Pengchong Tang

Rockville, MD 20852 Email: ferrarisf50@gmail.com Phone: (269) 365-6057 Github Linkedin Resume

SUMMARY

5+ years of solid working experience on SAS programming.
3+ years of substantial hands-on data wrangling, data analysis experience with Python and R.
1+ years of data engineering experience with SQL, NoSQL data warehouse and data lake.
Master of Science in Statistics and Bachelor of Science in Math.

SKILLS

Programming languages:	SAS, R, Python, MATLAB, Scala, VBA, Shell Script
Data science & big data tools:	Hadoop, PySpark, AWS EMR, Alteryx, Apache Airflow
Version control systems:	Git
Databases & platforms:	Oracle, SQL Server, PostgreSQL, Cassandra, MongoDB, AWS Redshift, Linux
Visualization tools:	ggplot2, Plotly, Tableau, Matplotlib, Seaborn
Technology knowledge:	Machine learning, Data Analysis, Data Mining, Data Wrangling, Predictive Modeling, Statistics, Natural Language Processing (NLP), Operations Research, Experimental Design

EXPERIENCE

Eurofins Lancaster Laboratories / SAS programmer/Scientist II at Kalamazoo, MI **11/2014 – 12/2019**

- Support clients with animal health data management, design SAS/Python/SQL scripts to build data mappings, transcribe data from multiple sources (Internal, EDC, lab result) and various formats (Excel, XML, SAS, SQL Server, flat file) into SAS tables, manipulate, merge, concatenate, transpose and clean SAS tables using PROCs and DATA steps, automate 60% of the mapping pipelines by implementing machine learning algorithms (tf-idf) to classify lab spreadsheet data. (Main skills: SAS/MACRO, SAS/BASE, SAS/SQL, Scikit-learn, NLP, Random forest, SQL Server)
- Define data validation rules, revamp the old procedure by designing SAS and Python programs to detect the data issue systematically during the mappings, investigate data issues, perform ad hoc data analysis and report findings to the clients. (Main skills: PROC SQL, PROC SUMMERY, PROC FREQ, DATA STEP, etc.)
- Design and develop a Python dashboard, which utilizes packages such as Selenium, PyODBC, and Win32com.client to automate repeated data management tasks, including email management, classifying mass PDF files and web forms filling, successfully reducing labor hours for clients by 80%. (Main skills: VBA, Python, Web scraping)

PROJECTS

I-94 Immigration Data Lake **03/2020**

- Gathered 40 million US I-94 immigration records and airline IATA code dataset, stored the data on AWS S3.
- Built an ETL pipeline using PySpark, extracted data from S3, transformed data using Spark SQL, loaded data back into S3 as a set of dimensional tables.
- Deployed the ETL pipeline onto AWS EMR using Boto3.

NYC Taxi Trip Duration **01/2018**

- Wrangled 1.4 million NYC taxi trip routes information during the first half of 2016 from OSRM, assessed and cleaned the data using Pandas/Numpy, created data visualization dashboard with Tableau, Matplotlib and Seaborn, performed data analysis using Spark on AWS EMR.
- Implemented machine learning algorithms (Xgboost) using Scikit-learn to build a model to predict taxi trip duration, scoring LB 0.38.

Credit Card Fraud Detection **11/2017**

- Explored and analyzed 280,000 credit card transactions using R ggplot2 and reported findings using R Markdown.
- Identified transaction patterns and created univariate, bivariate and multivariate EDA reports.
- Built a classification model using Scikit-learn in Python. The model achieves 0.8 score on average on AUPRC, can

detect about 80% of frauds.

EDUCATION

- MS in Statistics, George Washington University, Washington DC — **09/2010 – 06/2012**
- BS in Mathematics, Sun Yat-sen University, Guangzhou, China — **08/2006 – 06/2010**

CERTIFICATIONS

- Udacity Data Engineering Nanodegree Program — **02/2020 – 03/2020**
- Udacity Data Analyst Nanodegree Program — **11/2017 – 01/2018**
- Coursera Data Visualization and Communication with Tableau — **11/2017**
- Coursera R Programming — **03/2015**
- SAS Certified Base Programmer for SAS 9 — **02/2014**
- SAS Certified Advanced Programmer for SAS 9 — **11/2012**