

**Porkalai Duraisamy**

**DATA SCIENTIST/ML ENGINEER**

[**d.porkalai@gmail.com**](mailto:d.porkalai@gmail.com)

**Ph: 972-472-1995**

**https:**[**//www**](http://www.credly.com/users/porkalai-duraisamy)**.c**[**redly.com/users/porkalai-duraisamy**](http://www.credly.com/users/porkalai-duraisamy)

**https://www.linkedin.com/in/porkalai-d/**

EDUCATION

|  |  |  |  |
| --- | --- | --- | --- |
| **Degree** | **Stream** | **University** | **GPA** |
| Bachelor of Engineering (B.E.) | Electronics and Communications Engineering | Anna University (2005  – 2009) | 4.0/5.0 |

CERTIFICATION

|  |  |  |
| --- | --- | --- |
| **Certification** | **Certification No** | **Issuing Organization** |
| Microsoft Certified Azure Data Scientist Associate | H975-5442 | Microsoft |

PROFESSIONAL SUMMARY

* Around 8+ years of professional experience in Retail, Banking and Financial Services and Health Care with adept knowledge on Data Analytics, Machine Learning (ML), Predictive Modelling, Natural Language Processing (NLP) and Deep Learning algorithms.
* Proficient in Data cleaning, Exploratory data analysis (EDA) and Initial Data Analysis (IDA).
* Proficient in entire data science project life cycle and actively involved in all the phases of project life cycle including, Data Acquisition, Data Extraction, Data, Engineering, Data Pre- Processing, Data Cleaning, Features Scaling, Feature Engineering, Dimensionality Reduction, Statistical Modeling, Algorithm implementation, AB Testing & Validation and data visualization.
* Experienced knowledge in machine learning algorithms such as Ensemble Methods (Random forests), Linear, Logistic Regression, Regularized Linear Regression, SVMs, XGBoost, Deep Neural Networks, Extreme Gradient Boosting, Decision Trees, K-Means, KNN, Hierarchical models, Naïve Bayes.
* Created and optimized diverse SQL queries, working knowledge of Azure Synapse Analytics, Azure Blob storage, Azure SQL, MySQL.
* Expertise in Natural Language Processing (NLP) in python including tokenization, lemmatization, stemming, parts of speech tagging, sentimental analysis.
* Proficient in deep learning models like Recurrent Neural Networks (RNN) using Python. * Applied different techniques to evaluate using different Machine learning models.
* Hands-on with Python Libraries like NumPy, Pandas for Data Manipulation and Matplotlib, Seaborn for Data Visualization, Scikit Learn for model building.
* Highly competent at wide varieties of Data Science programming languages and Big Data tools such as Python, SQL, Tableau, Sci- kit Learn, Spark, and Hive.
* Experience in Extracting data for creating Value Added Datasets using Python, Azure, AWS and SQL. * Used Cloud technologies such as Azure and AWS for data retrieval.
* Strong analytical skills proven over multiple projects.
* Good understanding of computational environments like cloud Microsoft, Azure, Apache Spark.
* Statistical programming languages such as Python, MATLAB, and query language like SQL. Proficient in using databases like Azure Cosmos DB, Mango DB, PostgreSQL, MySQL, Amazon Aurora, Amazon RDS, Amazon Dynamo DB.
* Experience in Enhancing data collection procedures to include information that is relevant for building analytic systems.
* Experience with complex Data processing using Azure pipelines, Azure Data Factory including ETL and Data ingestion dealing with unstructured and semi-structured Data.
* Worked on IDE's like Jupyter Notebook, Spyder, PyCharm, Subline Editor, Atom also expertise in tools like Microsoft Excel, Tableau.
* Well versed with dealing with Structured and Unstructured data, Time Series data and statistical methodologies like Hypothesis Testing, ANOVA, multivariate statistics, regression, classification, modeling, decision theory, time-series analysis and Descriptive
* statistics.
* Experienced in AWS pipeline, Amazon Sage Maker, Amazon Redshift and Amazon S3 for end-to-end Data Preprocessing, Data Processing, Model Building & Evaluation.
* Experienced with Machine Learning, Regression Analysis, Clustering, Boosting, Classification, Principal Component Analysis and Data Visualization Tools.
* Proficient in Data transformations using log, square-root, reciprocal, cube root, square and complete box-cox transformation depending upon the dataset.
* Concrete mathematical background in Statistics, Probability, Differentiation and Integration, Linear Algebra and Geometry
* Proficient at a wide variety of Data Science programming languages Python SQL, Pyspark, Tableau, Sci-kit Learn, NumPy, SciPy and Pandas.
* Experienced in creating Data Visualizations for KPI's as per the business requirements.
* Good Knowledge in Data Validation, Data Cleaning, Data Verification and Identifying data mismatches.
* Adroit at employing various Data Visualization tools like Tableau, Matplotlib, Seaborn.
* Good understanding of creating Conceptual Data Models, Process/ Data Flow Diagrams, Use Case Diagrams, Class Diagrams and State Diagrams.
* Good communication and presentation skills, willing to learn, adapt to new technologies and third- party products.

