Yasir Modak

Data Scientist - ProKarma

Portland, OR - Email me on Indeed: indeed.com/r/Yasir-Modak/d1ea5847e0bfd914

Willing to relocate: Anywhere
Authorized to work in the US for any employer

WORK EXPERIENCE

Data Scientist

ProKarma - Portland, OR - January 2015 to Present

Currently I am working for a leading automobile manufacturer headquartered at Portland, OR. I am involved in multiple data science projects, recently I worked on a project that asked for building a predictive model that utilizes text as features to predict repair hours for trucks. Further we also formulated a recommender system which facilitated an enhanced approach for configuration of trucks. This reduce the number of clicks to build a customized trucks on dealer system by 80%. Further performed sentiment analysis and text summarization.

Responsibilities:
• Examined the existing database, collected statistics to learn about user behavior.
• Merged user data from multiple data sources.
• Performed Exploratory Data Analysis using.
• Prototype machine learning algorithm for POC (Proof of Concept).
• Performed Data Cleaning, features scaling, features engineering.
• Formulated a novel approach to build machine learning algorithm and implemented it in production environment.
• In real-time association rules were implemented which uses prior probabilities.
• Performed Data Mining in R (TM package, LSA package) using SAP HANA platform.
• Established dimensionality reduction by SVD, 1500 data codes were transitioned into 24 different unique features.
• Developed Performance metrics to evaluate Algorithm's performance.'
• Performed data visualization on the front end by using SAP Lumira.

Environment: TERADATA, Oracle, HADOOP (HDFS), R Studio, Python, SAP HANA, JAVA, HIVE, SAP LUMIRA.

Data Scientist

Walmart - Bentonville, AR - March 2014 to December 2014

Walmart has millions of customers who shop in store as well as online over a range of thousands of products. It produces enormous amount of data which can provides insights about the products which are in demand, consumer habits, consumer demands, marginal profits from the sales etc. Our aim is to make wise use of such data to develop a recommender engine which can learn from past data of the customer transactions and recommend relevant options of new products to the customers. This enhances the customer experience as well as increases sales for the Walmart. By implementing this project, we were able to achieve 5% increase in online sales revenue.

Responsibilities
• Examined the existing database, collected statistics to learn about user behavior.

• Merged user data from multiple data sources.
• Used Collaborative Filtering with Latent Factors model to build a recommender engine.
• Performed extensive implicit as well as explicit data collection.
• Performed Exploratory Data Analysis using R and Hive on Hadoop HDFS.
• Prototype machine learning algorithm for POC (Proof of Concept).
• Performed Data Cleaning, handled missing data, outliers, features scaling, features engineering.
• Developed Performance metrics to evaluate Algorithm's performance.
• Calculated RMSE score, F-SCORE, PRECISION, RECALL, and A/B testing to evaluate recommender's performance.
• Addressed the over-fitting by adding regularization (lasso / ridge) term in the algorithm.
• Fine-tuned low bias and high variance trade off.
• Performed post-hoc data analysis (tukey test) in R.

Environment: TERADATA, Oracle, HADOOP (HDFS), PIG, MySQL, R Studio, Python, MAHOUT, JAVA, HIVE.

Data Scientist

Providence Health - Beaverton, OR - January 2012 to February 2014

Providence is one of the leading providers of health insurance in the west coast. Naturally it has to maintain millions of records of its customers which includes insurance claims, phone call conversations with the insurance agent, customer service preferences, and customer feedback. Our idea was to build smart categories that accurately captures customer's interactions based on the previous records. And to build an algorithm that learns from interactions between the customer and providence agent. Since this helps to provide an overview about the customer, it helped to reduce the call time by 40% and enhanced the customer experience.

Responsibilities:
• Performed exploratory data analysis by using R and SAP HANA.
• Designed, implemented and automated modeling and analysis procedures on existing and experimentally created data.
• Parsed data, producing concise conclusions from raw data in a clean, well-structured and easily maintainable format.
• Implemented Topic Modelling.
• Perform tfidf weighting, normalize.
• Performed scheduled and adhoc data driven statistical analysis, supporting existing processes.
• Developed clustering models for customer segmentation using R.
• Performed Time Series Modeling to forecast future sales and revenue.

Environment: R, SQL, Python, TABLEU, SAP HANA, SAS, JAVA, PCA & LDA, regression, logistic regression, random forest, neural networks, Topic Modeling, NLTK, SVM (Support Vector Machine), JSON, XML, HIVE, HADOOP, PIG, MAHOUT.

Data Scientist

Bank of America - Pittsburgh, PA - January 2010 to December 2012

Formulated a predictive credit risk model that can correctly determine whether the given loan is safe or risky based on the various features of the customer. The project was to build an algorithm that accurately classifies credit card holders among multiple classes based on the historical data available on multiple features. Further, the aim was to improve bank's efficiency by reducing default rate while offering new products. Also performed Time Series Modeling for forecasting future revenue.

Responsibilities:

• Responsible for predictive analysis of credit scoring to predict whether or not credit extended to a new or an existing applicant will likely result in profit or losses.
• Data was extracted extensively by using SQL queries and used R packages for the data mining tasks.
• Collected data from application information as well as behavioral information of the customers.
• Performed data cleaning and handled missing data by utilizing multiple methods (knn, OLS etc).
• Conducted principal component analysis, using PRINCOMP to reduce the dimension of the data. Used WOE to convert categorical data to continuous data
• Computed Credit Risk Parameters such as Probability of Default (PD) and Loss Given Default (LGD) and Exposure at Default (EAD).
• Improved the profit margin by 2% by predicting profit or loss due to existing or new costumer.
• Built a logistic regression model, performed clustering and evaluated its performance on test metrics.
• Used Kolmogorov-Smirnov test (K-S test or KS test) to measure the quality of the models.
• Used PROC GRAPH for generating various graphs and charts for analyzing the different features.
• Used k-fold cross validation on the dataset to avoid over fitting.
• Evaluated the overall performance of credit risk model by using ROC- curve, AUC.

Environment: .Net, SAS, R, Python, Oracle, IBM DB2, MS SQL, HIVE, HADOOP, PIG, MAHOUT.

EDUCATION

Bachelors in Mechanical Engineering

Shivaji University

Masters in Mechanical Engineering

University of Texas - Arlington, TX