jessica Ashraf Anis

The Modern Data Landscape

Data stands as the fundamental asset of modern organizations, driving innovation and strategic decisions. Companies today generate an unprecedented volume of data from diverse touchpoints:

- Customer interactions and profiles
- Transactional records
- Website clicks and user behavior
- Service logs and operational metrics

The paramount challenge lies in transforming this raw data into **actionable, real-time business decisions**.



The Business Imperative: Preventing Churn

Companies face substantial financial losses annually due to critical issues such as customer churn, suboptimal customer experiences, and delayed data-driven decisions.

Customer Churn

Millions lost as customers leave.

Poor CX

Dissatisfaction leads to attrition.

Late Decisions

Missed opportunities and reactive strategies.

Traditional systems often fall short: they cannot efficiently process real-time data, struggle to scale with big data volumes, and fail to accurately predict complex customer behavior.

> **Our Core Question:** How can we proactively predict customer churn before it impacts the business?

Project Objective: A Holistic Solution

This project aims to engineer a comprehensive platform designed to tackle these challenges head-on.

01

End-to-End Big Data & AI Platform

A unified system for all data and AI needs.

02

Batch & Real-Time Processing

Handling both historical and live data streams.

03

Machine Learning for Churn Prediction

Leveraging AI to foresee customer behavior.

04

Live Dashboards

Empowering decision-makers with instant insights.

Beyond the Basics: An Enterprise-Grade Platform

Our project transcends typical academic exercises, offering a fully integrated enterprise platform that harmonizes multiple advanced disciplines.

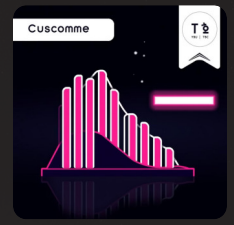
This is not merely a Spark, Flink, or Machine Learning project; it's a holistic solution.

Comprehensive Data Sourcing

We meticulously collected data from a diverse array of real-world sources to ensure robustness and realism.

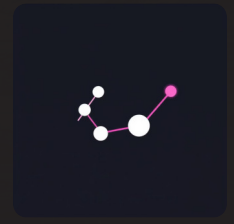
- Transactional customer data (purchases, subscriptions)
- Clickstream behavior (website navigation, app usage)
- Customer complaints and feedback
- Detailed customer surveys
- System logs and performance metrics

Real-World Complexity: Our approach embraces messy, multi-format data, mirroring true business environments.



Transactional Data

Clickstream Data



Survey Data

System Logs

Dynamic Data Ingestion Layer

Our ingestion strategy ensures continuous and seamless data capture from all sources.

Batch Ingestion

Leveraging **Azure Data Factory** for scheduled, bulk data transfers.

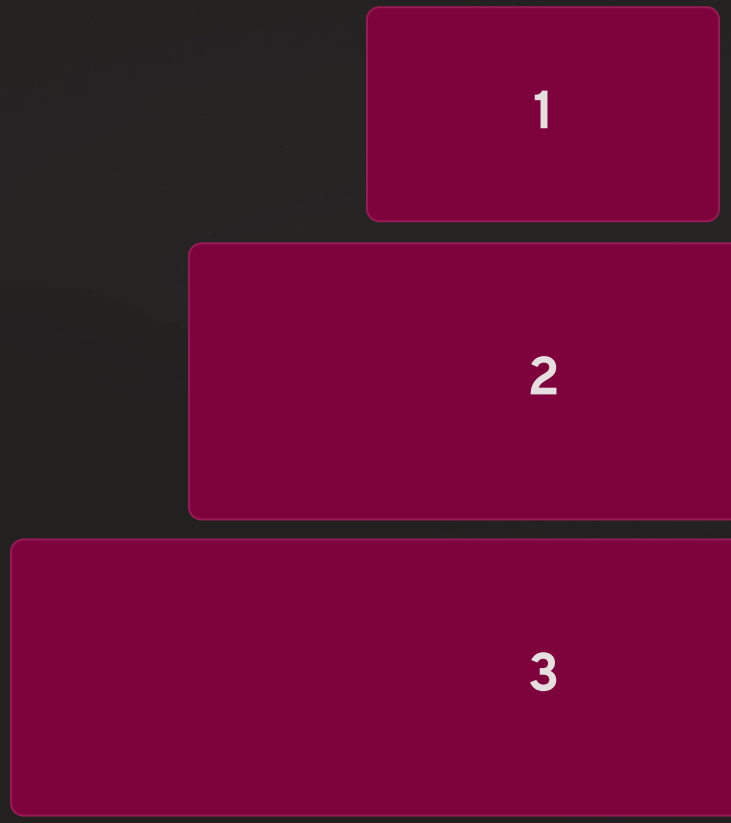
Real-Time Streaming

Utilizing **Apache Kafka / Event Hub** for high-throughput, low-latency data streams.

