Ammar Alyousfi

Kuala Lumpur, Malaysia • ammar5[redacted]@gmail.com • +6011620[redacted]

Website & blog: ammar-alyousfi.com
Github: ammar1y • Kaggle: ammar111
in: ammar-alyousfi • SO: 2282785

Education

University of Malaya — Kuala Lumpur, Malaysia
Master of Data Science degree with distinction (GPA: 3.98/4) — Feb 2018 - Oct 2019

- Research project: built and deployed a **machine-learning model** to predict taxi-trip duration before the trip starts. The process involved preparation, cleaning, and in-depth **exploratory data analysis** of taxi-trips data; it also involved feature engineering, hyperparameter optimization, cross-validation, PCA dimensionality reduction, and ensemble methods. The final model was formed by multi-layer stacking of different models and then it was **deployed** as a Flask web application. Tools used include Python, Pandas, Matplotlib, Seaborn, Scikit-learn, XGBoost, and Jupyter Notebook. Project code and report are available on http://bit.ly/ds-prj
- Relevant courses: Data Analytics, Data Mining, Machine Learning for Data Science, Programming for Data Science, Big Data Application and Analytics, Big Data Management.

Princess Sumaya University for Technology — Amman, Jordan
B. Sc. degree in Computer Engineering; ranked first (GPA: 93.9%) — Sep 2012 - Sep 2016

- Relevant courses: Data Structures and Introduction to Algorithms, Database Systems, Visual Programming, Discrete Mathematics, Computer Architecture and Organization, Operating Systems.

Certifications

- Data Analysis with Python—IBM (2019)
- Databases and SQL for Data Science—IBM (2019)
- Data Science Methodology—IBM (2019)
- Machine Learning with Python—IBM (2019)
- Data Visualization with Python—IBM (2019)
- Open Source tools for Data Science—IBM (2019)

Experience

Apigate — Kuala Lumpur, Malaysia
Data Science and Analytics Intern — April 2019 - July 2019

- Performed **extensive data analysis** on millions of records of company data stored in Google BigQuery warehouse to get useful **business insights**. Python, SQL, Pandas, and Matplotlib were used to analyze the data and produce compelling visualizations, tables, and **dashboards** that were used to present the analytics results to the management and then to the CEO.
- Created an **ETL** Python script to automatically transfer and **sync** company data from Salesforce to the data warehouse of the company on Google BigQuery. The program was scheduled to run weekly and it used API services to access Salesforce and BigQuery data.

Skills

Programming Languages: Python, R, JavaScript.

Data Analysis and Machine Learning: Pandas, NumPy, Scikit-learn, TensorFlow/Keras, XGBoost, LightGBM, etc. Also familiar with SAS and Hadoop ecosystem.

Data Visualization and Dashboards: Matplotlib, Seaborn, Google Data Studio, Follium for maps, etc.

Databases: SQL. **Web Scraping:** BeautifulSoup, Selenium, HTTrack.

NLP: NLTK. **Web Development:** Flask, Django, HTML, CSS, JavaScript.

Cloud Computing: Google Cloud Platform (BigQuery for big data, Compute Engine, and Storage), Amazon Web Services (EC2, S3, and Route 53).

Languages: English: fluent (TOEFL iBT: 102). Arabic: native.

Projects

End-to-end data-science projects: In addition to the master-degree project mentioned above, a project was conducted to build a machine-learning model to predict house prices based on many characteristics like house size, construction year, etc. The project data was prepared and cleaned before applying exploratory data analysis. Then multiple models were built and compared including Linear Regression, K Nearest Neighbors, Support Vector Machines, Neural Network, Random Forest, and Gradient Tree Boosting (XGBoost). Project report and code can be found on `http://bit.ly/hp-pdf`.

YouTube Trending Videos Analysis: data of 40,000+ videos was analyzed and explored to get insights on YouTube trending videos and to identify the common characteristics among them. Toward that end, informative visualizations and tables were generated using Python, Pandas, Seaborn, and other tools. The Jupyter Notebook that contains the analysis code and results can be accessed on `http://bit.ly/YT-analysis`.

Kaggle Competitions: Participated in many Kaggle machine-learning competitions with regression and classification tasks. Some of them are:

- **Help Navigate Robots:** In this multi-class classification problem, participants were asked to detect the type of surface the robots are standing on using time-series data collected from IMU sensors. Used Python, LightGBM for modeling, and Tsfresh package to extract features from the time-series data. Ranked in the top **5%** among 1470+ competitors. Code used and results: `http://bit.ly/help-robo`.
- **VSB Power Line Fault Detection:** In this binary classification problem, participants were asked to predict whether signals acquired from power lines have partial discharge patterns. Used Python, Pandas, NumPy, and SciPy for data pre-processing and feature extraction. For modeling, the stacking ensemble-method was used with multiple models including XGBoost, LightGBM, Scikit-learn Neural Network, and Scikit-learn Logistic Regression. In this competition, the best score was 0.71899, the worst was -0.28109, and my score was 0.58655. Solution explanation and code used: `http://bit.ly/vsb-pl`.

Pair & Compare: A web application that makes it easier to compare fonts and font-pairs. It allows using all 800+ Google fonts without downloading or installing any of them. It was built using HTML, CSS, JavaScript, Vue.js, etc. It can be visited on `http://bit.ly/p-and-c`.

Focus Phase: An open-source time-tracking command-line application with statistics and visualizations. It is built using Python and published on the Python Package Index. Github link: `http://bit.ly/focus-phase`.

S3upload: an open-source Python application that makes it faster to upload a large number of files to AWS S3. Github link: `http://bit.ly/s3upload`.