SOUJANYA CHERUVU

SR. DATA SCIENTIST
Linkedin | Tableau | Github
+1 647-897-8459
Soujanya.cheruvu19@gmail.com

PROFILE SUMMARY

- Overall 6+ years of hand on experience in Data Science in implementing AI solutions to optimize business processes & minimize costs.
- Proficiently manage end-to-end project cycles: use case framing, data collection, exploration, visualization, model creation, performance evaluation, and deployment.
- Skilled in a diverse array of machine learning models: Regression (Logistic, Linear, Non-Linear), K-Means, Random Forest, Decision Trees, Support Vector Margin, Time Series (AR, MA, ARIMA), and Deep Learning architectures like ANN and CNN.
- Efficient in preprocessing data in Python and R using Visualization, Data cleaning, Correlation analysis, Imputations, Feature Selection, Scaling and Normalization, and Dimensionality Reduction methods.
- Knowledge in Text Mining, Topic Modelling, Association Rules, Sentiment Analysis, Market Basket Analysis, Recommendation Systems, Natural Language Processing (NLP).
- Proficient in data visualization tools such as Tableau, Power BI, Python Matplotlib, Python Seaborn, to create visually powerful and actionable interactive reports and dashboards.
- Skilled in Data chunking, Data profiling, Data Cleansing, Data mapping, Creating Workflows and Data Validation using data integration tools like Informatica during the ETL process.
- Worked on various databases like Big Data, Hadoop, Teradata, Oracle, MYSQL, SQL Server, DB2 and expertise in creating tables, data population and data extraction from these databases.
- Sound knowledge and Experience on SDLC methodologies in Agile and Waterfall.

TECHNICAL SKILLS

Statistical Methods	Hypothetical Testing, ANOVA, Normality Testing, Bayes Law, Principal Component Analysis (PCA)
Machine Learning	Linear Regression, Logistic Regression, Naïve Bayes, Decision Trees, Random Forest, Support Vector Machine(SVM), Neural Networks, Sentiment Analysis, K-Means Clustering, K-nearest Neighbors (KNN), Time Series (ARIMA), Ensemble Methods, Gradient Boosting Trees, XGBoost, Natural Language Processing, ANN, CNN
Data Visualization	Tableau, Power BI, Looker Studio, QlikView, Python (MatPlotLib, Seaborn), R(ggplot2)
Languages	Python (Pandas, Numpy, SciPy, Matplotlib, streamlit, scikit-learn, pickle, BeautifulSoup, Scrapy, request, geopandas, NLTK, Plotly, Seaborn, TensorFlow, Pytorch, Flask), R Programming, SQL, C, COBOL
Databases	Big Data, Hadoop, Hive, Teradata, ORACLE, SQL Server, DB2, Spark, BigQuery
Tools/Technologies	R Studio, Jupiter Notebook, PyCharm, Data Bricks, Informatica, QC, Remedy, ServiceNow, MS Project, SAP Solution Manager (SolMan), VISIO, SharePoint, SalesForce
SDLC Methodologies	Agile, Waterfall

PROFESSIONAL EXPERIENCE

SENIOR DATA SCIENTIST | X-AXIS SOLUTIONS | CANADA

Mar 2021 - Present

Project description:
Developed a Fraud Detection and Prevention System (FDPS) for insurance transactions, utilizing machine learning for real-time analysis of transactional data. Applied practical techniques such as anomaly detection and predictive models to swiftly identify and prevent potential fraud. Strengthened user authentication and enhanced overall security in the insurance sector.

Responsibilities:

- Acquired and analyzed unstructured and semi-structured data using AWS, employing Amazon S3 for storage and AWS EMR for distributed data processing.
- Engineered feature sets from diverse sources, primarily medical records, using Python libraries (Pandas, NumPy) for comprehensive risk assessment.
- Conducted sentiment analysis using AWS Comprehend and entity recognition with Google Natural Language API to extract insights efficiently.
- Developed and implemented machine learning algorithms using scikit-learn and TensorFlow for real-time analysis of transactional data. Implemented ensemble methods (Random Forest, XGBoost) from scikit-learn for predictive modeling in insurance claims fraud detection.
- Integrated the PyOD library for anomaly detection, identifying unusual patterns indicative of potential fraud.
- Utilized various techniques such as Apriori algorithm, K-means clustering, RNN, NLP techniques for behavioral analysis, and integrated them with our internal system to enhancing the system's ability to recognize suspicious behavior.
- Conducted rigorous A/B testing methodologies to evaluate and optimize predictive underwriting models.
- Collaborated closely with domain experts, stakeholders, and DevOps teams throughout the project lifecycle. Ensured effective communication and integration of the developed models into existing systems, guaranteeing smooth deployment and utilization of the implemented solutions.
- The implemented models demonstrated a 20% improvement in fraud detection accuracy, contributing to a more robust and efficient fraud prevention system. The collaborative efforts also facilitated quick adaptability to emerging fraud patterns, showcasing the effectiveness of the integrated models in real-world scenarios.

