Vivek Mohan Singh

Associate Analyst

Gramener
A Data Science Company

PROFILE SUMMARY

- A Data Science post-graduate with knowledge of Exploratory Data Analysis, Statistics, Machine Learning, Natural Language Processing and Big-Data(HDFS, EC2, EMR using Hive and PySpark)
- Experience of working on a project end-to-end: Problem scoping, data gathering, delivering advanced analytics solutions and insights.
- Exposure to handling variety of datasets from different domains including Stock Markets, Twitter tweets, Loan Applications, Unemployment Data, Elections data etc.
- Worked as a lease administration analyst at NTT DATA IPS for client JLL Inc., having proficient experience in Predictive Modelling, Hypothesis testing, Big Data (HDFS, Spark, Hive) and related visualization using BI tool Tableau, for analytics driven Portfolio management.

EDUCATION

PG-D in Data Science – (3.4/4.0)
International Institute of Information Technology - Bangalore (2017-2018)

Bachelor of Technology in Computer Science & Engineering – (7.0/10.0)
Jaypee University of Engineering and Technology (JUET) (2013-2017)

CERTIFICATIONS & TRAININGS

Learning Python for Data Analysis and Visualization – Udemy
R Programming: Advanced Analytics in R – Udemy
Spark and Python for Big Data with PySpark– Udemy
Data Analysis with Python and Pandas – Udemy
Cluster Analysis on Distributed Data using AWS EC2 & EMR – Udemy
Application Development with IBM Watson Bluemix – Udemy
JSON Web App developer API & Java Object Literal – Udemy
Building Responsive Data Visualizations with D3.js - Udemy
Advanced DAX and related visualizations using PowerBI – Udemy

TECHNICAL SKILLS

- Proficient experience in R and Python.
- Machine Learning: Classification (linear and nonlinear separable data), Regression, Clustering, Scalar Vector Machines and Decision Tree Ensembles.
- Big Data – HFDS, EMR using Hive, PySpark, EC2
- Exploratory Data Analysis- (Tidyverse and Pandas)
- Time Series Analysis(ARIMA and Auto ARIMA)
- Data Visualization (Matplotlib, seaborn, ggplot2, D3.js)
- Web Scrapping- (Beautiful Soup and Rvest)

NON-TECHNICAL SKILLS

- Analytical Problem solving
- Curiosity Driven
- Automation
- Team Player
- Communication skills
- Creative (Art and Design)

TOOLS

- R
- Python
- PySpark, Scala
- Tableau, PowerBI, Excel
- SQL, Hive
- AWS: EC2, EMR, RDS, S3

WORK EXPERIENCE

Associate Analyst - GRAMENER, Bengaluru |JUN 2019 – Present

- Worked with Dimension Data as an analyst, for their IT consulting BU.
- Integrated the existing Analytics backend from R to Python3.
- Further automated the quarterly data refresh to a one click approach on PowerBI, the dashboard get automatically refreshed by running python code on PowerBI which in turn is synced with company server for ingesting fresh data for every quarter.
- Explored and analysed the data for significant scope in cross-sell across region and up-sell within the same geographical regions. Accordingly offered them success plays to target.

Sr. Associate Consultant (Analytics) - NTT DATA Information Processing Services, Noida/Bengaluru |Aug 2017- Apr 2019

- Managing clients as a Lease Administration Analyst at NTT DATA IPS.
- Providing full administrative support for Marketing Officer for Leasing and Portfolio administration.
- Analysing and forecasting Interest rates using predictive modelling and marketing mix on historical data.
- Preparing and distributing critical date reports; maintains tickler system for tracking important dates and tasks.
- Conducting financial analysis/accounting as appropriate, including AR/AP, rent rolls, and budgets in accordance with required accounting principles (GAAP, Tax, Cash, etc.)
- Abstracting and interpreting commercial lease documents.
- Reporting by deploying Tableau preparing storyboards, and related visualization.

