Scott Burger

Data Scientist, Microsoft

Bellevue, WA - Email me on Indeed: indeed.com/r/Scott-Burger/5f6344559b029095

website: svburger.com
Willing to relocate: Anywhere
Authorized to work in the US for any employer

WORK EXPERIENCE

Data Scientist

Microsoft - Redmond, WA - January 2017 to Present

With this group, I was tasked with rebuilding the entire end-to-end data pipeline to provide a stream of insight to feed our analytics dashboards. I had inherited a poorly run system that was then rebuilt and highly optimized in a more logical capacity from a database engineering perspective. I rebuilt the system from the ground up to have all the data processing done in the cloud instead of in a local system's dashboarding documents. I also leveraged R code for machine learning insights into the data. From this work, we were able to evolve the group from dashboarding static reports to generating trends and insights in the data to use for predictive purposes.

Data Scientist

Microsoft - Redmond, WA - February 2016 to January 2017

Microsoft's Retail and Channel Marketing team focuses on sales unit modelling for various product groups across a very stratified set of data, including several different form factors, distributors, and supply-chain business logic scenarios. With the sales team, I've led a small team of data scientists to guide modelling approaches as well as designing analytical models myself. We have successfully completed many business objectives ranging from data warehousing strategy to complex machine learning model forecasts. The team had originally been relying on Excel as a crutch for all their data needs. One of our most impactful fixes was to use a more intelligent SQL system to warehouse the data. Another key impact we drove was to utilize machine learning association rules to better understand our data and how to recommend it to customers.

The highest impact deliverable for this project was developing a market sizing model to determine where gaps are within internal data and how to approach an extrapolation to the full market picture with minimal reliance on third party data. The model is primarily a data reshaping exercise that is built in an R environment. We use the final flat sizing model table as a base layer for visualization to key business stakeholders and to help drive decision making.

I have done extensive work in applying machine learning modelling to sales projections from our data warehousing outputs. By tying together different datasets from vastly different business ends of the spectrum and reshaping appropriately, we first build a unified data structure across multiple verticals including time, form factor (ie, desktop or notebook), and sales channel, we can then apply GDP-based economic factors to the data and train our machine learning models accordingly. We have done this exercise with great accuracy for predicting unit sales at a yearly and quarterly level.

Other projects in this space include theoretical economic market cap estimation, ARIMA modelling and seasonality decomposition for use in sales forecasting, cross-channel retail consumption models based on extracted linear models for each vertical slice of the data, and designing sales growth ranking algorithms to bucketize distributions into high, medium, and low performance categories.

Principal Data Scientist

Microsoft - Redmond, WA - May 2014 to January 2016

With the highly agile and fast-paced Engineering, Community, and Online (ECO-IT) team, I am responsible for leading a team of Solution Managers, UX Researchers, Engineering Leads/PMs, customer and business SMEs to determine what's important to the customer/user and understand the experience (work flows, customer/user journey), target low level drivers, measurements, metrics and success drivers. I work with engineering teams to help design instrumentation needs for engineering systems and improve data acquisition and collection strategy by doing gap analysis.

Duties include conducting statistical analysis to determine key factors for planning and conducting experiments to prove causality using prescriptive and predictive analytics by application of appropriate machine learning algorithms (decision trees, classification, clustering, regression trees, logistic regression, random forest, regression etc.)

I am also responsible for A/B Testing of experiments designed to validate clear hypotheses regarding measurable outcomes, and for modelling the telemetric requirements for the KPIs, drivers and True North Metrics for any strategic priority. Additional duties include: analytical model management in databases between WEDCS, COSMOS, Hive, SSAS data cubes, and HDInsight clusters running on Azure.

Responsible for adhoc data queries using T-SQL scripting, and adhoc data integration in MSIT provided servers by building Data models using SSAS and C#. I am also conducting statistical analysis to determine key factors for planning and conducting experiments to prove causality using statistical tools such as R, Matlab, JMP and Data Visualization tools such as SSRS, PowerView.

Data Scientist

Microsoft - Redmond, WA - November 2012 to November 2013

The Data & Decision Sciences Group (DDSG) is chartered to provide data-focused analytics leadership and promote data-driven decision making throughout Microsoft.

