**SAURABH PATIL**

Somerset, NJ 08873 | 469-994-4684 | [saurabh.arun.patil@gmail.com](mailto:saurabh.arun.patil@gmail.com) | [https://www.linkedin.com/in/patil-saurabh](https://www.linkedin.com/in/kshitija-kulkarni)

**SUMMARY**

Data Scientist with a strong background of Machine learning, Statistics and Computer Science. I have 4+ years of industry experience (Banking & Financial Services- 3 years & Marketing- 1.5 years) along with 2 years of extensive Graduate schoolwork in Business Analytics (concentration: Data Science). Experienced with building Data pipeline, handling large structured and unstructured data sets, Developing analytical products and deriving optimal solutions using state- of- the art **Data science, Software development, machine learning, Natural Language Processing, deep learning tools in R/Python, TensorFlow, KERAS, Hadoop, Hive, Spark, MongoDB, SQL, Tableau,** **d3.js** etc. in creating end-to-end data analytics projects and decoding data to obtain actionable insights in production environment.

**Professional Summary**

* Actively participated in all phases of the project life cycle including **data acquisition (Web Scraping), data cleaning, Data Engineering** (dimensionality reduction- PCA & LDA, normalization, weight of evidence, information value), **feature selection, features scaling & features engineering, Statistical modeling (decision trees, regression models, neural networks, generalized linear models), testing and validation (ROC plot, k-fold cross validation) and data visualization.**
* Experience working full data insight cycle – from discussions with business, understanding business logic and business drivers, **Exploratory Data Analysis, identifying predictors, enriching data, working with missing values, exploring data dynamics, meaning or building predictive data models (if predictability can be found)**
* Excellent data visualization experience either with proprietary code in **R or Python**, or using other visualization tools like **Tableau, D3.js, QlikView, R Shiny**; ready for insight digestion by business and decision making to senior management (Global CTO, Global BI Leadership level)
* Extensive experience using R packages like (GGPLOT2, CARET, DPLYR, STATMODEL, QDAP, TIDYTEXT, RHADOOP, GGMAP)
* Extensive experience using PYTHON packages like (SCIKIT-Learn, NLTK, Matplotlib, Beautiful Soup, TensorFlow, SciPy)
* Experience collating sparse data into single source, working with unstructured data, writing custom data validation scripts
* Hands on experience in implementing **SVM, Naïve Bayes, Logistic Regression, LDA, Decision trees, Random Forests, Ensemble methods, XG-Boost, recursive partitioning (CART), K-Means clustering, NLP, text classification, Topic models**
* Experienced with Big Data Tools like Hadoop (HDFS), Hive, Spark, Elasticsearch & Apache Lucene
* Experience in working with **Data Management** and **Data Governance** based assignments.
* Extensive knowledge in **Data Validation** in **SQL Server,** **Oracle** and **MySQL** by writing **SQL queries.**
* Knowledge of ETL tools such as **Informatica and SSIS** for data integration and processing
* Experience in **Health care Management, Retail** with excellent Domain knowledge in **financial industry’s** financial instruments and financial markets (Capital & Money) and **marketing industry’s** predictive model for personalized marketing, CTR and impressions optimization, defining marketing strategy and presenting recommendations.
* Excellent communication, analytical and, presentation skills; expert at managing multiple projects simultaneously.
* Experience working with **on-shore**, **offshore**, **on-site** and **off-site** individuals and teams.
* Strong understanding of Software Testing Techniques especially those performed or supervised by BA including **Black Box testing, Regression testing,** and **UAT**.

**BUSINESS EXPERIENCE**

|  |  |  |
| --- | --- | --- |
| **1. Algorithm Ventures Inc, Somerset NJ** | **Data Scientist** | **May 2017- Present** |

**Audience Street**

A personalized marketing platform that analyzes millions of ad requests, clicks and conversions to build persona about unique audiences which includes demographics, geographical, psychographics, behavioral and In-market preferences of everyone. This platform then runs ad campaign on these highly targeted audience to achieve optimum conversions.

