Multivariate Time Series Prediction

Khyati Parekh
Niharika Sharma
Rajiv Veeraraghavan
Team - The Time Turners
University of Washington, Seattle

Submitted to -
Advisor - Megan Hazen, University of Washington
Sponsor - Luca Cazzanti, Amplero

Content

Executive Summary

The goal of this project is to perform a multivariate time series analysis on the data to find the probability of a user churn out. This project will be completed in a span of 3 months and the cost will depend on how long we need to use AWS for the analysis. We aim to give Amplero a fair idea about what the probability of churn of a user is and when they are most likely to churn out. Since the timing of an interaction with a user can have positive effects for retention, user engagement, and revenue that are difficult to capture with standard model performance metrics; specialized metrics and key performance indicators may need to be developed. This is one of the challenges we face while developing models to analyse the data given to us. Assessment of the data requirements and model complexity are also a part of the deliverables.

Statement of Problem

Amplero is an Artificial Intelligence Marketing (AIM) company that enables business-to-consumer (B2C) marketers at global brands to optimize customer lifetime value at a scale that is not humanly possible. Unlike traditional rules-based marketing automation systems, Amplero's Artificial Intelligence Marketing Platform leverages machine learning and multi-armed bandit experimentation to dynamically test thousands of permutations to adaptively optimize every customer interaction and maximize customer lifetime value and loyalty. With Amplero, marketers in competitive, customer-obsessed industries like telecom, banking, gaming and consumer tech are currently seeing measurable lift across key performance indicators—including 1-3% incremental growth in customer topline revenue and 3-5x lift in retention rates [2].

The subscription economy challenges marketers to address the longitudinal behavior of consumers. Our goal is to predict the future behavior of customers from multivariate time series. Why are we using multivariate time series? The assumption is that multivariate time series describe the longitudinal behavior of users along various dimensions and carry predictive power beyond what univariate models provide. Possible ideas and techniques to do so include, but are not limited to, the latest techniques in deep learning (for example LSTMs), or an innovative approach to Hidden Markov Models (HMMs). We will choose techniques with broad discretion and try to not confine ourselves to existing methods of implementation.

Amplero's compressed dataset consists of 3 types of files:

1) 10 usage*.jsonish.gz files that contains the usage data
2) 10 social*.jsonish.gz files that contains social data
3) State_recharge_etc.jsonish.gz that contains state, event, and some usage data

Using, Machine Learning and/or Deep Learning methods(section Technical Approach) we will predict user churn among prepaid users given a multivariate time series. Also, we will conduct research to measure the similarity between two time series depending on magnitude by doing Value comparison, Trend comparison, Distribution comparison, Distance Analysis, and other Statistical Test like t-test. Even a 1-3% decrease in user churn can tremendously increase in the revenue for the mobile network.

Objectives

Our main goal is Multivariate Time Series Prediction of User Behavior. The subscription economy challenges marketers to address the longitudinal behavior of consumers. For instance, let's consider the prepaid mobile case - Users choose when to replenish their accounts, and the amount of the top up, according to their usage and needs. A marketer who can predict the timing and amount of a user's future recharge events can craft personalized offers that are more likely to resonate with the user.

Similarly, being able to predict a user's propensity to churn out of a subscription, and the timing of the churn event, makes marketing more efficient. We aim to provide a mostly accurate prediction for Amplero using various machine learning techniques which we have learned during the course of this degree.

Our objective is, among other things, perform a multivariate time series analysis on the data to find the probability of a user churn out. We are also considering other machine learning algorithms like Hidden markov models (HMM), Long short term memory networks (LSTM) etc. for the same goal.

Technical Approach

This section describes the technical approaches that we will adopt to productionize the solution. This section talks about how we will identity needs of our customer, identify target specification, detailed literature on some models used for multivariate time series analysis, and design concept that we intend to use and deploy.

Identifying Needs of Customers

Correctly identifying our customer's needs is essential for ensuring customer satisfaction and expectation. To identify our customer's need we will:

- Based on the given data, we would iterate our results with customers. Eg. we would develop a model and immediate look for feedback on its performance and also better understand the features the customer actually care about.
- We believe that using an iterative approach to developing the machine learning model will really help in better understanding the needs of customer.
- In a lot of scenarios, getting domain expertise from the customer will be tremendously useful and we would like to acquire the domain knowledge through frequent interaction with the customer.

Identifying Target Specifications

We look to leverage concepts such as time series analysis and machine learning to tackle the problem statement. With time series analysis, we aim to explore a similarity metric between various time series with the hope that we can accurately distinguish between churned and non-churned users. From a machine learning viewpoint, we look to implement various algorithms such logistic regression, HMM, LSTM to predict the probability of churn out for a user. Since we dealing with time series that can modeled as a markov chain problem, algorithms such as HMM and LSTM can play a valuable role in predicting the probability of user churn.