TECHNICAL SKILLS AND TOOLS:

|  |  |
| --- | --- |
| Machine Learning | : Classification, Regression, Feature Engineering, Clustering, Neural Networks, Regression analysis, Naive Bayes, Decision Tree, Random Forest, Support Vector Machine, Neural Network, KNN, Ensemble Methods, K-Means Clustering, Natural Language Processing (NLP), Sentiment Analysis, Collaborative Filtering, ML packages, ChatGPT APIs. |
| Cloud | : Azure, AWS |
| Visualization | : Azure ML Studio, AWS Sage Maker, Matplotlib, Power BI |
| Selected Coursework | : Linear Algebra, Multivariate Calculus, Probability and Statistics, Time Series Analysis. |
| Databases | : Azure SQL DB, MySQL, SQL, Azure Cosmos DB, Mango DB, PostgreSQL, MySQL, Amazon Aurora, Amazon RDS, Amazon Dynamo DB, Amazon S3. |
| IDE | : Jupyter-Notebook, Spyder. |
| Big Data | : Azure Data Brick, Azure Data Lake, Azure HD Insights (MapReduce, Spark, Kafka, Azure Synapse Analytics), Amazon Redshift |
| Statistical Analysis | : Time Series Analysis, Regression models, Principal Component Analysis and Dimensionality  Reduction, cluster analysis. |
| Programming Languages  Version Control  Continuous Delivery Service Delivery Methodologies | : Python (Pandas, NumPy, SciPy, Scikit-learn, Seaborn, Matplotlib, NLTK), SQL, Pyspark, and MS Azure, AWS, Docker.  : Git, Subversion (SVN)  : AWS Code Pipeline, AWS Code Commit, AWS Code Build, AWS Code Deploy, Azure DevOps Pipeline, Azure Repos, Azure DevOps Release,  : Agile - Scrum, Agile – Kanban, Waterfall |

PROFESSIONAL EXPERIENCE:

**Data Scientist**

**Client Name: Cervello, Dallas, TX Mar 2022 to Current**

* Responsible for analyzing large sets of data to draw conclusions and make recommendations to the business using statistical models.
* Involved in discussions with business teams to understand the business process mapping and came up with business use cases to improve sales, predictive product market penetration using ML models.
* Worked closely with project managers to develop and expand business use cases ideas for implementations.
* Used XGBoost model to identify the features that predict the sales metrics and KNN algorithm to generate a comparable set of accounts that behave similarly.
* Used AWS S3 bucket to load the data from SQL database. Worked in development of product penetration model enhancements that is deployable and schedulable in AWS sage maker and Amazon Code Pipeline.
* Created End to End ML models based on State rules to handle unique rules around account purchase patterns and regulations.
* Worked in generalizing model to streamline penetration evaluation for a user driven product. (Brands, Sku groups, Product types etc.)
* Lead the evaluation and inclusion of appropriate affinity metrics in penetration based on region and account demographics.
* Deployed 10+ models of product penetration models in AWS cloud platform by implementing diverse Amazon pipelines catering to multiple regions.
* Automated Production Penetration model to run based on new account data and update the model to provide feedback on sales pattern change over time and insight in to market expansion.
* Leveraged ChatGPT API to process audio, video, text files and generated themes, summary, sentiment score and weekly customer sentiment reports for customer service center directors.
* Involved in fast paced & complex demand sensing program to acquire data from 20+ external sources and 25+ internal sources and then cleansed data, scaled features, reduced dimensions then ingesting data sets in into data bricks environment.
* Crafted models to accurately predict (less 7% variance) the demand for upcoming 12 months including safety stock and dead stocks. Derived specific features based on feature importance which contributes to dead stocks and eventually reduction in it.
* Utilized machine-learning algorithms such as Linear Regression, Logistic Regression, KNN, XG Boost for data modelling. * Used Stratified sampling and statistical modeling to identify sample data for accuracy validation.
* Employed Exploratory Analysis and Accuracy Tuning to identify actionable insights to improve efficiency and customer experience.
* Created Data Dictionaries and Data Model for data analysis.
* Wrote and optimized SQL queries to perform data extraction from Big Data platforms.
* * Worked with Directors, Senior Managers-Data and Sr. Data Analysts to understand and gather specific requirements and extract business relevant requirements.
* * Used Hive, SQL for Data Analysis and Data Wrangling &amp; Transformation.
* * Built RDBMS tables in cloud SQL for large datasets pulled from various data sources for insight analysis.