This dual approach guarantees that both historical archives and live operational data are continuously integrated into our platform without interruption.

Structured Data Storage: Data Lake Gen2

All ingested data is systematically organized and stored within **Azure Data Lake Gen2**, employing a layered zone approach for optimal data management, quality, and governance.



Raw Zone

Original, immutable, unprocessed data as it arrives from sources.

Curated Zone

Cleaned, validated, and structured data, optimized for analytical workloads.

Enriched Zone

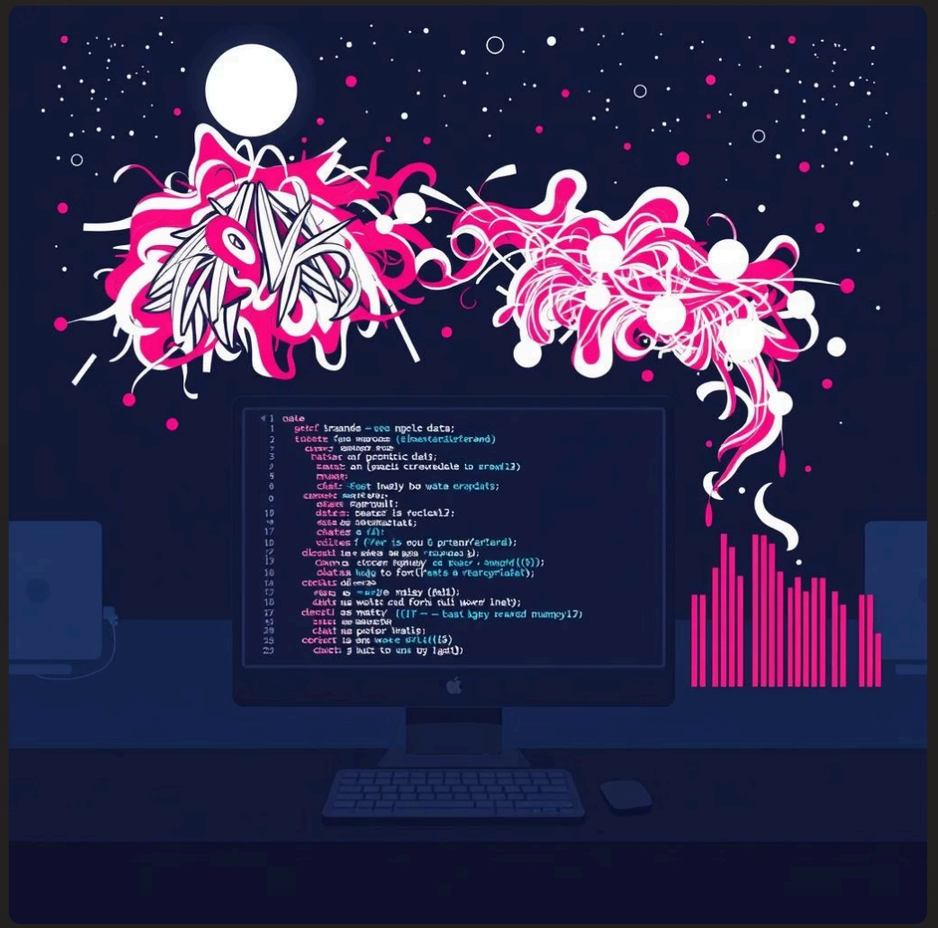
Fully processed, aggregated, and business-ready data, primed for consumption by AI models and dashboards.

This architecture ensures data integrity, accessibility, and readiness for advanced analytics.

Data Cleaning & Preprocessing: The Foundation of Insight

This critical phase, which consumed the majority of our project's effort, involved rigorous data cleaning and preparation to ensure the highest quality input for our models.