The group interfaced with many internal corporate clients to provide analytical insights with the use of R, SAS, Python, and JMP. I've been utilizing data mining techniques, statistical modeling, segmentation methods, profiling, and targeting to provide solutions to business problems. I helped develop predictive models and clustering/graph algorithms to detect at-risk product keys & global piracy patterns and built regime-changing models to predict code signing time for large builds.

I was also responsible for using the group's Hadoop cluster to architect, design, and develop actionable analytics with associated Hadoop technologies like Pig and Hive.

Delivered a training course to senior management level employees on topics including: Theoretical Probability Distributions, Hypothesis testing, Correlation, Experimentation Principles, and tutorials on specific tools like Sigma-XL, Minitab, JMP, and Excel.

Delivered a training course to senior management level employees on topics in data visualization, such as: design principles for visualization, basic and advanced visualization principles, analysis principles for data visualization, correlation, data exploration, and big data analytics. Tutorials on specific tools like powerview and pivot charts were done in small groups with attendees.

Software Engineer

Market Fish - Seattle, WA - April 2012 to July 2012

With Market Fish, I used Unix-based list operations to clean, normalize, map, ingest and index lists of customer data into the company's platform. I also extensively used Big Data-related

programming tools such as Hadoop, Splunk and Elastic Map Reduce for statistical analysis of large data stored on the company's Amazon Web Service storage.

Research Assistant

University College London - London, OH - October 2010 to August 2011

Here I worked as a research assistant in the Dark Energy Survey group at UCL. Here I wrote functions in Perl, C, Matlab and Fortran to analyze power spectra and correlation statistics. My analysis set constraints on error levels in various cosmological parameters by comparing data sets with and without covariance for a future multi-million dollar, international galaxy survey. Results from my thesis work on this topic suggest that the Dark Energy Survey must have parameter estimation accuracy in the realm of point spread functions of at least 95% for data without covariance taken into account. With covariance taken into account, purity of samples could be as low as 85% and still be within 2-σ.

Research Assistant

Western Washington University - Bellingham, WA - September 2007 to August 2010

Responsibilities

At WWU I worked primarily in numerical and data analysis with Perl, Fortran and C. I engineered functions and scripts to probe faint-end luminosity functions of galaxy clusters, then determined their parameters with use of the Sloan Digital Sky Survey Data Release 7.

EDUCATION

Masters of Science in Astrophysics

University College London - London
2010 to 2012

Bachelor of Science in Physics

Western Washington University - Bellingham, WA
2006 to 2010

SKILLS

Artificial Neural Networks, C, C#, C++, Emacs, Emacs-speaks-statistics, Fortran, Git, Hadoop, Hive, Java, JMP, LATEX, Matlab, Numpy, Perl, Pig, Powersim, Python, PyCuda, R, SAS, SQL

LINKS

http://www.linkedin.com/profile/view?id=205774719

PUBLICATIONS

An Introduction to Machine Learning with R

December 2017

Machine learning can be a scary concept for the uninitiated. This book aims to provide a solid foundation of introductory principles used in machine learning with R. Starting with the basics like Regression, we move into more advanced topics like Neural Networks, then into the frontier of machine learning in the R world with packages like Caret.

A Quantitative Satisfaction Index to Replace NSAT

November 2017

In this paper, I derive, from basic mathematical principles, a more robust metric for gauging customer satisfaction based on Lickert-style form surveys. I show that the current implementation of NSAT is negatively biased and the improved QSAT method can, in some cases, be higher upwards to 20 percent.

ADDITIONAL INFORMATION

Resourceful analytical professional with experience with bleeding-edge technologies, programming, and advanced statistical analysis. A skilled communicator and experienced with developing and leading technical workshops.

Experience in applying machine learning and statistical modeling techniques to solve business problems. Expert in distilling vast amounts of data to meaningful discoveries at requisite depths at scale with a wide repertoire.

A strong technical voice for his team who is recognized as a "go to" person for Hadoop, Hive and Pig. He cares about doing things right, sharing what he has learned, and exploring new options.

Specialties:

Logistic regression, categorical data analysis, general linear models, nonparametric methods, factor analysis, cluster analysis, time-series forecasting & large-scale data analysis, machine learning algorithms (support vector machines, naive-bayes, classification and regression trees), KD-tree algorithms, design of experiment principles, screening experiment analysis, sequence mining, proficient in MatLab, JMP, SQL, R, comfortable with Java, Python, experience working on very large databases and working with MapReduce technologies