- Documented project specifics, including tool selection and utilization, emphasizing code reusability and modularity, implementing version control best practices for collaborative coding.
- Integrated Tableau for advanced data visualization, providing stakeholders with interactive dashboards to comprehend fraud detection trends and patterns.

Environment: Python, AWS (SageMaker, S3, EMR, Lambda, Comprehend) Google Natural Language API, scikit-learn, TensorFlow, PyOD, Apache Spark, Apriori algorithm, K-means clustering, Tableau

SENIOR DATA SCIENTIST | BIRCH COMMUNICATIONS | CANADA

Dec 2019 - Mar 2021

Project description:

Create a predictive model aimed at anticipating high-priority (P1 & P2) IT tickets, facilitating preemptive measures and timely issue resolution. Simultaneously, forecast quarterly and annual incident volumes across diverse domains, enhancing preparedness for resource allocation and strategic technology planning.

Responsibilities:

- Curated data from diverse sources, seamlessly integrating them using Amazon S3 for centralized storage and integration, enhancing data acquisition efficiency.
- Engineered a personalized Streamlit interface in Python within an internal data application, allowing real-time updates (leveraging Amazon Lambda) and analytics for predicting and preventing priority 1 & 2 IT Service Management incidents.
- Forecasted incident volumes across various fields quarterly and annually, contributing to strategic resource and technology planning for proactive measures.
- Oversaw the entire lifecycle of the prediction model, from initial development and regulatory validation to deployment, performance monitoring, and updates to adapt to changing business environments.
- Improved model accuracy by 10% through optimization techniques like GridSearch, RandomizedSearch, and Cross-Validation to fine-tune hyperparameters.
- Employed the Hadoop Distributed File System (HDFS) for efficient and scalable storage of incident data, ensuring optimal data management.
- Applied sentiment analysis on customer reviews, implementing SMOTE techniques to handle class imbalances and achieving a 25% increase in accuracy for sentiment classification, along with a 30% improvement in nuanced customer sentiment identification.
- Orchestrated the deployment of Amazon EMR (Elastic MapReduce) for scalable and distributed data processing, enhancing overall data processing capabilities.
- Collaborated with the IT team to deploy and monitor the model after rigorous testing.
- Leveraged MLflow, an open-source platform, to track experiments, manage model versions, and establish a reliable framework for reproducibility. This facilitated effective management of different model iterations and streamlined the model development process.
- Resulted in a 20% reduction in resolution time for high-priority tickets and a 30% improvement in accuracy for resource and technology planning, as facilitated by the forecasted ticket volumes.

Environment: Python, Amazon S3, Streamlit (Python), Amazon Lambda, GridSearch, RandomizedSearch, Cross-Validation, Hadoop Distributed File System (HDFS), SMOTE, Amazon EMR, MLFlow

DATA SCIENTIST | AXIS BANK | INDIA

Sept 2018 - Nov 2019

Project Description:

The project aimed to forecast customer churn within the credit card customer base, identifying influential variables prompting customer migration. The objective was to redirect marketing efforts towards these customers, reengaging them with the products. The focus was on studying customer behavior to pinpoint patterns leading to disengagement from the product.

