Curriculum Vitae

Lavajiit Singh

Mobile : +91-7775859367
Address **:** Delhi NCR
E-Mail : lavajiit@yandex.com

PROFESSIONAL SUMMARY

- Data Science intern with 1+ year of experience comprising Data Science, Data Engineering and Full Stack Development.
- Built company's first working ML model that beat baseline accuracy.
- Improved the data engineering process time by 30% and model's accuracy by 2% in PySpark.
- A quick learner and self-motivated individual, who adapts as well as challenges status quo.

EXPERIENCE

A) **Data Science Intern – SIPI-IP, Noida – (4+ months)** ***from September 2018 till*** *date*

a) Understand the data obtained from research teams and then further conduct comprehensive **data cleaning** and meticulous **EDA**.
b) **Formulate experiments** to check the possibility and usability of the data obtained from all teams.
c) **Regularly interact** with research teams to give and receive suggestions and to better understand the business process cycle.
d) **Beat the baseline accuracy** for one team. Currently in the process of doing the same for all teams and improve the accuracy gradually.

B) Data Engineering Intern - RedCarpetUp, Delhi – 2 months ***(****from* ***July 2018*** *till* ***August 2018****)*

e) Carry out **ETL process** with raw data **extraction** from csv files, then **cleaning** it and **transforming** it for **machine learning model** (*fraud detection model*) and **loading** into PostgreSQL database. Both ETL and ML modelling were done in PySpark's dataFrame and sparkML libraries respectively.
f) Load the data from PostgreSQL and build the machine learning model for **fraud detection**.
g) **Actively communicate** with fraud review team, KYC team and web developers for **continuous refining** of whole data science pipeline.
h) **Increased ML model's accuracy by 2%** and **reduced ETL processing time by 30%** using Spark's best practices and regular communication with the teams.

C) Full Stack Development Intern - Karmaa Lab, Bangalore – 7 months ***(****from* ***June 2017*** *till* ***December 2017****)*

a) **Built** the **complete web project** from scratch in Django.
b) **Added the web app** for website **article summariser** using web scraping using BeautifulSoup and NLTK.
c) **Added geo-fencing capability** using google-map APIs in Python and JavaScript.
d) **Deployed** the **project** on AWS EC2.
e) Conducted **daily scrums**.
f) **Assisted** in new interns' **recruitment**.

PROJECTS

A) Credit Risk Modelling

a) The objective of the project was to **predict whether to clear the loan** for the new customer or not based on the historical data of 800 customers. Firstly, without looking at the data, **noted down the hypothesis** i.e. questions on what can increase/decrease the chances of loan getting approved. **Using pandas, matplotlib and seaborn**, **conducted univariate & bivariate data analysis** to find out the answers of the questions popped up from the hypothesis. EDA further generated some questions as well as answered the hypothesis. **Converted the categorical features into numerical features and then correlation was tested**.

b) The **missing values** were **imputed** and **outliers** were **treated**. The **skewness was corrected** using log transform.

c) As it was a classification problem, ML modelling was started with **logistic regression** with **stratified k-fold cross validation**. During **feature engineering**, three new features were derived which represented the data in a better way and improved the model performance. Then **Decision Tree** and **Random Forest and XGBOOST** were also tested and random forest turned out to be the most accurate.

B) MNIST Handwritten Digit Recognition

a) Dataset contain digits 0-9 with a total of 60,000 training samples and 10,000 test samples. Each sample image is 28x28 and linearized as a vector of size 1x784. So, the training and test datasets are 2-d vectors of size 60000x784 and 10000x784 respectively.

b) The modelling was done using ANN using **tensorflow.**

c) **Accuracy** achieved was **97%.**

C) Linear Regression and Logistic Regression from scratch

The motivation behind these projects was to **apply understanding of the algorithms** and to use simple dataset for them. The main objective was not to use any machine learning library. The modelling was carried out in the lines of Andrew Ng's ML course on Coursera and using *numpy*. Further, it was done using **scikit-learn** too.

D) Belgian Traffic Sign Image Classification

The dataset is a collection of 62 different types of traffic signs used on Belgian roads in 62 sub folders in training and test folders. The images are in .ppm format and are read from the subfolders and converted into *numpy* arrays. Images are labelled as the subfolder names. **Feature extraction** was needed as they had different sizes. So, images were **rescaled** to 28x28 and then converted to grayscale. An ANN was built using tensorflow after flattening the input image. Then a fully connected layer of logits was constructed. The model gave the accuracy of 61% which call for further improving the model by adding hidden layers in ANN or using CNN.

E) Cats & Dogs Image Classification

The problem is similar to Belgian Traffic sign problem. Here we just have 2 classes of images so it's a **binary** classification problem. The motivation was to use **keras** and **CNN** for modelling. Due to RAM limitation only one epoch was completed which gave accuracy of 71% which gives appositive sign that CNN should have given much better accuracy after completing the set 20 epochs. One interesting finding about keras capability was its intuitive image processing method: **ImageDataGenerator** which loads the images from the folders and creates training and test set with the functionality of resizing the images, normalizing and assigning classification type on the fly. It makes building the model in few lines.

SKILLS, TOOLS & HANDS-ON

- **Python** (**P**ySpark, **P**andas, Numpy, **M**atplotlib, **S**cikit-learn, **T**ensorflow, **K**eras, **D**jango, **G**oogle Map API)
- **SQL** (**M**ySQL, **P**ostgreSQL), Hands-on **NoSQL** (**M**ongoDB).
- **Linux, AWS EC2.**
- Quick **Learner** and always look for **problem solving**.
- Excellent **Communication** and **Organizational** skills.

WEB LINKS

Github : https://github.com/lavajiit
LinkedIn : https://linkedin.com/in/lavajiitsingh

MOOC's

- **Machine Learning by Stanford** University - Coursera
- Machine Learning A-Z™ **Hands-On Python & R In Data Science** - Udemy
- Machine Learning Crash Course - Google
- **Spark** and **Python** For **Big Bata** With PySpark - Udemy
- Foundations of **Data Analysis** — Part 1: **Statistics Using R** - edX
- Foundations of **Data Analysis** — Part 1: **Inferential Statistics** - edX
- **Django 2** & Python, The Ultimate Web Development Bootcamp - Udemy

EDUCATION

2014-2016 **M.Tech**, Defence Institute of Advanced Technology (DU), Pune
2008-2012 **B.Tech**, MDU.

Reference(s)

Available on request.