Eric Ji

Data Scientist - Capital One

Stony Brook, NY - Email me on Indeed: indeed.com/r/Eric-Ji/18f583849fdd2ac4

● Professional qualified Data Scientist/Data Analyst with over 6 years of experience in Data Science and Analytics including Machine Learning, Data Mining and Statistical Analysis
● Involved in the entire data science project life cycle and actively involved in all the phases including data extraction, data cleaning, statistical modeling and data visualization with large data sets of structured and unstructured data
● Experienced with machine learning algorithm such as logistic regression, random forest, XGboost, KNN, SVM, neural network, linear regression, lasso regression and k-means
● Implemented Bagging and Boosting to enhance the model performance.
● Strong skills in statistical methodologies such as A/B test, experiment design, hypothesis test, ANOVA
● Extensively worked on Python 3.5/2.7 (Numpy, Pandas, Matplotlib, NLTK and Scikit-learn)
● Experience in implementing data analysis with various analytic tools, such as Anaconda 4.0 Jupiter Notebook 4.X, R 3.0(ggplot2, Caret, dplyr) and Excel […]
● Solid ability to write and optimize diverse SQL queries, working knowledge of RDBMS like SQL Server 2008, NoSql databases like MongoDB 3.2
● Strong experience in Big Data technologies like Spark 1.6, Sparksql, pySpark, Hadoop 2.X, HDFS, Hive 1.X
● Experience in visualization tools like, Tableau 9.X, 10.X for creating dashboards
● Excellent understanding Agile and Scrum development methodology
● Used the version control tools like Git 2.X
● Passionate about gleaning insightful information from massive data assets and developing a culture of sound, data-driven decision making
● Ability to maintain a fun, casual, professional and productive team atmosphere

WORK EXPERIENCE

Data Scientist

Capital One - New York, NY - July 2016 to Present

Credit Card Fraud Detection
Capital One Financial Corporation is an American bank holding company specializing in credit cards, home loans, auto loans, banking and savings products.

The purpose of this project was to fight against credit card fraud. My team mainly focused on rebuilding credit card fraud detection model, monitoring the model in production, taking action if model performance degrades and working closely with business team to onboard new model.

Responsibilities:
● Communicated and coordinated with other departments to collection business requirement
● Worked on miss value imputation, outliers identification with statistical methodologies using Pandas, Numpy
● Participated in features engineering such as feature creating, feature scaling and One-Hot encoding with Scikit-learn

● Tackled highly imbalanced Fraud dataset using undersampling with ensemble methods, oversampling with SMOTE and cost sensitive algorithms
● Improved fraud prediction performance by using random forest and gradient boosting for feature selection with Python Scikit-learn
● Implemented machine learning model (logistic regression, XGboost) with Python Scikit- learn
● Optimized algorithm with stochastic gradient descent algorithm
● Fine-tuned the algorithm parameter with manual tuning and automated tuning such as Bayesian Optimization
● Validated and select models using k-fold cross validation, confusion matrices and worked on optimizing models for high recall rate
● Implemented Ensemble Models with majority votes to enhance the efficiency and performance
● Designed rich data visualizations with Tableau 9.4

Environment:
Python 3.5 (Numpy. Pandas, Scikit-learn), SQL server 2014, Jupyter Notebook 4.3

Data Scientist

New York Life Insurance - New York, NY - May 2015 to June 2016

New Business Premium Optimization
New York Life Insurance Company (NYLIC) is the largest mutual life-insurance company in the United States, and one of the largest life insurers in the world, ranking #61 on the 2016 Fortune 500 list, with about $550 billion in total assets under management.

In this project, my team (Advanced Modeling team) was responsible to build a predictive model that can help underwriters to identify potential new business case, improve the Submission-Bound ratio and Hit ratio using machine learning algorithms. The final model helped increase around 8% New Business Premium.
Responsibilities:
● Collected historical data and third party data from different data source with Big Data tools
● Conducted data exploratory analysis to identify where the P&C insurer has been successful in writing new business in the past
● Worked on data cleaning and ensured data quality, consistency, integrity using Pandas, Numpy
● Worked on outliers identification with box-plot, studentized, K-means clustering using Pandas, Numpy
● Participated in features engineering such as feature intersection generating, feature normalize and Label encoding with Scikit-learn preprocessing
● Performed univariate and multivariate analysis on the data to identify any underlying pattern in the data and associations between the variables
● Built the machine learning model include: logistic regression, random forest, XGboost to score and identify the potential new business case with Python Scikit-learn
● Ensured that the model has high accuracy, validated model by ROC and AUC
● Designed rich data visualizations to model data into human-readable form with Tableau 9.2, R ggplot2, R Shiny
● Used Git 2.X for version control and coordinating with the team