Literature Review

A time series is a series of data points recorded in time order. As the name suggests, it involves working on time (years, days, hours, minutes) based data, to derive hidden insights to make informed decision making. Data Mining methods are being increasingly used in forecasting time series data, in addition to conventional statistical approaches. Two important aspects of time series data mining can be identified as classification and forecasting. To build and try stochastic models on the time series data, we need to make sure that the dataset is stationary. The following are the three properties of a stationary series:

1. The mean of the series should not be a function of time rather should be a constant.
2. The variance of the series should not a be a function of time.
3. The covariance of the i th term and the (i + m) th term should not be a function of time.

If the time series data is not stationary, then there are multiple ways of bringing this stationarity, for instance, Detrending, Differencing etc. Exploring data becomes most important in a time series model – without this exploration, we will not know whether a series is stationary or not.

Auto Regressive Moving Average (ARMA; Peter Whittle (1951)) models are commonly used in time series modeling and AR or MA are not applicable on non-stationary series.Given a time series of data X_t , the ARMA model is a tool for understanding and, perhaps, predicting future values in this series. The model consists of two parts, an autoregressive (AR) part and a moving average (MA) part. The AR part involves regressing the variable on its own lagged (i.e., past) values. The MA part involves modeling the error term as a linear combination of error terms occurring contemporaneously and at various times in the past.

Over the past few decades, artificial neural networks have attracted great attention in the time series forecasting community. ANNs exhibit superior performance on classification and regression problems in the field of machine learning. Compared to statistics-based forecasting techniques, neural network approaches have an upper hand and several unique characteristics, like

1. NNs are nonlinear and complicated model;
2. NNs are data-driven being flexible and universal; and
3. NNs are nonparametric as they have no requirement for an explicit underlying model.

Neural networks have been used widely for a comprehensive range of applications in time series forecasting varying from weather, financial, business, to disaster management.

SVM-based forecasting methods use a class of generalized regression models, such as Support Vector Regression (SVR) and Least-Squares Support Vector Machines (LSSVMs; Smola and Scholkopf (2004)), that are parameterized using convex quadratic programming methods (Balabin and Lomakina, 2011). SVMs are categorized into linear, Gaussian or RBF, polynomial, and multilayer perceptron classifiers. A linear regressor is then constructed by minimizing the structural risk minimization (the upper bound of the generalization error), leading to better forecasting performance than conventional techniques (Cao, 2003). Recently, extreme learning machine (ELM), a new type of neural network has been introduced (Huang et al.,) for regression and classification problems.

Most of the models involve batch processing, where the model is fit and updated intermittently using batches of historical data. However, the curse of dimensionality due to the prohibitive computational effort, memory requirements, and large data sizes hampers their applicability to many real-world problems, especially for real-time process monitoring. A variety of sequential (also known as online or recursive) forecasting models, such as Hidden Markov Models (HMMs; Rabiner (1989)), are investigated to surmount this limitation. A Hidden Markov model (HMM) is a statistical Markov model in which the system being modeled is assumed to be a Markov process with hidden states. An HMM can be considered as the simplest dynamic Bayesian network.

Design Concept

On a high level, we look to analyze and apply machine learning techniques on a subset of the the 400Gb compressed dataset using a large compute machine such as AWS ec2.

For sampling the large dataset, we are going to look at those users that have changed state (active to inactive) over the the period of 180 days. Also, we are going to further sample the dataset based on the user's mobile usage in the last 30 days before churn out. In particular, we are going to ignore users that did not have any mobile usage in the last 30 days before churn out. This is being done so that we can better understand and accurately predict "why users that have had activity in the last 30 days still churn out?".
Based on the above sampling, we believe the dataset will be of appropriate size to run machine learning techniques in a computationally efficient manner.

When it comes to programming and actually implementing the machine learning techniques, we will using python as the base programming language and leverage packages such as scikit learn to run the algorithms. Scikit learn has modules for ML techniques such as logistic regression, SVM and kernel based methods. For implementing Hidden Markov models in Python, there exists an available open source github library [8] for implementation. We are also looking to leverage other open source libraries - eg. LSTM for time prediction [7] is available on github and we believe we can improvise on the current algorithm to better suit our use case.

For testing the effectiveness of our methodology, we will be training on a part of the sampled dataset and testing on the remaining (preferable 80:20). While training, we believe regularization is going to play a key role in deciding the performance of a particular machine learning technique.

