IRFAN ALI KHAN

Senior Data Scientist | Data Analyst | Econometrics & Revenue Modelling Expert

Email: *irfanaliguarulhos@gmail.com* ***Portfolio Website:*** *https://irfanaliguarulhos.github.io/irfanali.github.io/*

LinkedIn: *https://www.linkedin.com/in/irfan-ali-khan-93b52b159/* *WhatsApp: +5511969328044 (SP, Brazil)*

SUMMARY

Seasoned **Data Scientist and Senior Analyst** with over **10 years of experience** combining advanced **econometric modelling (TransFlow, PyTorch), Statistical Analysis, Machine Learning, AI, LLMs**, and **ETL pipeline engineering** to drive data-informed business decisions. Currently pursuing a **PhD in Economics**, with specialization in **Big Data, Forecasting Models, Revenue Optimization, and Statistical Computing**. Skilled in using **Python, R, SQL, MySQL, SPSS, Tableau, Power BI, and Excel** for high-impact analytics, reporting, and data transformation.

Extensive experience leading data projects in corporate, consulting, and academic environments globally, including roles at **KPMG, Accenture, and nonprofits**. Adept at simplifying complex data for stakeholders and applying **scientific research methods** to solve real-world problems. Passionate about building robust data systems and **Predictive pipelines using ML Algorithms (AutoML)** that support strategic planning and performance improvement.

SKILLS

- **Data Analysis & Programming: Python, PySpark, SQL, R, MySQL, NoSQL**
- **Data Visualization: Tableau, Power BI,** PowerPoint, **Matplotlib, Plotly, DAX, LOD Expressions**
- **Big Data & ETL: Apache Spark, ETL Processes, Real-Time Analytics, KPIs, Data Wrangling & Cleaning, AI, MLflow Tracking, AutoML, LLM, Hyperparameter Tuning,**
- **Machine Learning & Statistical: Scikit-learn, Stats models, Predictive Modelling, TensorFlow, PyTorch, Deep Learning, LLMs, NLP, A/B Testing, XGBoost, GLM/GAM, Feature Engineering, LightGBM, Langchain, SPSS, Excel Adv.**
- **Methodologies:** Agile, Scrum, Kanban, **Jira**
- **Languages: *English,*** *Portuguese, Spanish, Arabic, Hindi, Urdu, Pashto.*

EDUCATION

- **PhD in Economics**, USP – Universidade de São Paulo, 2021 – 2025
- **Postgraduate Degree in Big Data**, FACUMINAS, Dec 2020 – May 2023
- **MPhil in Economics**, Preston University, 2016 – 2019
- **MBA in Accounting and Finance**, Preston University, 2011 – 2013

CERTIFICATIONS: *• Google Advanced Data Analytics Specialization (Big Query) • IBM Certified Data Engineer - Big Data • Data Scientist and Machine Learning (AWS) • Advance Big Data Science & ML Certificate.*

EMPLOYMENT HISTORY

SWB, Senior Data Scientist & Analyst– Florida, United States (Apr 2024 to Jan 2025)

SWB is a dynamic non-profit organization focused on leveraging data to drive impactful community initiatives.

Activities and Responsibilities:

• Designed and developed **API-driven data pipelines integrating structured/unstructured data** from multiple sources. Automated data transformation & tagging processes, ensuring high accuracy in merchant analytics.

• Developed robust **ETL pipelines using SQL and Python to extract, transform, and load data** from multiple sources, with a focus on data normalization and standardization for maximum accuracy and **validity,** and loading data-frame in JSON/CSV format.

• **Applied forecasting models (ML, TensorFlow) for rent prediction and business planning**.

• Engineered and maintained detailed data models that supported **predictive feature engineering (LLM)** and reinforced data integrity throughout the **analytics lifecycle**.

• **Experienced in Power BI and Tableau**, specializing in interactive ***dashboards, data storytelling, performance optimization, and predictive modelling, utilizing advanced calculations, filtering techniques, and business intelligence strategies for impactful decision-making.***

Projects:

• *Renters Insight Dashboard* – Engineered an end-to-end analytics solution that consolidated data from diverse sources, incorporating advanced data modelling and rigorous ETL processes to deliver actionable insights that informed strategic decisions.

