SERGIO LUIZ DOS SANTOS

Data Scientist | Risk-Oriented Analytical Products | Machine Learning in Production

Nova Lima, MG — Brazil

Email: santossergiorealbjj@outlook.com

Phone: +55 31 99370-8399

LinkedIn: https://www.linkedin.com/in/santossergioluiz

Portfolio: https://portfoliosantossergio.vercel.app

PROFESSIONAL SUMMARY

Data Scientist focused on building analytical products aimed at risk reduction and measurable financial impact. Combines 20+ years of experience in regulated and mission-critical systems with end-to-end machine learning development.

Experienced in translating business problems into data-driven solutions across the full ML lifecycle: problem framing, exploratory data analysis (EDA), feature engineering, model training, validation, deployment, and steady-state monitoring.

Strong ability to connect statistical performance to business decision-making, especially in risk-sensitive and compliance-driven environments.

English: Upper-Intermediate (advanced technical reading; improving spoken fluency)

TECHNICAL SKILLS

Programming Languages

Python | SQL | C# | .NET

Data Science & Machine Learning

Pandas | NumPy | SciPy | Scikit-learn

Gradient Boosting (XGBoost, LightGBM – exploratory projects)

Supervised Learning (Regression, Tree-based models)

Exploratory Data Analysis (EDA)

Feature Engineering

Cross-validation techniques

Model evaluation metrics (MAE, RMSE, R^2)

Model selection and trade-off analysis

Bias-variance awareness

Overfitting prevention

Data leakage prevention

Basic knowledge of Deep Learning (TensorFlow, PyTorch – academic studies)

Foundations of causal inference and limitations of predictive modeling

Data & Cloud Infrastructure

Azure Databricks | Supabase | Render | Docker

Windows Server | Linux Server

Data Visualization

Power BI | Matplotlib | Seaborn

Governance & Critical Environments

Data traceability

Access control

Operational continuity

Technical documentation

Work in regulated environments

FEATURED PROJECT

Predictive Operational Risk Platform

GitHub: github.com/Santosdevbjj/analiseRiscosAtrasoObras

Project built using structured synthetic data based on realistic operational assumptions.

Business Problem

Approximately 35% of deliveries were completed late, generating contractual penalties and indirect operational costs. Decisions were reactive, made only after delays occurred.

Core question:

Which operations present higher risk, and where should the company act first to reduce financial exposure?

Baseline

Forecasts based only on historical average delay.

Mean Absolute Error (MAE): 12 days.

No structured risk prioritization mechanism.

Solution Strategy

• Designed layered data architecture (raw → analytics → products)

• Consolidated operational variables (weather, soil type, supplier rating, schedule)

• Performed feature engineering oriented toward risk indicators

• Treated missing values and standardized categorical variables

• Applied cross-validation for robustness

• Compared Linear Regression and RandomForestRegressor

• Evaluated trade-offs between interpretability and predictive performance

RandomForest was selected due to its robustness to heterogeneous variables and resilience to noisy data.

Risks assessed during modeling:

• Overfitting mitigation through validation strategy

• Data leakage checks during feature construction

• Awareness of non-causal relationships in predictive modeling

Technical Performance

MAE (Baseline): 12 days

MAE (Model): 4.97 days

Error reduction: 59%

Consistent reduction of uncertainty in a high-variability operational environment.

Business Performance

Translated statistical error into financial exposure:

Estimated reduction in uncertainty: ~60%

Estimated annual financial impact: ≈ R$ 248,400

Decision shift: Reactive → Preventive

Each prediction error was converted into residual financial risk, enabling structured prioritization and proactive intervention.

Production Deployment

• Model deployed in cloud environment (Render)

• REST API implemented using FastAPI

• Model containerized with Docker

• Model versioning and reproducibility practices during deployment

• Telegram bot for operational queries with automated PDF reports

• Executive simulator built in Streamlit

• Data access layer managed via Supabase

Focus on accessibility for non-technical stakeholders, traceability, and real-world usability.

Key Learning

Translating model error metrics into defensible financial impact and clearly communicating statistical trade-offs to decision-makers in risk-sensitive environments.

PROFESSIONAL EXPERIENCE

Data Scientist – Applied Projects

Self-employed | 2023 – Present

• End-to-end ML lifecycle execution: problem definition → EDA → feature engineering → validation → deployment → monitoring

• Development of production-ready analytical products

• Model versioning and reproducibility practices in production environments

• Translation of statistical performance into measurable business impact

• Documentation aligned with data governance and traceability practices

• Evaluation of model trade-offs, limitations, and operational risks

Critical Systems & Infrastructure Consultant

Self-employed | 2008 – Present

• IT consulting in corporate mission-critical environments

• Windows and Linux server administration

• Integration of legacy systems with modern solutions

• Support for legal and administrative departments

• Work in regulated and compliance-sensitive contexts

Banco Bradesco S.A.

Clerk → Programmer → Systems Analyst | 1985 – 2008

Career progression from administrative operations to systems analysis and technical leadership in a large-scale banking institution.

Key Results:

• Automated internal processes, eliminating approximately 2,920 hours/year of manual work

• Supported systems serving 500+ users with 99.5% availability

• Developed systems focused on traceability, reliability, and regulatory compliance

• Worked in highly regulated banking environment with operational risk exposure

EDUCATION & CERTIFICATIONS

Data Science with Python – 2025

Azure Databricks – Microsoft – 2025

Neo4J and Graph Analysis – 2025

IBM AI Fundamentals – 2024

Microsoft Azure Bootcamp – 2024

Power BI and Advanced Excel – 2024

CURRENT STACK

Python | SQL | Pandas | NumPy | Scikit-learn | XGBoost | Azure Databricks | Power BI | Neo4J | Docker | Streamlit | Supabase