As we progress with our machine learning modelling, that dataset considered might be large and in such cases, we are looking to use large ec2 instances from AWS where deep learning and scikit learn packages have already been installed to improve on the computational efficiency. Running ML algorithms on our local computers will be extremely cumbersome as they don't have enough processing power (RAM, CPU and memory) for such computationally intensive algorithms such as neural networks.

Project Management

The following are the specification regarding the timelines and deliverables -

i . Project duration - 09/27/2017 through 12/15/2017 - 01/03/2018 through 03/16/2018

ii. Milestones -
Milestone 1 - Exploratory analytics and experimental designing phase. By the end of this milestone, we will have a clear understanding of the whole dataset. We would have figured out the erroneous, difficulty with the dataset. Also, from data exploration, data visualization and statistical experimentation, we will select the models that can be used to solve the problem. Our main aim would be to get comfortable with accessing the dataset because the dataset is a true representation of large scale dataset (800 GB), whIch requires using scalable methods and frameworks like Apache Spark, Amazon EMR and S3.

Milestone 2 - Model down select
From n different models that we selected from milestone 1, based on their performance, speed and scale up capacity, we will select the final model and start implementing to productionize it.

Milestone 3 - Final Deliverable
This is the final deliverable, which consist of robust documentation of the study, Codebase, final model and result as per the industrial standards.

iii. Schedule - The roadmap in Figure 1 contains detail about all tasks related to data management and collection to model implementation, designing, and testing. All the three members of the team will do all the tasks. Our team has 3 members with a very similar background, which makes everyone comfortable with most of the task that is required to be done.

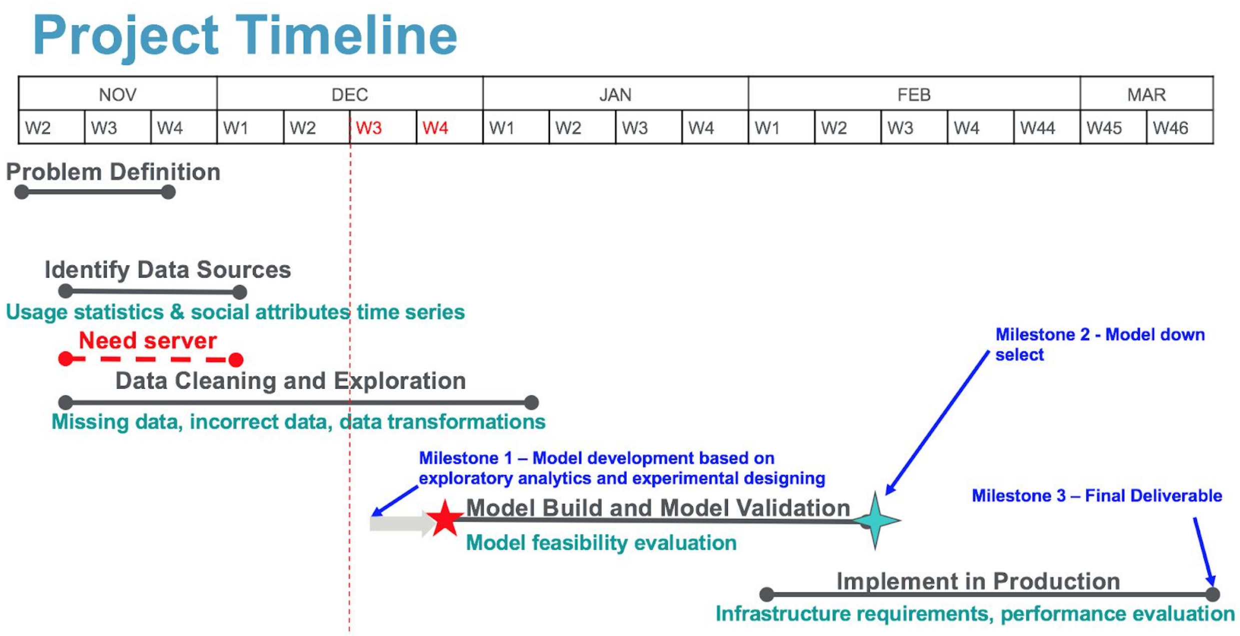


Fig 1 - Project Timeline

Deliverables

Based on the statement of a problem and other details provided to us during initial scrums, we propose the following deliverable for the project:

1. Similarity between time series depending on magnitude - Explore good similarity measures for time series that would tend to assign more similarity to time series of users in the same category (non-churners or churners) and less similarity to users from different categories.
2. Churn Classification - Classification of data-points (customers) as churner and non-churners. Provide probability of a customer either switching their existing plan or leaving the network altogether.

The above tasks are what we intend to do and deliver, but a lot is dependent on the data-set we have and what challenges we will come across while dealing with it. Hence, the definition of the deliverables might change in future with mutual consent and agreement between all the parties.