UI-UX Developer- AICL Communications, Mumbai | June 2017 -Aug 2017

- Designed and Developed Web UI, Content Management System, mocks, wireframes, and flat designs for
- Digitalization of annual reports of various top tier companies such as Hero, Reliance Industries Ltd, JSW Steel etc.
- Developed multi stack websites from static HTML/CSS to Plugin driven PHP content management systems.
- Designed and implemented PHP and D3.js driven data plots both static and dynamic for report's tables and plots.

KEY PROJECTS

Uber Supply Demand Gap:

Analysed the problem of cab availability to and from airport i.e. Problem of cancellation of cars by the driver or non-availability of cars. Analysis was done on 6lacs+ customers' data provided by Uber by cleaning the raw data in R, segmenting & Bucketing into time specific metrics and separately visualizing Univariate and Bivariate variables board and into meaningful plots, story correlation matrix in Tableau, for a deeper understanding of the cause behind the Gap.

Duration 15 Days

Team Size 1

Technology/Tools R, Tableau

Activities

- Data Attributes: Req. ID, TimeOfReq, Pickup-time, Drop-off time, Driver ID, Status (completed, cancelled, unavailable)
- Data Cleaning and Date Formatting request time-stamp, pick-up and drop timestamp (POSIXCT)
- **Segmentation:** Based on time span of Demand/request time-stamp segregated the data into pre-morning, morning-rush, noon, evening, pre-night rush.
- **Feature Selection** – PCA, Gini Index for homogeneity.
- Various bivariate analysis done to analyse significant predictor for Supply and Demand.
- Analysed each of the segments and their supply demand curve.
- Derived the cost function for surge-up and surge-down during heavy demand hours by assessing supply demand curve at various time specific segments.
- As an output surge up and surge down cost function was derived for every time segment also suggestions such as drivers should be incentivised for toward airport journey as its very likely that they do not get a trip back to city

Risk Analysis of loans:

Did this analysis for a consumer finance company, analysed two major Risks for the bank:
i) If the applicant is **likely to repay the loan**, then not approving the loan results in a **loss of business** to the company
ii) If the applicant is **not likely to repay the loan,** i.e. he/she is likely to default, then approving the loan may lead to **financial loss** for the company.
Performed Exploratory Data Analysis to understand the **driving factors (or driver variables)** behind loan default. The company utilized this knowledge for its portfolio and risk assessment.

Duration 1 week

Team Size 1

Technology/Tools R/PowerBI

Activities

- Followed lenders buyer's club model. For generating the risk score for an applicant.
- Data Attributes: consumer attributes (applicant and demographics) and loan attributes (300+ features)
- Principle Component Analysis to recognise significant features.
- Used the Monte Carlo method (Markov Process) to run many trials (Bagging and boosting) with random market conditions as an ensemble decision tree, then

calculated portfolio losses for each trial. After this, used the aggregation of all these simulations to establish how risky an applicant is. And assign a score.

- Output: Dashboard visualising various important insights such as, Defaults rates for loan types (Debt consolidation, credit card loans, personal loans), Default rates by loan purpose, top 4 products over time, Term distribution, Avg. interest rates in top 4 products.
- Important default predictors were identified across Debt Consolidation, Home Improvement and Credit Card loan types.

Time-Series Analysis forecast for Sales and Demand:

Performed Time Series analysis for "**Global Mart**", which is an
Online store super giant having worldwide operations. The store caters 7 different market segments in 3 major categories.

Duration	1 week
Team Size	1
Technology/Tools	R

Activities

- The data was segmented into 21(7*3) unique market segment buckets and out of these major 2 were taken for analysis using coefficient of covariance among top 21 segments.
- The data was then mapped to Time series format in R, the model was built and forecasted using classical decomposition and auto-ARIMA to predict sales and demand for next 6 months. The forecast was then evaluated with MAPE.
- The most profitable segments for sales and demand were **Europe-Consumer** and **Asia Pacific-Consumer**.

Multiple Logistic Regression Analysis:

Undertook this project for an HR of a company to predict the cause behind the attrition of employees. The company has maintained a database containing employee/manager survey information,
The general data about the employee and in-time and out-time for 261 days for about 4400 employees. Analysis was done by segmenting the metadata into buckets of categorical independent variables as:
1. Employee Survey
2. Manager Survey
3. General Data
4. Working hour information using In-Data & Out-Data

After data preparation and EDA, dummies were created for categorical data and the final data was scaled for model building. A total of 210 logistic models were built, predicted and cross evaluated on test data. The predicted data was 89.0% accurate with train data and gave 11 significant driver variables for the cause behind attrition of employees.