Technologies: SQL, Python, Tableau, R, Data Pipelines, Data Warehousing, Excel, Google Sheet

Data Analyst - KPMG Australia · Contract (Aug 2022 - Apr 2024)

• **Data Quality Assessment:** Conducted thorough assessments of data quality and completeness, ensuring accurate analysis and trustworthy data foundations.

• .**Customer Demographics Analysis**: Analyzed customer demographics to identify high-value segments and uncover actionable insights for targeted marketing strategies.

• **Data Dashboards & Presentations:** Created visually appealing data dashboards and presentations, enabling stakeholders to make informed decisions.

• Collaborative Projects: Built high-impact visual **reports on customer demographics, pricing models, and marketing response. And facilitated data-driven decision-making processes**.

• **Data Cleaning & Modelling**: Honed skills in data cleaning, modelling, and visualization, utilizing tools such as Python, R, **Tableau, and Microsoft Power BI.**

• **ETL Processes**: Implemented and **optimized ETL processes** to ensure seamless data ingestion and transformation, maintaining data integrity. Implemented API-driven data pipelines to improve **eCommerce data integration and merchant analytics. Conducted retail data analysis, revenue performance studies, and merchant behaviors modelling (LLM).**

• **Data Processing Engineering**: Employed statistical significant analysis to validate data significance

and accuracy, enhancing the data preparation process.

• **Proficient in Power BI and Tableau, developing intuitive dashboards**, defining KPIs, and implementing data-driven analytics using custom measures, trend forecasting, and visualization best practices to enhance business insights.

Technologies: SQL, Python, Tableau and Power BI, Data Pipelines, Data Warehousing, Excel. **Databricks.**

Data Analytics & Visualization – Accenture (Apr 2021 – Jul 2022)

Activities and Responsibilities:

Conducted extensive **exploratory data analyses using SQL and Python** to extract actionable insights, employing statistical techniques to validate data significance and accuracy.

• **API-Based Data Integration: JSON/CSV data processing**, API calls, automated file transfers. Merchant data processing, payment system analysis, retail data insights

• **Designed, developed, and automated dynamic reporting dashboards** in Tableau that integrated seamlessly with underlying data models normalized for accuracy and consistency.

• **Implemented and refined efficient ETL processes** to ensure smooth ingestion and transformation of data from multiple sources, with stringent standardization procedures to maintain data validity.

• **Proficient in Tableau dashboard design**, data visualization, and business analytics, creating intuitive and interactive reports. Strong expertise in custom calculated fields, LOD expressions, dynamic filtering, and predictive analytics to enhance data insights. Skilled in dashboard storytelling, hierarchy-based drill-downs, and visual optimization, ensuring impactful and actionable business reporting.

• **Built comprehensive data models to support predictive analytics** and **engineered features** that **enhanced forecasting accuracy, MLflow Tracking and business intelligence**

• **Data Cleaning & Modelling**: Honed skills in data cleaning, modelling, and visualization, utilizing tools such as Python, R, Tableau, and Microsoft Power BI.

Technologies: SQL, **Python**, **Tableau, Power BI**, Data Pipelines, Data Warehousing, Excel. **Databricks.**

PROJECT HIGHLIGHTS

- **Demand Forecasting Engine:** Modeled product demand using time-series ML methods.
- **Revenue Optimization Model:** Implemented regression-based price optimization.
- **Geo-Mapping Distribution Model:** Built models for rural logistics performance.
- **Research Projects Supervision:** Mentored 200+ university students globally in scientific data research using Python, R, SPSS, and Excel

Irfan Ali Khan (PhD)

Data Scientist | Data & Analytics | Revenue Growth Analysis | Researcher

Business Problem

Price Prediction, Fraud Detection, Inventory Optimization, Market Trends, Demand Forecasting, Customer Segmentation,Customer churn prediction, A/B testing frameworks, sentiment analysis.