Budget

One of the key requirements for the project is the extensive usage of Amazon Web services.

1. S3 Bucket - Amplero provides us the data-set using S3 service. All the files are in JSON format and the total size is approximately 400 GB (compressed), capturing the snapshot of last 180 days' usage data.

2. EC2 Instances - At Least 1 instance for initially conducting the exploratory analytics, followed by deployment of models.
3. Nosql Database AMI + EC2 Instance or RedShift Cluster to store the data provided by Amplero.

Table 1: Requested items and funds for initial design

Service	**Quantity**	**Duration**	**Expected Cost monthly**
AWS EC2 Instance	1 - m3.2xlarge (On-demand usage)	09/27/2017 through 03/16/2018	$ 389.43
S3 Storage space	1000 GB	09/27/2017 through 03/16/2018	$23
AWS Redshift	1	01/03/2018 through 03/16/2018	No estimate yet
Storage: Amazon EBS Volumes	1 volume with 1000GB storage, General purpose SSD (gp2)	09/27/2017 through 03/16/2018	$ 100 - 3(free tier)

* The cost is estimated using AWS calculator - http://calculator.s3.amazonaws.com/index.html

Communication and Coordination with Sponsor

Rajiv will be primary source of contact with the sponsor. We have set up a recurring weekly meetings with the sponsor (Luca) to update him with the progress made and ask for feedback. In our meetings, we will also be deep diving into the machine learning techniques explored. This meeting will either take place remotely over a conference video call or at University of Washington main campus. We will also communicate frequently with the sponsor over email and ask for suggestions/feedback. Whenever we communicate with sponsor over email, all the three participants are included in the email list so that everyone is on the same page.

Team Qualifications

Khyati Parekh

I am a graduate student at the University of Washington pursuing a Master's degree in Data Science. I completed my undergraduate studies in Computer Science. Over the past year, I have learned and applied machine learning and other data analysis methods in various projects.

Niharika Sharma

I am a graduate student at the University of Washington pursuing Masters in Data Science. I have done my under graduation in Computer Science Engineering. As a part of Data Science program and past industrial experience, I work heavily with large-scale datasets utilizing a diverse array of technologies and tools as needed, to deliver insights, such as Java, Python, Tableau, AWS, SQL, and NoSQL database systems. Additionally, having approximately 2 years of industrial experience in the field of data science and software engineering.

Rajiv

I am a second year Master's student in Data Science at University of Washington. Prior to my Master's, I worked as a Quantitative strategist at Goldman Sachs from 2014 to 2016 where they developed statistical models to optimize firm wide trade operations. I have earned a Bachelor's degree in Computer Science and also has experience working on machine learning, applied statistics and software development.

Conclusion

In conclusion, through this project we hope to get a better idea of how to do time series analysis and what methods are best suited for it. We also hope to accurately predict the number of customers who churn out as even a 1 or 2% drop in churn out could result in significant savings for the company.