Responsibilities:

- Spearheaded and led a comprehensive customer churn prediction initiative at AXIS Bank, leveraging cloud services such as AWS for end-to-end data management, storage, and processing.
- Employed web scraping tools and APIs to efficiently collect extensive credit card customer data, subsequently stored in AWS S3 buckets for seamless accessibility.
- Utilized R for in-depth data exploration and cleaning, ensuring data integrity. Utilized Amazon Redshift, a cloud-based data warehouse, for efficient data querying and manipulation.
- Applied statistical techniques and domain knowledge to perform feature selection, optimizing the model's predictive power.
- Implemented machine learning models using a combination of scikit-learn, AWS SageMaker, and XGBoost, striking a balance between interpretability and performance.
- Employed a cloud-based ETL solution, AWS Glue, to seamlessly integrate data for the creation of training and testing sets.
- Utilized Power BI for visualizing model outputs and developing interactive dashboards, facilitating effective communication of insights to stakeholders.
- Collaborated closely with marketing teams to propose and implement targeted strategies, resulting in a notable 15% reduction in customer churn.
- Established a robust monitoring system using AWS CloudWatch and Lambda functions, allowing real-time assessment of strategy effectiveness.
- Iteratively improved the model, achieving a commendable 5% increase in accuracy through continuous feedback analysis.

- Documented the entire end-to-end process using Jupyter Notebooks, and AWS documentation practices, ensuring transparency, reproducibility, and knowledge transfer.
- Communicated findings and recommendations effectively to stakeholders through interactive Power BI reports, contributing to data-driven decision-making within the organization.

Environment:
AWS (S3, Redshift, SageMaker, Glue, CloudWatch, Lambda), R, Python (Pandas, scikit-learn, XGBoost), Power BI, Jupyter Notebooks.

JUNIOR DATA SCIENTIST | GEMINI EDIBLES AND FATS PVT. LTD | INDIA

June 2017 - Aug 2018

Project Description:
The project aimed to refine the marketing strategy by efficiently segmenting historical customer data using k-means clustering. Subsequent A/B testing analysis of past campaigns identified areas for improvement, leading to actionable insights and recommendations for optimizing future initiatives.

Responsibilities:
- Conducted an extensive analysis of five years' worth of historical customer data using SQL for efficient data retrieval and Python (Pandas, NumPy) for data manipulation. Performed data cleaning and exploratory analysis to gain insights into key features for further analysis.
- Utilized k-means clustering, a traditional statistical method, to segment the customer base, identifying five distinct customer segments. This approach resulted in a 15% improvement in engagement and a remarkable 20% increase in sales within our state.
- Developed and optimized data pipelines, incorporating SQL and Python (Pandas), reducing data processing time by 40%. This optimization ensured the timely availability of data for analysis, contributing to a more agile and responsive data-driven environment.
- Executed A/B testing on major marketing campaigns, leveraging statistical hypothesis testing techniques. The results revealed a 25% improvement in overall campaign effectiveness and an 18% increase in conversion rates, offering actionable insights for refining future campaign strategies.
- Created weekly and monthly Tableau dashboards for stakeholders, presenting key performance indicators and discussing current sales data and improvement strategies. Automated this process to provide real-time updates, facilitating clear communication of actionable insights to non-technical stakeholders and aiding in better decision-making across departments.
- Implemented regression analysis and predictive modeling to forecast sales based on campaign insights, resulting in a 12% improvement in the accuracy of sales predictions. This enabled more informed resource allocation for future campaigns.
- Generated detailed reports highlighting a 10% increase in sales directly attributable to the optimization of marketing strategies based on data-driven insights. These reports served as a comprehensive guide for decision-makers, providing a tangible understanding of the impact of data-driven marketing initiatives.
- Collaborated effectively with cross-functional teams, presenting findings and recommendations in bi-weekly meetings. This collaborative approach fostered a data-driven decision-making culture within the organization.
- Regularly updated stakeholders on project progress through weekly reports, ensuring transparent communication and alignment with business objectives. This proactive communication strategy facilitated a cohesive understanding of the project's impact and ensured continuous improvement in marketing strategies.

Environment:
Python, SQL, Jupyter notebook, Statistical Methods(A/B testing, Hypothesis Testing), Tableau, MS Excel, SAP

EDUCATION

University of Texas, Austin - Post Graduate Program in Data Science and Business Analytics — **July 2020- Aug 2021**

Statistical Data Analysis, Natural Language Processing (NLP), Exploratory Data Analysis, Data Mining, Data Modeling, Data Engineering, Statistics, Data Visualization

Western Sydney University, Sydney - Masters in Accounting — **July 2016 - Aug 2017**

Accounting, Finance, Law, Fraud Prevention, Forensic Accounting, Corporate Finance

Osmania University, India - Bachelor's of commerce (Hons) — **June 2013 - July 2016**

Statistics, Accounting, Finance, Law, Economics, Marketing