**B.Surendra Kumar**

**Data Scientist**

**+91-9493610434**

**sk.techie6@gmail.com**

**SUMMARY:**

* 3 years of overall IT Experience in Data Science including Machine Learning, Data Mining and Statistical Analysis.
* Experience in architecting and building comprehensive analytical solutions in Marketing, Sales and Operations functions across Retail, Logistics, worked closely with functional team leaders (in Product, Marketing, etc.) to explain analysis, findings, and recommendations.
* Experience in acquiring, merging, cleaning, analyzing and mining structured, semi-structured and unstructured data sets for analysis
* Strong track record of contributing to successful end-to-end analytic solutions (clarifying business objectives and hypotheses, communicating project deliverables and timelines, and informing action based on findings)
* Expertise writing production quality code in SQL and Python.
* Involved in independent research and experimentation of new methodologies to discover insights, improvements for problems. Delivered findings and actionable results to management team through data visualization, presentation, or training sessions. Proactively involved in roadmap discussions, data science initiatives and the optimal approach to apply the underlying algorithms.
* Experience building interpretable machine learning models, and building end to end data pipelines which included extracting, transforming and combine all incoming data with the goal of discovering hidden insight, with an eye to improve business processes, address business problems or result in cost savings
* Experience in Data Modelling retaining concepts of RDBMS, Logical and Physical Data Modelling until 3NormalForm (3NF) and Multidimensional Data Modelling Schema (Star schema, Snow-Flake Modelling, Facts and dimensions). Hands on experience in optimizing the SQL Queries and database performance tuning in Oracle, SQL Server.

**TECHNICAL SKILLS**

|  |  |
| --- | --- |
|  | **Exploratory Data Analysis:** Univariate/Multivariate, Outlier detection, Missing |
|  | value imputation, Histograms/Density estimation, EDA in Tableau |
|  | **Supervised Learning:** Linear/Logistic Regression, Decision Trees, Ensemble Methods, Random Forests, Naïve Bayes, XG Boost Machine learning NLP. |
|  | **Unsupervised Learning:** Principal Component Analysis, Association Rules, Factor |
|  | Analysis, K-Means, Hierarchical Clustering. |
| **Statistics/ML** | **Feature Engineering:** Stepwise, Recursive Feature Elimination, Relative Importance, |
| Filter Methods, Wrapper Methods and Embedded Methods |
|  |
|  | **Statistical Tests:** T Test**,** Chi-Square tests, Stationarity tests, Auto Correlation tests, |
|  | Normality tests, Residual diagnostics, Partial dependence plots and Anova. |
|  | **Sampling Methods:** Bootstrap sampling methods and Stratified sampling, SMOTE, |
|  | Under Sampling and Over Sampling. |
|  | **Model Tuning/Selection:** Cross Validation, AUC, Precision/Recall, Walk Forward |
|  | Estimation, AIC/BIC Criterions, Grid Search and Regularization |
|  | **Time Series:** ARIMA, Holt winters, Exponential smoothing, Bayesian structural time |
|  | Series. |
| **Machine Learning** | **Python:** pandas, NumPy, scikit-learn, SciPy, stats models, matplotlib, |
|  | TensorFlow, SpaCy, NLTK. |
| **ML Tools** | Dataiku, ML Studio (Azure), Jupyter Notebook, PyCharm |
|  |  |
| **Databases/ETL/Query** | SQL Servers, Oracle, Subqueries, joins, DDL/DML statements |
|  |  |
| **Visualization** | Seaborn, Plotly, MS-Office (MS - Excel, MS-Word, MS-PowerPoint) |
| **Prototyping** | PowerPoint |

**EDUCATION**

* **B.E. (Electronics and Communication Engineering) |2012-2016** – Yogananda Institute of Technology and Science.

**WORK EXPERIENCE**

# LENOVO

**Role: Data Scientist**

**Project Description:** They are responsible for selling electronic products seasonally. Lenovo products like Laptop,tablet computers, smartphones, workstations, supercomputers, electronic storage devices …. These products are highly selling products. Because of these highly selling products customers started calling to customer care related to queries regarding the products. Multiple calls are receiving related the products we wanted to an automated way solving these queries and Sales forecast on Lenovo products.

# Solution Approach:

1. **Customer service chatbots can enhance the customer experience as well as reduce costs of service delivery**.

Model: RASA NLU / Stemming, Lemmatization, TF- IDF Vectorizer, Text mining.

**b) Sales forecast on Lenovo products.**

Model: Applied Time Series model – ARIMA to predict the Lenovo products.

**WIPRO LIMITED**

**Client: DHL.**

**Role: Senior Associate.**

**Project Description:** Goods shipped internationally into a country are subjected to Customs clearance process. Clearance is the process where a package entering a receiving country is reviewed by the customs and cleared to be sent to receiver after evaluating the documents of the package.

This system provides DHL several Corporate wide benefits which are as follows:

* Reduces customs clearance processing time.
* Enables management to expedite the express clearance of held packages.
* Ensures the integrity of the Cage inventory.
* Improves customer service.
* Reduces customs penalties and fines.
* Provides timely and meaningful information to management.

However, Due to incomplete shipment information like missing documents, incorrect duties and taxes are paid, invalid International Harmonized System Code etc., **DHL** had been paying customs around $1.4 M per annum in the form of penalties and fines.

By implementing the below AI & ML techniques, we are trying to minimize the penalties and fines.

1. **Entity Profiling** – We have created individual Feature Stores for Location/Commodity/Party Profiling from the historical shipping information. This shipping information is transformed, aggregated, and stored in Feature Store tables for further use by Cage Risk Models. Additional features can be generated using clustering Models to segment the
2. Location/Commodity/Parties who exhibit similar behavior to different categories

Model: K-Means Clustering

1. **Shipment Property Anomaly Detection (SPAD) service** - Early detection of anomalous shipment information increases the resolution time available till a shipment failure event and provides an automated feedback mechanism to rate the quality of information provided for international shipments.

Model: Auto Encoder & Statistical Approach

1. **Document Identifier** – This is a Multi class Model to predict the Document Type. While uploading the supporting documents for an item to be shipped, we can proactively suggest the ‘Document Type’ to the customer without which he could have uploaded the document as type ‘OTHER’. The ‘OTHER’ type of documents may increase the risk of caging resulting in the additional expense for the Client.

Model: OCR (OpenCV) & XG Boost

1. **Named Entity Recognition (NER**) - NER Model to identify the common entities in shipment documents such as Name, Location, Document Name, Product description, Harmonized System Code etc., for structured data extraction Model: SpaCy NER
2. **Cage Risk Model** –Built predictive models that can reliably predict the risk of a package being caged during different stages of a shipment and also highlight the possible cause/s.

Model: RF Classifier (Binary) and MLP Classifier (Multi Label)

1. **Cage Dwell Model**- Developed a machine learning model that will predict the number of days an “at risk” shipment will be held in clearance which would benefits the business by providing advance notification and allowing for proactive intervention to reduce delay in international shipping lanes. It also triggers recalculation of Estimated Date of Delivery and notifies to both the client and **DHL** operations.

Model: XG Boost