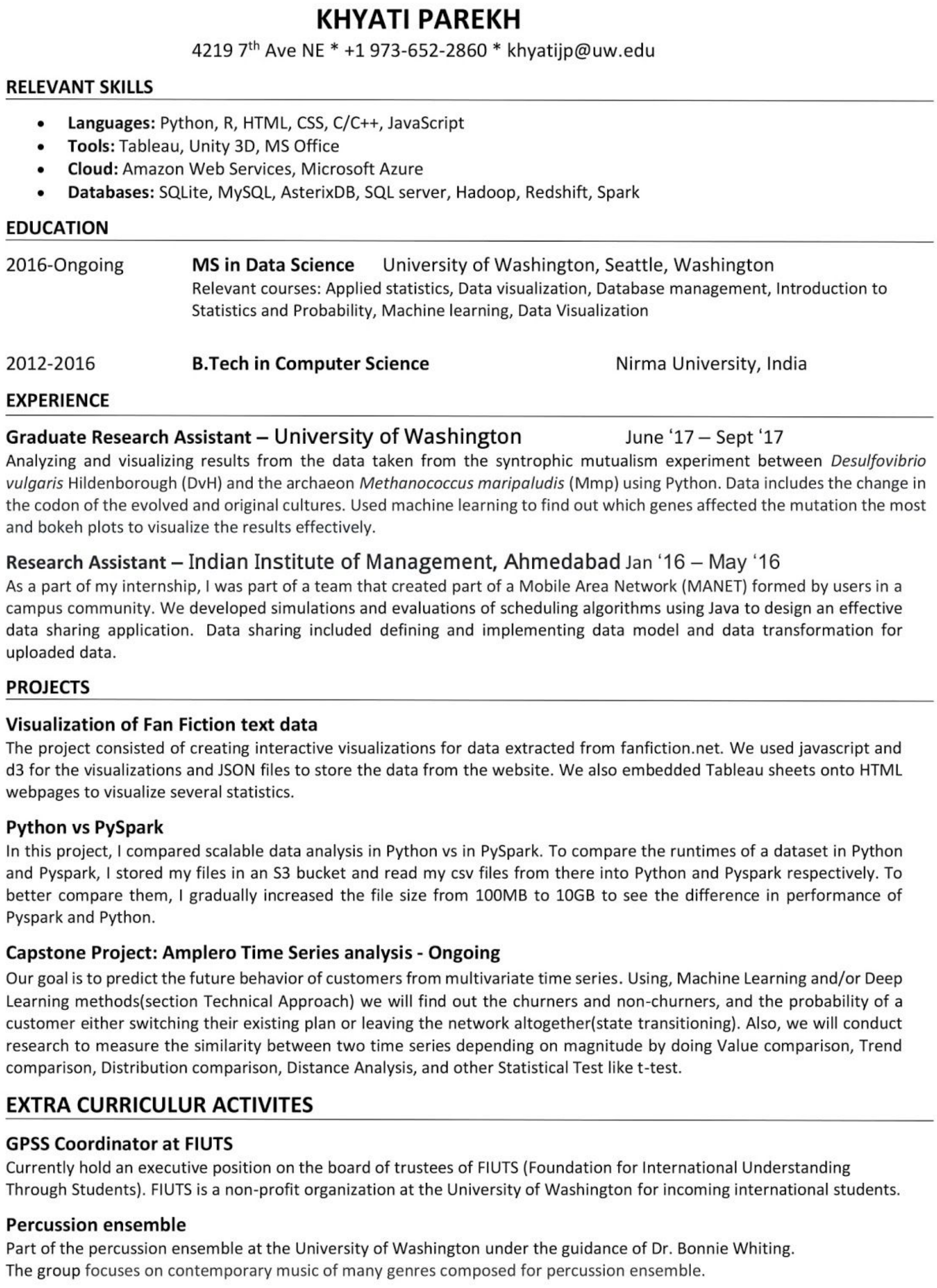
In addition to getting a better idea about time series analysis, we also hope to learn how to deal with large datasets using AWS and deep learning techniques. By working with Amplero on this project, we will also be in a better position to enter the job market and understand how to work with real world data science problems as opposed to working on projects in a particular course. In conclusion, we hope to gain further insight into solving data science problems by working on this project.

References

1. Amplero
2. meganursula/DATA590A
3. http://shodhganga.inflibnet.ac.in/bitstream/10603/75388/19/19_appendix.pdf
4. http://www.ijmer.com/papers/Vol4_Issue7/Version-2/IJMER-47020105.pdf
5. https://en.wikipedia.org/wiki/Autoregressive_model
6. https://www.analyticsvidhya.com/blog/2015/12/complete-tutorial-time-series-modeling/
7. https://github.com/jaungiers/LSTM-Neural-Network-for-Time-Series-Prediction
8. https://github.com/hmmlearn/hmmlearn

Appendix A

Khyati Parekh



KHYATI PAREKH

4219 7th Ave NE * +1 973-652-2860 * khyatijp@uw.edu

RELEVANT SKILLS

- **Languages:** Python, R, HTML, CSS, C/C++, JavaScript
- **Tools:** Tableau, Unity 3D, MS Office
- **Cloud:** Amazon Web Services, Microsoft Azure
- **Databases:** SQLite, MySQL, AsterixDB, SQL server, Hadoop, Redshift, Spark

EDUCATION

2016-Ongoing **MS in Data Science** University of Washington, Seattle, Washington
Relevant courses: Applied statistics, Data visualization, Database management, Introduction to Statistics and Probability, Machine learning, Data Visualization

2012-2016 **B.Tech in Computer Science** Nirma University, India

EXPERIENCE

Graduate Research Assistant – University of Washington June ‘17 – Sept ‘17

Analyzing and visualizing results from the data taken from the syntrophic mutualism experiment between *Desulfovibrio vulgaris* Hildenborough (DvH) and the archaeon *Methanococcus maripaludis* (Mmp) using Python. Data includes the change in the codon of the evolved and original cultures. Used machine learning to find out which genes affected the mutation the most and bokeh plots to visualize the results effectively.

Research Assistant – Indian Institute of Management, Ahmedabad Jan ‘16 – May ‘16

As a part of my internship, I was part of a team that created part of a Mobile Area Network (MANET) formed by users in a campus community. We developed simulations and evaluations of scheduling algorithms using Java to design an effective data sharing application. Data sharing included defining and implementing data model and data transformation for uploaded data.