- **Handling missing values:** Employing imputation and strategic removal.
- **Removing duplicates:** Ensuring unique and accurate data records.
- **Outlier detection:** Identifying and managing anomalous data points.
- **Encoding categorical features:** Transforming qualitative data for ML models.
- **Feature scaling and normalization:** Standardizing data for optimal model performance.



Tools Utilized: Python, Pandas, NumPy. This meticulous process forms the bedrock for reliable predictions.

Exploratory Data Analysis (EDA) & Feature Engineering

Our comprehensive EDA provided deep insights into customer behavior, while targeted feature engineering significantly enhanced model predictive power.

Key EDA Discoveries:

- **Churn distribution:** Identified significant imbalance within the dataset.
- **Feature correlations:** Revealed relationships between variables.
- **Customer behavior patterns:** Uncovered segments and trends.
- **Usage vs. churn relationship:** Quantified impact of product usage.

Key Results: The dataset exhibited high imbalance, and both customer complaints and tenure were strong indicators of churn.

Impactful Feature Engineering:

- **Monthly usage rate:** Quantified product engagement.
- **Customer engagement score:** Holistic measure of interaction.
- **Complaints frequency:** Aggregated service issue occurrences.
- **Estimated customer lifetime value:** Projected future revenue.

Outcome: These newly engineered features significantly improved the accuracy and robustness of our machine learning models.

Impactful Feature Engineering

We engineered several new features to enhance the predictive power of our models, transforming raw data into meaningful predictors:

Monthly Usage Rate

Quantified user activity and product engagement over time.

Customer Engagement Score

A holistic metric capturing user interaction and satisfaction levels.

Complaints Frequency

Aggregated occurrences of reported service issues or dissatisfaction.

Estimated Customer Lifetime Value

Projected future revenue a customer is expected to generate.

These new features were crucial, significantly improving the accuracy and robustness of our machine learning models.

ETL & Orchestration: Automated Data Flow

Our platform leverages **Apache Airflow** for robust workflow orchestration, ensuring every step of the data pipeline is automated, monitored, and executed with precision.

Data Cleaning

Automated removal of inconsistencies, errors, and duplicates.

Feature Transformation

Standardizing and enriching data for optimal model readiness.

Data Validation

Ensuring data integrity and adherence to predefined quality rules.

Data Quality Checks

Continuous monitoring to maintain high standards of data reliability.

This comprehensive automation eliminates manual processing, reduces human error, and ensures timely, reliable data availability for all downstream applications.

Business Value

Our end-to-end platform delivers tangible business benefits, transforming how companies manage customer relationships and drive growth.

Reduced Churn

Minimize customer attrition by identifying at-risk customers early, allowing for timely, targeted retention efforts.

Higher Retention

Foster long-term relationships and build brand loyalty through deeper understanding and personalized customer experiences.

Informed Decisions

Empower stakeholders with real-time data and predicted insights for faster, more strategic business choices.

Proactive Engagement

Shift from reactive problem-solving to anticipatory action with early warnings and predictive models.

Revenue Growth

Directly contribute to the bottom line by optimizing customer lifecycle management and increasing customer lifetime value.