**Environment**: Python, SQL Server, S3, AWS sage maker, NumPy, SciPy, Pandas, Matplotlib, Seaborn, Scikit-learn, AWS

Code Pipeline, AWS Code Commit, AWS Code Build, AWS Code Deploy, Git

**Data Scientist/ML Engineer**

**Client Name: Bank of Oklahoma, Tulsa, Oklahoma Sep 2020 to Feb 2022**

* Responsible for performing Machine-learning techniques regression/classification to predict the outcomes.
* Used Python to get the data in right format, which used by other internal applications to calculate the thresholds.
* Utilized machine-learning algorithms such as logistic regression, multivariate regression, K-means, & Recommendation algorithms for data modelling.
* Involved in team meetings and discussions with business teams to understand the business use cases.
* Worked with Data Engineers and Data Analysts to understand and gather specific requirements and extract business relevant stories.
* Used Python, SQL to create Statistical algorithms involving Multivariate Regression, Linear Regression, Logistic Regression, PCA, Random Forest models, Decision trees, Support Vector Machine for estimating the risks of welfare dependency.
* Used Azure HD Insight for Data analysis to meet the requirements. * Used neural networks to build deep learning CNN models.
* Collaborated with database engineers to implement ETL process, wrote and optimized SQL queries to perform data extraction and merging from SQL server database.
* Worked in creating different visualizations in Tableau using Bar charts, Line charts, Pie charts, Maps, Scatter Plot charts, and Table reports.
* Developed NLP with Deep Learning algorithms for analyzing text improving over their existing dictionary-based approaches
* Creating browser-based visualization tools and dashboards using Tableau
* Using Gitlab repository to clone the code and commit the changes and push it to the develop branch from feature branch. Employed statistical tests such as hypothesis testing, t-test, confidence intervals, for error measurements.
* Built RDBMS tables in cloud SQL for large datasets pulled from various sources using API's to develop NLP machine learning models.
* Data cleaning, Exploratory data analysis, Data Wrangling and preprocessing data using Pandas NumPy and data visualization using
* Pandas, NumPy, seaborn, SciPy, Matplotlib, Seaborn.
* Analyzed data using data visualization tools and reported key features using statistical tools and supervised machine learning techniques such as Naïve Bayes, Decision Trees, Random Forest and Gradient Boosting for predictive analysis using python Scikit- Learn to achieve project objectives.
* Performed Data Wrangling, feature engineering, apply statistical analysis, data visualization, data mining using Python (Pandas, NumPy, Matplotlib, seaborn, Scikit-learn).
* Developed ETL workflow which pushes web server logs to an Azure blob storage.
* Computations by using Azure ML studio and designer Local Computer cluster system with multiple hyper parameters tuning.
* Most of the work is done by programming language like Python, SQL (query language) and command lines.
* Built an automated data preprocessing tool for text data cleaning by using Python and Pandas
* Maintained data warehouse tables through the loading of data and monitored system configurations to ensure data integrity. * Generated reports in case of Decline Claims using Tableau.
* Validated and correlated the collected information using visualizations techniques in Python, Azure ML Studio. * Successfully deployed the developed predictive model for production use through GitHub.
* Monitored Real time analytics Dashboard (State Street's Cloud Development Tool (CDT) after implementation of statistical model that provides the customer in taking a better decision).
* Used HD Insights to analyze datasets and perform transformation according to requirements. * Utilized machine learning techniques to perform ad-hoc predictive analytics.
* Used Optimization Technique Simulated Annealing and Decision Tree ML concepts. Statistical concepts were widely used including Central Limit Theorem, Probability Concept, Probability Distribution (Binomial, and Poisson Distribution).
* Created data visualizations and reports to convey results and analyze data using Tableau and Matplotlib. * Maintained a log of all the iterations performed in Python during the data modeling process.
* Validating and correlating the collected information using visualization techniques. * Responsible in creating CI/CD pipeline for ML ops in Azure.
* Followed data flow path based on Machine Learning platform using logistic regression and decision trees to build credit risk model
* Used statistical techniques for hypothesis testing to validate data and interpretations.
* Carried out Statistical Analysis such as hypothesis testing and Chi-square tests.
* Built models using supervised classification techniques like K-Nearest Neighbor (KNN), Support Vector Machine(SVM),Logistic Regression and Random Forests with Principal component analysis to identify important features.