Duration	2 weeks
Team Size	1
Technology/Tools	R

Web Scrapping and Data Analysis for US General Elections (Trump vs. Clinton)

Web Scrapping to get exit polls data and donor data set from Democratic and Republican campaigns for a dashboard and perform EDA to get insightful stories.

Duration 2 weeks

Team Size 1

Technology/Tools Python3

Activities

- Data Scrapping from Huffington post using beautiful soup, Python
- Data cleaning - Pandas
- Descriptive Analytics – insights and stories from candidates' data
- Analysed the cause behind Trump surprising victory and major predictors
  For the same, for e.g. Hispanic voters, Afro-Americans and Undecided voters effects.
- Did a time series analysis around debates dates and analysed its effect on exit polls.
- The Donor Data set tells the contributions to Republican Party as a strong driving variable for Trump's victory.

ETL Automation for ticketing support system at NTT DATA for JLL:

Successfully implemented ETL Automation for lease data import for our company product JLL Portfolio. Automation incorporates Leverton API along with Text Analytics using Python.

Duration 6 month

Team Size 6

Technology/Tools Python, Leverton Robocon API, ADex, tSQLt

Activities

- Text Recognition using Leverton Robocon API and its automated load to region specific buckets backend using python NLTK libraries which recognises a lease document based on distinct currency ISO codes.
- Data Attributes: Lease Data for properties across 7 regions
- tSQLt an SQL server unit testing framework was utilised to perform assertive/schema and quality checks in the imported regional lease data.
- Build a web app using .net core MVC which enables the client to download one of many support ticketing templates to raise the issue, the template is then digitally red Leverton API and required changes are implemented using ADex(contract intelligence platform powered by AI)

Sentiment Analysis of ABC news headlines:

Analysed 15 years data from ABC news channel for headline sentiments.

Duration 1 week

Team Size 2

Technology/Tools R, tidy-text, Excel, lexicons, tidyR

Activities

- Identification word frequencies and trained the lexicons accordingly.
- Calculated the sentiment score for each year 2003 -16 using tidyR.
- Identified the most common Negative/Positive sentiment in the past 10 years.
- Build plot to gauge contribution of top 10 words (+ve and -ve) in Bing lexicon score when iterated over last 10 years data.

Annual Report HERO MOTO CORP 2015-16(website):

Designed and developed the annual report website for HERO MOTO CORP 2015-16. The Website is broadly built to succinctly visualize statistics and non-statistic part of the report as a digital media for customers and shareholders. It's powered by various J-Query plugins, uses bootstrap as a frontend framework & PHP-PDO for handling database.

Website-Link

Market Mix Modelling – (Dissertation Project IIIT-B)

Developed a market mix model for Eleckart (an e-commerce firm specialising in electronic products) to observe the actual impact of different marketing variables over the last year. Using the understanding of the model, recommended the optimal budget allocation for different marketing levers for the next year.

Duration	2 months
Team Size	2
Technology/Tools	R,Hive, tidy-text, Excel Tableau, AWS RDS, EC2
Activities	• Objective was to identify the most impactful channels for marketing spends • Identified the marketing KPIs for Pricing, Discount, Promotions and Marketing and built models around such KPIs such as multiplicative, linear and Kyocks model to better predict sales after tuning the hyper parameters for these KPI. • All the modelling was done as an AWS EMR Cluster, getting real time distributed data. • Analysed pricing model and optimised (using 4 force model) pricing ladder for minimising cannibalization of products and products to be priced in a manner to reduce effect of cannibalisation. • Analysed shelf price inflation. • Also did analysis on competitive benchmarking for similiar products and market share. • Also visualized the market share funnel for Eleckarts' products. • Also built a Recommendation Engine- utilized R's recommender lab package to build an item based collaborative filter (IBCF) by converting the KPI for similar products into a real rating matrix, to generate recommendations for similar items when viewed.