Machine Learning Model Training

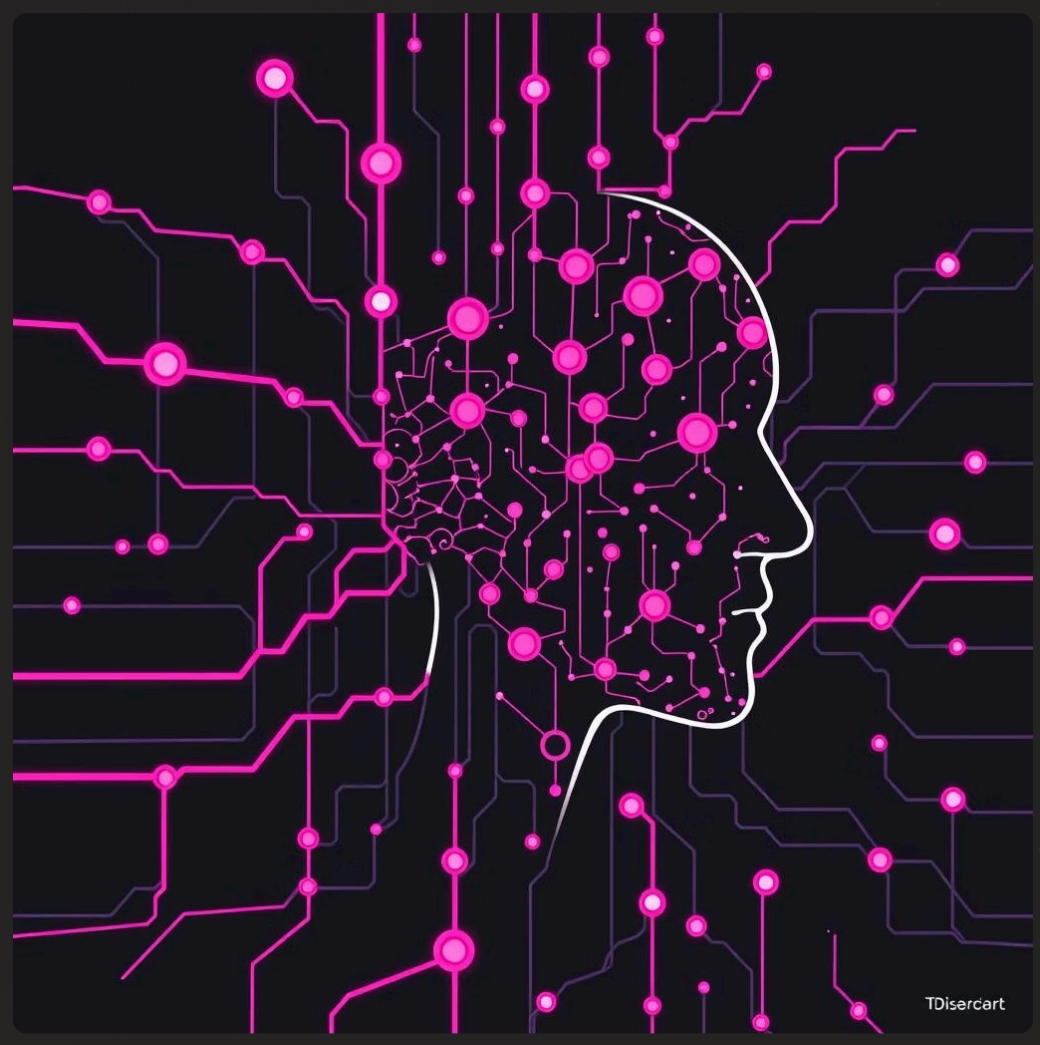
Our rigorous approach to model training involved evaluating multiple algorithms and fine-tuning their parameters to ensure optimal predictive accuracy for churn detection.

Models Trained

We selected a diverse set of powerful machine learning algorithms to capture various patterns in customer behavior:

- **Logistic Regression:** For its interpretability and efficiency as a baseline.
- **Random Forest:** Leveraging ensemble learning for robust predictions and feature importance insights.
- **XGBoost:** A highly optimized gradient boosting framework known for its superior performance in structured data.

Training Steps



01

Train/Test Split

Carefully partitioning our cleaned dataset into training and validation sets to ensure unbiased model evaluation.

02

Cross-Validation

Implementing K-fold cross-validation to robustly assess model performance and reduce overfitting risk.

03

Hyperparameter Tuning

Utilizing techniques like Grid Search and Random Search to optimize model parameters for peak performance and generalization.

Key Outcome: Through this systematic process, the best model was selected based on its superior performance across critical metrics for churn prediction.

Model Evaluation

Our churn prediction model underwent rigorous evaluation using a comprehensive set of metrics to ensure its reliability and effectiveness.

Accuracy

The overall correctness of the model's predictions.

0.9280

Precision

Proportion of correctly identified churners among all predicted churners.

0.9315

Recall

Proportion of actual churners correctly identified by the model.