PROJECTS

Visualization of Fan Fiction text data

The project consisted of creating interactive visualizations for data extracted from fanfiction.net. We used javascript and d3 for the visualizations and JSON files to store the data from the website. We also embedded Tableau sheets onto HTML webpages to visualize several statistics.

Python vs PySpark

In this project, I compared scalable data analysis in Python vs in PySpark. To compare the runtimes of a dataset in Python and Pyspark, I stored my files in an S3 bucket and read my csv files from there into Python and Pyspark respectively. To better compare them, I gradually increased the file size from 100MB to 10GB to see the difference in performance of Pyspark and Python.

Capstone Project: Amplero Time Series analysis - Ongoing

Our goal is to predict the future behavior of customers from multivariate time series. Using, Machine Learning and/or Deep Learning methods(section Technical Approach) we will find out the churners and non-churners, and the probability of a customer either switching their existing plan or leaving the network altogether(state transitioning). Also, we will conduct research to measure the similarity between two time series depending on magnitude by doing Value comparison, Trend comparison, Distribution comparison, Distance Analysis, and other Statistical Test like t-test.

EXTRA CURRICULUR ACTIVITES

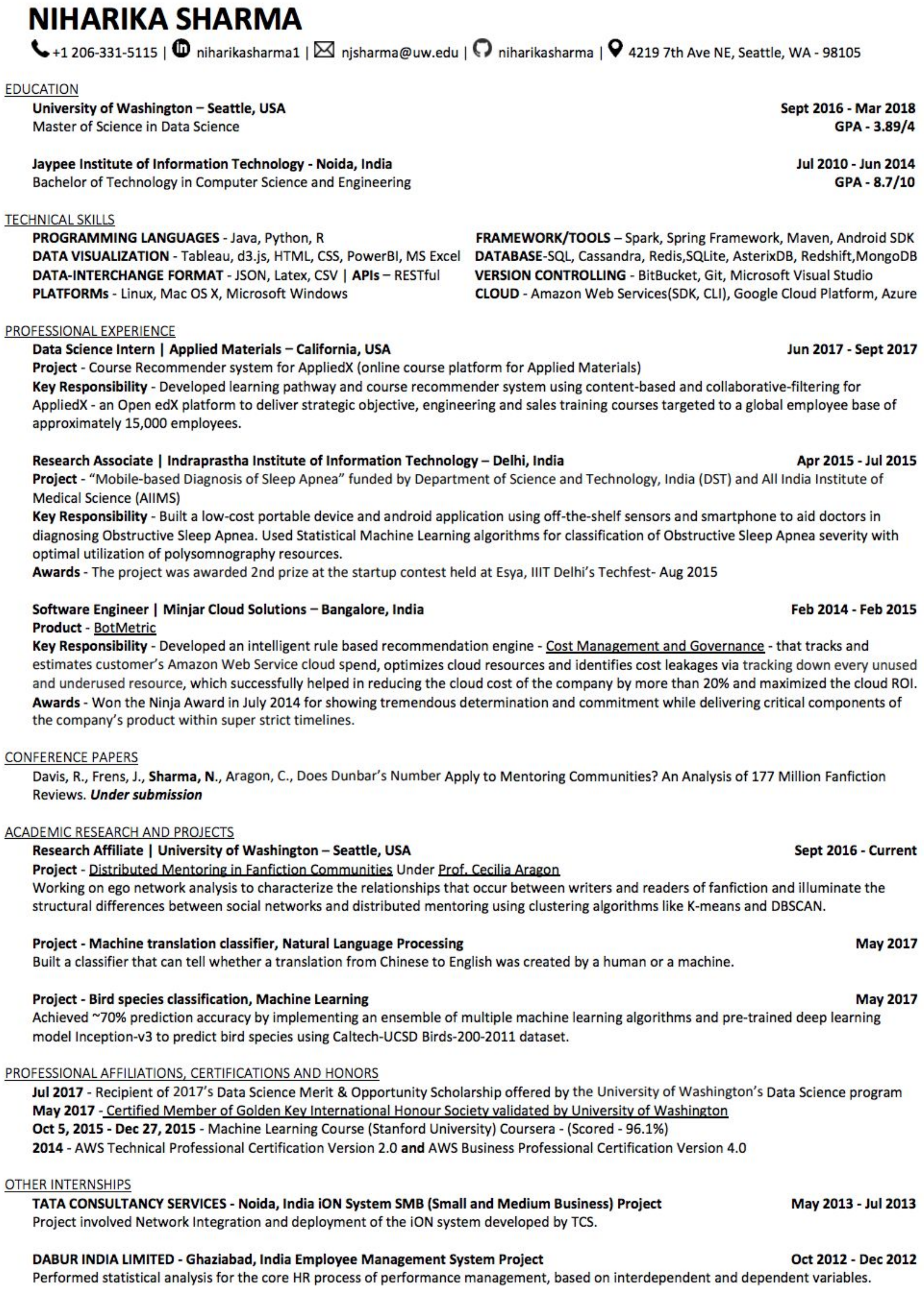
GPSS Coordinator at FIUTS

Currently hold an executive position on the board of trustees of FIUTS (Foundation for International Understanding Through Students). FIUTS is a non-profit organization at the University of Washington for incoming international students.

