# Gurpreet Singh Deol

Southall, West London | 07379806602 | Gurpreet.S.Deol@outlook.com | LinkedIn | GitHub

## PROFESSIONAL SUMMARY

MSc Data Science and BSc Physics graduate with a solid foundation in data analytics, machine learning and ETL pipeline development. Skilled in Python, SQL and Power BI with a proven history of creating predictive models, building automated data pipelines, designing interactive dashboards for analysis and comprehensive report writing. Professional experience in multiple customer facing roles with projects completed in various industries including finance, public health, logistics and environmental conservation.

## EDUCATION

**Royal Holloway University of London - Egham, Surrey**

***MSc Data Science and Analytics: Pass with Merit*** **Oct 2021 - Oct 2022**

Key Modules: Data Analysis, Machine Learning, Artificial Intelligence, Databases and Business Intelligence.

- First-class dissertation: Analysed historical football data with unsupervised machine learning techniques to identify tactical trends over 13 years.
- Gained expertise in developing machine learning models and gained experience in Python, SQL, R and MATLAB. Completed several projects using machine learning techniques to build predictive models and uncover insights in complex datasets.

***BSc Physics: Second Class Honors (Upper Division), 2:1*** **Sep 2018 - Jul 2021**

Key Modules: Mathematical Methods, Statistics, Quantum Mechanics, Particle Physics and Energy and Climate Science.

- First-class dissertation: Simulated electrical resistance in 2D geometries using Python.
- Developed strong analytical and critical thinking skills by tackling complex physics problems and developed a solid foundation in data analysis.

## SKILLS

**Programming Languages:** *Python, PostgreSQL, R and MATLAB.*
**Cloud Tools:** *AWS (Lambda, EC2, S3) and Azure.*
**Data Libraries:** *Pandas, Numpy, Scikit-Learn, Keras, TensorFlow, Geopandas and Psycopg2.*
**Reporting & Data Visualisation:** *Microsoft PowerBI, Excel, Matplotlib, Seaborn and LaTex.*
**Big Data Tools:** *PySpark and Hadoop.*
**Other Tools:** *Git/GitHub, Docker, Visual Studio Code, PostGIS and Jupyter Notebook.*

## PROJECTS

**Public Health and Air Quality Impact Analysis Data Pipeline | GitHub**

- Developed an ETL pipeline in Python to process air quality and health data, over 20 metrics from 10 global cities, integrating various APIs for data collection.
- Automated collection with Docker and AWS Lambda, and stored results in a PostgreSQL database.
- Performed time-series analysis to uncover correlations between pollution levels and respiratory illnesses, presenting insights via interactive Power BI dashboards for real time comparisons and plots using Matplotlib and Jupyter Notebook.

**Predicting Illegal Fishing using a Random Forest algorithm | GitHub**

- Built a 97% accurate Random Forest model to classify vessel fishing activity based on AIS data and geospatial features (speed, distance from shore/port, coordinates).
- Integrated real-time data ingestion via the Global Fishing Watch API to automate collection of vessel event data.
- Used PostGIS and GeoPandas to spatially filter vessel positions within Marine Protected Areas (MPAs) and open ocean zones.
- Detected and labelled potential illegal fishing activity by intersecting predicted fishing locations with protected waters.
- Created map visualizations to highlight high-risk zones, color-coded by fishing legality using Matplotlib and shapefiles

**Exploratory Data Analysis: Customer Loans in Finance | GitHub**

- Conducted exploratory data analysis on a dataset of 54,000 financial loan records, focusing on identifying trends and anomalies in loan repayment and customer risk profiles.
- Cleaned the data using Pandas and applied visualization techniques in Matplotlib to analyse repayment statuses and forecast financial risks.
- Highlighted insights on customer behaviour and financial performance trends to improve risk assessment.

**Sports Data Analysis: Football and Machine learning | GitHub**

- Scraped football data over a period of 13 years, creating a dataset with over 2,000 records and 100+ attributes, cleaned the data and applied feature engineering to prepare the dataset for analysis.
- Applied Principal Component Analysis and K-means clustering for a time series analysis to identify tactical shifts and trends in football across the decade, visualizing results using Matplotlib.
- Presented findings through a comprehensive report in LaTeX, highlighting potential factors influencing these trends.

## BOOTCAMPS AND CERTIFICATES

**AI Core Data Analytics Intensive Bootcamp** **Dec 2023 - Feb 2024**

- Completed 14 learning courses on topics including EDA, data cleaning with pandas, SQL, object-oriented programming, Power BI, Azure, AWS and two projects on python programming.