0.9370

F1-Score

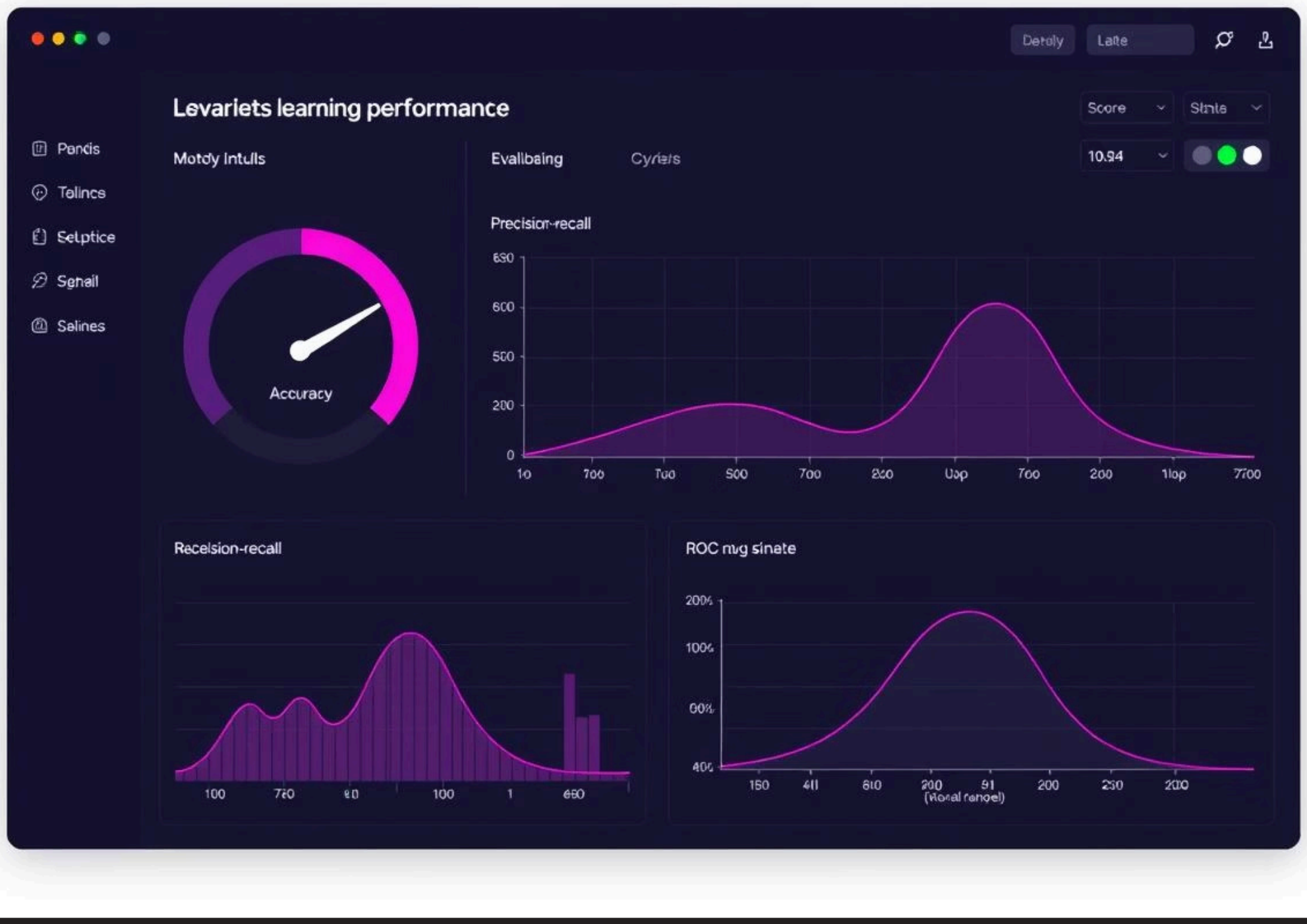
Harmonic mean of precision and recall, balancing both metrics.

0.9342

ROC-AUC

Ability of the model to distinguish between churners and non-churners.

0.9878



Strategic Priority: High Recall is specifically prioritized to ensure that potential churn customers are captured early, enabling proactive retention efforts.

Model Deployment

Our robust model deployment strategy ensures that churn predictions are not just accurate, but also actionable and integrated into daily operations, delivering insights when and where they're needed most.

Real-time Churn Prediction

Leverage advanced stream processing to identify and flag potential churn risks as they emerge, allowing for immediate intervention.

Batch Customer Scoring

Periodically score the entire customer base to categorize churn likelihood, enabling proactive, targeted retention campaigns.

Integrated Streaming Pipelines

Seamlessly feed predictions into existing data pipelines, enabling automated responses, triggers, and personalized customer engagements.