**Environment**: Python, Azure Synapse Analytics, Azure ML Studio & Designer, Azure Data Bricks, MYSQL, Statistical Analysis, K-means clustering, Regression, Naïve Bayes~~,~~ Tableau, Azure SQL DB, Azure Blob storage, Time Series Forecasting, Azure Kubernetes service, Azure Data factory, Azure ML ops.

**Data Scientist**

**Client Name: Dell Technologies, Round Rock (Austin), Texas Nov 2017 to Aug 2020**

* Performed Exploratory Data Analysis (EDA) to maximize insight into the dataset, detect the outliers and extract important variables numerically and graphically.
* Applied concepts of probability, distribution, and statistical inference on the given dataset to unearth interesting findings using comparison, T-test, F-test, R-squared, P-value etc.
* Analyzed data to identify glitches and cleaning it to reduce the distortions.
* Performed various data manipulation techniques in statistical analysis like missing data imputation, indexing, merging, and sampling.
* Performed Training, Test and Evaluated the model on a cloud Platform Azure ML after building the predictive model. Normalized the data and transformed all the values to a common scale.
* Visualize, interpret, report findings & develop strategic uses of data Using Python, Tableau.
* Built advanced Machine Learning classification models like KNN, SVM and clustering algorithms Hierarchical Clustering.
* Performed univariate and multivariate analysis on the data to identify any underlying pattern in the data and associations between the variables.
* Performed data cleaning and worked with developers to ensure the integrity of the data pipeline.
* Worked with key stakeholders to ensure all telemetry and appropriate KPIs are tracked and analyzed.
* Using NLP techniques identified possible matching words that can provide a base to categorize the protocol of required words * Analyzing of positive, Negative and Neutral sentiments of the consumers to provide a base to the predictive model.
* Responsible for Data Cleaning, features scaling, features engineering by using NumPy and Pandas in Python.
* As a part of the data science team, I have to collect the data from different databases then performed required operations like cleaning, wrangling, table joints, updating data bases and many more.
* Collaborated with database engineers to implement ETL process, wrote and optimized SQL queries to perform data extraction and merging from SQL server database.
* Used K-Means Algorithm Model with different clusters to find meaningful segments on customers and calculated the accuracy of model.
* Built multiple classification algorithms, such as Logistic Regression, Support Vector Machine (SVM), Random Forest, Ada boost and Gradient boosting using Python, Scikit-Learn and evaluated the performance on customer discount optimization.
* Worked on creating both regression and classification models and used advanced techniques like PCA, RFE, and ensembling for feature selection which helped to improve the accuracy of the model.
* Applied multiple Machine Learning (ML) and Data Mining techniques to improve the quality of product ads and personalized recommendations.
* Performed data mining on data using very complex SQL queries and discovered pattern and used extensive SQL for data profiling/analysis to provide guidance in building the data model.
* Responsible for mining large data sets and connected data from different sources in order to identify insights and designs. * Analyzed data and predicted end customer behaviors and product performance by applying machine learning algorithms.
* Implemented various machine learning models for classification- Logistic, SVM, Decision Tree, Random Forest, XGBoost, PCA and selected the best fitting model.
* Created visualization dashboards with tableau for the business users.
* Used Python, SQL to create Statistical algorithms involving Multivariate Regression, Linear Regression, Logistic Regression, PCA, Random Forest models, Decision trees, Support Vector Machine for estimating the risks.
* Involved in creating Data Lake by extracting customer's Big Data from various data sources (from Excel, Flat Files, SQL Server, Mongo DB, and also log data from servers).
* Generated ad-hoc SQL queries to fetch data from SQL Server database systems.
* Standardized the data, populated missing valued and selected relevant features as part of data preprocessing.
* Created dashboards in Tableau desktop based on the data collected from MS-excel and CSV files, with MS SQL server databases.
* Prepared and presented complex written and verbal reports, findings, and presentations by using various visualization tools such as Matplotlib.