**Responsibilities:**

* Developed Python batch script (Spark, Pandas, NumPy, NLTK, etc.) to clean 10 million+ records for audience clickstream and real-time bidding (RTB) data which are streaming each day on Amazon AWS EC2 Hadoop servers
* Performed web scraping for URLs in each record and created text corpus with term frequency for each record
* Analyzed each record to predict audience demographics (age, gender); geo-tagging (work & office location); and behavioral profiles (interests, purchasing habits) by making several models in R and Python
* Formulated the machine learning algorithms for age and gender predictions from the ground up in R programming (JSONLITE, DPLYR, NLP, tidy-text, QDAP, GGMAP, Google API) using NLP, segmentation & feature engineering
* Implemented sentimental analysis in Python (Pandas, NumPy, SciPy, NLTK, SCIKIT-Learn) which implies the audience interests, introvert/extrovert behavior & purchasing habits using Latent Dirichlet Algorithm and clustering
* Constructed proprietary DMP with above audience attributes for personalized marketing & campaigns creation
* Charted AB testing plan using two dummy campaigns and evaluated using number of conversions/purchase yields
* Resultant model showed nearly $50,000 reduction in marketing costs for several clients in Singapore & Europe

|  |  |  |
| --- | --- | --- |
| **2. TCS : ABN AMRO Bank, Amsterdam & Mumbai** | **Business Data Analyst /Data Scientist** | **Jan 2014 – Jul 2016** |

**ABN AMRO Mortgages**

ABN AMRO Bank NV is the third-largest bank in Netherlands which has over 3 million mortgage customers and 20 mortgage products. The annual revenue from ABN AMRO Mortgages is around €274 million euros. Hence, reducing defaulters using risk models with the help of machine learning approaches played a pivotal role in annual savings for the bank. This objective was achieved through solutions (risk modeling for mortgage products, carefree living for customers).

**Responsibilities:**

* Analyzed risk and mortgages data of customers using Python & segmented the customers on 1-5 risk levels
* Forecasted complex product risks using time series regression using ARIMA model & neural network models
* Designed testing plan for impact analysis and used statistical inference with hypothesis to validate these plans
* Presented the findings using interactive dashboards in Tableau 10.2 to various stakeholders in the bank
* Optimized data collection procedures & generated reports for Risk domain which saved $200,000+ to the company
* Gathered requirements from business stakeholders & functional team for impact analysis & effort estimations
* Developed ETL mappings of business requirements in Informatica & SQL queries for Mortgages, designed dashboards for quantifiable metrics of different bank’s products during agile monthly production life cycles
* Reported interesting findings on carefree living roadmap for different customer segments based on their financial status and other liabilities using machine learning models in Python
* These risk models for bank and carefree living roadmaps for bank’s customers help bank save €20 million euros annually due to reduction in mortgage defaulters by 8% during fiscal year 2015.
* **ACHIEVEMENTS-** Young Achiever’s Award, Special Initiative Award, Best Team Award

**EDUCATION**

**University of Texas at Dallas** May 2018  
*M.S., Business Analytics (Concentration: Data Science)*

**University of Mumbai** June 2013  
*B.E., Electronics Engineering*

**TECHNICAL SKILLS**

|  |  |
| --- | --- |
| **Databases** | Hadoop (HDFS), Hive, Spark, Oracle, MS SQL, MySQL, Elasticsearch, Lucene |
| **Languages** | Python, R, SAS, SPSS, Java, HTML, JavaScript, VB Script, SOA (XML) |
| **Data Analytics & BI Tools** | Tableau 10.5, D3.js, QlikView, Power BI, Informatica, SSIS, Git |
| **Packages** | Pandas, NumPy, SciPy SCKIT-Learn, Beautiful SOUP, Matplotlib, GGPLOT2, CARET, DPLYR, RSHINY, STATMODEL, GLM, GGMCMC, QDAP, RHADOOP etc. |
| **Documentation Tools** | SharePoint 2013, MS Office (Word/Excel/Power Point/Access/Outlook) |

**COURSEWORK**

Data management, Data mining, Predictive analytics, Machine learning, Data visualization, Regression, Classification, Clustering, Ensemble trees, Neural networks, Data modelling, Time series analysis & Econometrics, Operations research