The deployed models are **directly connected to live dashboards**, providing immediate visibility and empowering timely business decisions.

Visualization & Dashboards

Our platform culminates in intuitive, real-time dashboards, empowering stakeholders with immediate access to critical insights for proactive decision-making.

Key Visualization Tools:

- **Power BI:** For comprehensive, interactive business intelligence reports and enterprise-level analytics.
- **Streamlit:** For rapid development of custom, lightweight data applications and real-time model monitoring interfaces.

These tools ensure both broad organizational access to data and specialized views for data scientists.

Dashboard Displays Include:

- **Live Churn Probability:** Real-time scores for each customer, indicating their likelihood of churning.
- **Customer Behavior Analytics:** Detailed breakdowns of user interactions, product engagement, and segment-specific trends.
- **Key Business KPIs:** A consolidated view of critical performance indicators, such as customer retention rates, average revenue per user, and acquisition costs.

Automation & MLOps Concept

Our platform leverages robust MLOps practices to ensure a fully automated and continuously optimized AI lifecycle.

Automated Ingestion

Seamlessly captures raw data from diverse sources without manual intervention.

Automated ETL Pipelines

Cleans, transforms, and loads data, preparing it efficiently for analysis and modeling.

Automated Retraining

Continuously updates and improves ML models with fresh data for optimal performance.

Automated Predictions

Delivers real-time churn predictions directly to business applications and dashboards.

Model Monitoring

Tracks model performance, drift, and data quality to ensure ongoing accuracy and reliability.

This comprehensive approach guarantees a **fully automated AI lifecycle**, from raw data to actionable insights.

Full System Architecture

Our comprehensive platform provides an end-to-end solution, integrating every critical stage from raw data acquisition to insightful visualization.

Data Sources

Diverse raw data inputs from various enterprise systems.

Ingestion

Batch and real-time capture of all incoming data streams.

Data Lake

Scalable storage for raw and processed data.

ETL & Orchestration

Data transformation, loading, and workflow management.

Processing

High-performance computing for data transformations.

Machine Learning

Advanced models for churn prediction and insights.

Visualization

Actionable dashboards for real-time decision-making.

This **Complete End-to-End platform** is meticulously designed to deliver seamless data flow and intelligence.

Business Value

Our Big Data & AI Platform delivers tangible business benefits, transforming challenges into opportunities across the organization.

Minimize Churn

Proactively identify and retain at-risk customers, preventing significant revenue loss.

Boost Loyalty

Foster stronger customer relationships through personalized experiences and timely interventions.

Informed Decisions

Empower stakeholders with real-time insights for strategic and operational choices.

Maximize Revenue

Drive growth by optimizing pricing, identifying upsell opportunities, and reducing operational costs.

Strategic Foresight

Shift from reactive problem-solving to proactive strategy development and execution.

Target Industries

Our platform is meticulously designed to address the unique challenges and opportunities within several key industries, where data-driven insights are paramount for competitive advantage and churn prevention.

Telecommunications

Optimizing customer retention and service personalization.

Banking & Financial Services

Enhancing fraud detection, risk assessment, and customer loyalty programs.

E-commerce

Improving conversion rates, personalizing recommendations, and reducing cart abandonment.

SaaS Companies

Predicting subscription churn and driving product adoption.

FinTech

Revolutionizing payment systems, lending, and investment platforms with real-time analytics.

These sectors stand to benefit most from our holistic Big Data and AI capabilities.

What We Proved as a Team

This project successfully demonstrated our collective proficiency across critical domains essential for modern data and AI initiatives.

Data Engineering

Big Data Processing

Machine Learning & AI

Real-Time Systems

Business Intelligence

Enterprise Architecture

Our integrated approach and expertise provided a robust, end-to-end solution designed for scalability and operational efficiency.

Final Conclusion

This project stands as a testament to our commitment to delivering impactful, production-ready solutions:

Real Industrial Big Data System

Designed to handle massive datasets with enterprise-grade reliability and scale.

Real AI Prediction System

Leveraging advanced machine learning for accurate, proactive churn prediction.

Real Business Solution

Directly addressing critical business challenges with measurable impact.

Full End-to-End Platform

A fully integrated, production-ready platform from data ingestion to actionable insights.

> **This is not just a graduation project—it's a real enterprise solution ready for deployment.**

Thank You!

Questions?