**Environment**: Python, Machine Learning, SQL Server, Hive, Azure, NumPy, SciPy, Pandas, MatPlotLib, Seaborn, Scikit-learn, NLTK, Tableau, Docker

**Data Analyst/Engineer**

**Client Name: Tractor Supply Company, Brentwood, Tennessee May 2015 to Oct 2017**

* Compiled data from various sources public and private databases to perform complex analysis and data manipulation for actionable results.
* Worked closely with the Data Governance Office team in assessing the source systems for project deliverables. * Used SQL queries to pull the data from disparate systems and Data warehouse in different environments.
* Worked on data pre-processing and cleaning the data to perform feature engineering and performed data imputation techniques for the missing values in the dataset.
* Creating aggregates, filters, quick filters, calculated measures.
* Designed and deployed reports with drill up and drill down menu options and linked reports using Tableau.
* Performed financial product cost/benefit analysis to support Credit Union senior management in financial product development and management.
* Presented DQ analysis reports and score cards on all the validated data elements and presented- to the business teams and stakeholders.
* Involved in defining the Source to Target data mappings, business rules, data definitions. * Successfully interpreted data to draw conclusions for managerial action and strategy.
* Used statistical techniques for hypothesis testing to validate data and interpretations.
* Collaborated with various business stakeholders to create Business Requirement Document (BRD), translated gathered high-level requirements into a Functional Requirement Document (FRD) to assist implementation side SMEs and developers, along with data flow diagrams, user stories and use cases Part of a Scrum Agile team.
* Applied concepts of probability, distribution, and statistical inference on the given dataset to unearth interesting findings using comparison, T-test, F-test, R-squared, P-value etc.
* Used Excel and Tableau to perform different analysis like descriptive, exploratory, and predictive to get the insights of data more accurately.
* Participated in Data Acquisition with Data Engineer team to extract historical and real-time
* Tested complex ETL mappings and sessions based on business user requirements and business rules to load data from source flat files and RDBMS tables to target tables.
* Prepared and presented data quality report to stakeholders to give understanding of data. * Responsible for data aggregation, data pre-processing, descriptive and inferential analysis.
* Used Data Quality validation techniques to validate Critical Data elements (CDE) and identified various anomalies.
* Performed advanced data extraction, reporting, analysis, and consultative services in order to drive and evaluate consumer and commercial financial product marketing efforts.
* Provided summary expense and performance data to support multiple levels of product management in daily operations management and service delivery optimization and provide other data reporting as required to support special projects and internal auditing functions.

**Environment:** Python, MS Excel, Hadoop, Spark (MLlib), Predictive Analytics, Statistical Analysis System, Azure, Tableau.

**Python/Data base Admin**

**Client Name: Stanford Healthcare, Palo Alto, California.**

**Jan 2014 to Apr 2015**

* Designed the applications using Python, JSON and jQuery. Worked on backend of the application. * Involved in the Design, development, and test and deploy the website.

* Developed entire backend modules using Python on Django Web Framework. * Handled all the client-side validation using JavaScript.

* Implement code in Python to retrieve and manipulate data.

* Responsible for debugging and troubleshooting the web application.

* Developed Unit test cases and performed integration and system testing. * Performed database operations and queries using MySQL.

* Generated property list for every application dynamically using Python.

* Used Python and Django to interface with the jQuery UI and manage the storage and deletion of content.

* Designed and developed components using Python with Django framework. Implemented code in python to retrieve and manipulate data.

* Written Python scripts to parse XML documents and load data in database. * Optimized the database queries to improve the Performance.

* Designed and developed Forms, templates and mapped URLs using Django guidelines. * Rewrite existing Python/Django module to deliver certain format of data.

* Executed various MySQL database queries from Python using MySQL connector and MySQL database package.

**Environment**: Python, MySQL, jQuery, GitHub, Web Services, Unit Test.

Badges

https://[www.credly.com/users/porkalai-duraisamy](http://www.credly.com/users/porkalai-duraisamy)