Percussion ensemble

Part of the percussion ensemble at the University of Washington under the guidance of Dr. Bonnie Whiting.
The group focuses on contemporary music of many genres composed for percussion ensemble.

Niharika Sharma



NIHARIKA SHARMA

☎ +1 206-331-5115 | in niharikasharma1 | ✉ njsharma@uw.edu | niharikasharma | 📍 4219 7th Ave NE, Seattle, WA - 98105

EDUCATION

University of Washington – Seattle, USA **Sept 2016 - Mar 2018**
Master of Science in Data Science **GPA - 3.89/4**

Jaypee Institute of Information Technology - Noida, India **Jul 2010 - Jun 2014**
Bachelor of Technology in Computer Science and Engineering **GPA - 8.7/10**

TECHNICAL SKILLS

PROGRAMMING LANGUAGES - Java, Python, R
DATA VISUALIZATION - Tableau, d3.js, HTML, CSS, PowerBI, MS Excel
DATA-INTERCHANGE FORMAT - JSON, Latex, CSV | **APIs** – RESTful
PLATFORMs - Linux, Mac OS X, Microsoft Windows

FRAMEWORK/TOOLS – Spark, Spring Framework, Maven, Android SDK
DATABASE-SQL, Cassandra, Redis,SQLite, AsterixDB, Redshift,MongoDB
VERSION CONTROLLING - BitBucket, Git, Microsoft Visual Studio
CLOUD - Amazon Web Services(SDK, CLI), Google Cloud Platform, Azure

PROFESSIONAL EXPERIENCE

Data Science Intern | Applied Materials – California, USA **Jun 2017 - Sept 2017**
Project - Course Recommender system for AppliedX (online course platform for Applied Materials)
Key Responsibility - Developed learning pathway and course recommender system using content-based and collaborative-filtering for AppliedX - an Open edX platform to deliver strategic objective, engineering and sales training courses targeted to a global employee base of approximately 15,000 employees.

Research Associate | Indraprastha Institute of Information Technology – Delhi, India **Apr 2015 - Jul 2015**
Project - "Mobile-based Diagnosis of Sleep Apnea" funded by Department of Science and Technology, India (DST) and All India Institute of Medical Science (AIIMS)
Key Responsibility - Built a low-cost portable device and android application using off-the-shelf sensors and smartphone to aid doctors in diagnosing Obstructive Sleep Apnea. Used Statistical Machine Learning algorithms for classification of Obstructive Sleep Apnea severity with optimal utilization of polysomnography resources.
Awards - The project was awarded 2nd prize at the startup contest held at Esya, IIIT Delhi's Techfest- Aug 2015

Software Engineer | Minjar Cloud Solutions – Bangalore, India **Feb 2014 - Feb 2015**
Product - BotMetric
Key Responsibility - Developed an intelligent rule based recommendation engine - Cost Management and Governance - that tracks and estimates customer's Amazon Web Service cloud spend, optimizes cloud resources and identifies cost leakages via tracking down every unused and underused resource, which successfully helped in reducing the cloud cost of the company by more than 20% and maximized the cloud ROI.
Awards - Won the Ninja Award in July 2014 for showing tremendous determination and commitment while delivering critical components of the company's product within super strict timelines.

CONFERENCE PAPERS

Davis, R., Frens, J., **Sharma, N.**, Aragon, C., Does Dunbar's Number Apply to Mentoring Communities? An Analysis of 177 Million Fanfiction Reviews. ***Under submission***

ACADEMIC RESEARCH AND PROJECTS

Research Affiliate | University of Washington – Seattle, USA **Sept 2016 - Current**
Project - Distributed Mentoring in Fanfiction Communities Under Prof. Cecilia Aragon
Working on ego network analysis to characterize the relationships that occur between writers and readers of fanfiction and illuminate the structural differences between social networks and distributed mentoring using clustering algorithms like K-means and DBSCAN.

Project - Machine translation classifier, Natural Language Processing **May 2017**
Built a classifier that can tell whether a translation from Chinese to English was created by a human or a machine.

Project - Bird species classification, Machine Learning **May 2017**
Achieved ~70% prediction accuracy by implementing an ensemble of multiple machine learning algorithms and pre-trained deep learning model Inception-v3 to predict bird species using Caltech-UCSD Birds-200-2011 dataset.