Environment:
Python 2.7(Numpy. Pandas, Scikit-learn), R(ggplot2, Shiny), Jupyter Notebook 4.0, SQL Server 2008, Tableau 9.2

Data Analyst

Ping An Insurance - Shanghai, CN - July 2013 to August 2014

Customer Credit Score
Ping An Insurance is a holding company whose subsidiaries mainly deal with insurance, banking, and financial services. The company was founded in 1988 and has its headquarter in Shenzhen. It is one of the top 50 companies in the Shanghai Stock Exchange. Ping An is also a component of Hang Seng Index, an index of the top 50 companies in the Hong Kong Stock Exchange.

My team is responsible for the P2P platform evaluating each borrower's credit score using past historical data and assign an interest rate to the borrower.

Responsibilities:

- Built next-generation prediction model for personal loan, which utilized multiple sources of information to produce optimal loan risk assessment
- Rebuilt personal loan valuation model, reconstructed statistical drivers, provided data science consulting support to business teams
- Selected models implemented include: Logistic Regression, Decision trees, Random Forest, SVM with Python 3.3 sklearn
- Used F-Score, Precision, Recall, and A/B testing to evaluate Model performance
- Designed rich data visualizations to model data into human-readable form (map, Tableau, etc.)

Data Analyst

Century 21 Real Estate LLC - Hangzhou, CN - February 2012 to June 2013

Hangzhou, China Feb 2012 - Jun 2013
Data Analyst
Housing Price Prediction
Century 21 Real Estate LLC is an American real estate agent franchise company founded in 1971. The system consists of approximately 6,900 independently owned and operated franchised broker offices in 78 countries and territories worldwide with over 106,000 sales professionals.

In this project, my team was responsible to make more accurate and efficient housing rent price prediction, and predict how popular an apartment rental listing is based on the renting price, listing content like text description, number of bedrooms, etc.

Responsibilities:

- Worked on data transformation and accessed raw marketing data in varied formats with different methods for analyzing and processing
- Worked with BI team in data investigation, responsible for interpreting variables and data visualization
- Tested the models for problems like goodness of fit, over-fitting, multicollinearity, residual normality, etc
- Involved in variable selection, model optimization and model selection
- Prepared model instructions and detailed report for managers, including explaining variables and scenario tests of predictive models

Environment:
Python 2.7, Rstudio, SVN, SQL Server 2008

Data Analyst

Huawei - Beijing, CN - November 2010 to February 2012

Huawei Technologies Co. Ltd. is a Chinese multinational networking and telecommunications equipment and Services company headquartered in Shenzhen, Guangdong. It is the largest telecommunications equipment manufacturer in the world, having overtaken Ericsson in 2012.

Responsibilities:

● Worked on data transformation and accessed raw marketing data in varied formats with different methods for analyzing and processing

● Translated business needs into data analysis, business intelligence data sources and reporting solution for different types of Client

● Join tables by employing SQL inner join, intersect, union all and union syntax

● Used complex query statements like, sub queries, correlated queries, derived tables, case functions to insert the data depending.

● Developed an A/B test dashboard with data from three different testing platforms

Environment:
Python 2.7, Rstudio, Git 2.0, SQL Server 2008

SKILLS

Python (5 years), SQL (4 years), R (4 years)

ADDITIONAL INFORMATION

TECHNICAL SKILLS:

Machine Learning Algorithms: Programming Language:
Logistics Regression, Naive Bayes, Decision Python 2.7/3.5 (Numpy, Scipy, Pandas,
Tree, Random Forest, KNN, Linear Regression, Seaborn, scikit learn, NLTK), R 3.0, SAS 9.1,
Lasso, Ridge, SVM, Regression Tree, XGboost, SQL
K-means

Analytic Tools: Hadoop Ecosystem
Anaconda 4.0 (Jupyter NoteBook 4.X, Spyder), Hadoop 2.X, Spark 1.6+ (Pyspark, Sparksql,
Rstudio MLlib), MapReduce, Hive 1.X,

DataBase: Data Visualization:
SQL Server […] MongoDB 3.2 Tableau 9.4, Python- matplotlib