Data Extraction & Sources

SQL | APIs | Web Scraping | IoT Data | SaaS (Salesforce, Snowflake) | Cloud Storage (AWS S3, Azure Blob, GCP BigQuery)

Data Engineering & ETL Pipeline

Apache Airflow | Spark | Databricks | Pandas | NumPy | ETL Pipelines

Bronze Layer

Data Ingestion (Connect data sources) Minimal Transformation Data Quality Checks

Gold Layer

Data Cleaning Transformation(Standardize data types, perform data integration) Intermediate Data Products

Diamond Layer

Data Modeling, high-value Data insights and data products for decision-making

Exploratory Data Analysis (EDA)

Matplotlib | Seaborn | Plotly | Descriptive Stats | Correlations | Outlier Detection

Descriptive Analysis, Statistical Analysis, Univariate Analysis, Bivariate Analysis, Multivariate Analysis, Correlational Analysis, Outlier Detection, Time-Series Analysis, Missing Value Analysis, Visualization and Data Storytelling, Dimensionality Reduction, Comparative Analysis

Feature Engineering

Scaling (MinMax, StandardScaler) | Encoding | Dimensionality Reduction (PCA, t-SNE) | Feature Selection (RFE, SHAP)

Filter Methods, Wrapper Methods, Embedded Methods, Recursive Feature Elimination, Forward Selection, Backward Elimination, SelectKBest, Mutual Information, Chi-Squared Test, Correlation Analysis, Variance Threshold, LASSO Regularization, Ridge Regression, Tree-Based Feature Importance, Permutation Importance, SHAP Values, PCA, Sequential Feature Selection, Univariate Feature Selection, Multicollinearity Analysis, Bayesian Feature Selection, Stability Selection, Dimensionality Reduction

Exploratory Data Analysis (EDA)

Matplotlib | Seaborn | Plotly | Descriptive Stats | Correlations | Outlier Detection

Descriptive Analysis, Statistical Analysis, Univariate Analysis, Bivariate Analysis, Multivariate Analysis, Correlational Analysis, Outlier Detection, Time-Series Analysis, Missing Value Analysis, Visualization and Data Storytelling, Dimensionality Reduction, Comparative Analysis

Model Building

Linear Regression | Random Forest | XGBoost | LightGBM | TensorFlow | PyTorch | NLP (NER)

Linear Regression, GLM, GAM, Random Forest, XGBoost, LightGBM, TensorFlow, PyTorch, scikit-learn, statsmodels, Apache Spark MLlib, PySpark, NLP, Sentiment Analysis, NER, Hugging Face Transformers, spaCy, NLTK, Large Language Models (LLM), BERT, GPT, H2O.ai, Azure AutoML

Model Optimization & Tuning

GridSearchCV | Optuna | Bayesian Optimization | AutoML (H2O.ai, Azure AutoML)

These optimization techniques collectively ensure that the selected model not only fits the data well but also generalizes to unseen data, making them indispensable in building robust, high-performance machine learning systems.

Deployment & Serving

Flask | FastAPI | Docker | Kubernetes | AWS SageMaker | Azure ML | MLflow Tracking

Each of these tools plays a crucial role in ensuring that machine learning models are not just built, but are also efficiently deployed, served, and monitored in a production environment, thereby enabling seamless and scalable delivery of data-driven solutions.

Monitoring, Testing & MLOps

Prometheus | Grafana | Drift Detection | Retraining Pipelines | CI/CD Automation

Combined with ***automated retraining pipelines and robust CI/CD practices***, these tools and techniques help maintain high model accuracy and system stability over time.

Reporting, Insights & Business Strategy

Dashboards (Tableau, Power BI, Streamlit) | Automated Reporting | NLG (Natural Language Generation)

Reporting, Insights & Business Strategy stage not only delivers comprehensive analytics but also ensures that the data-driven insights are actionable and truly aligned with the overall business strategy. This holistic approach transforms raw data into strategic decisions, ensuring that every insight drives impactful outcomes across the organization.