PROFESSIONAL AFFILIATIONS, CERTIFICATIONS AND HONORS

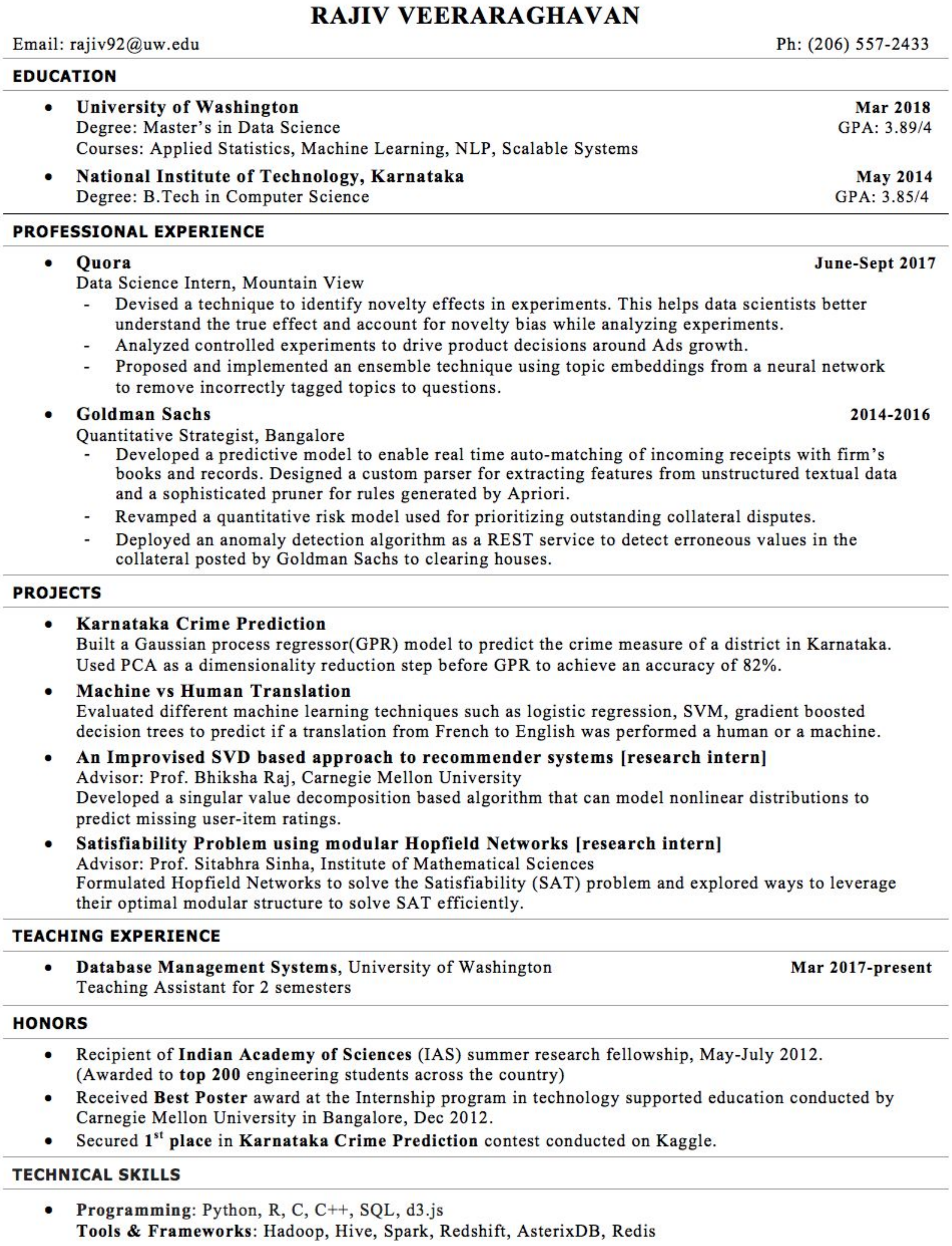
Jul 2017 - Recipient of 2017's Data Science Merit & Opportunity Scholarship offered by the University of Washington's Data Science program
May 2017 - Certified Member of Golden Key International Honour Society validated by University of Washington
Oct 5, 2015 - Dec 27, 2015 - Machine Learning Course (Stanford University) Coursera - (Scored - 96.1%)
2014 - AWS Technical Professional Certification Version 2.0 **and** AWS Business Professional Certification Version 4.0

OTHER INTERNSHIPS

TATA CONSULTANCY SERVICES - Noida, India iON System SMB (Small and Medium Business) Project **May 2013 - Jul 2013**
Project involved Network Integration and deployment of the iON system developed by TCS.

DABUR INDIA LIMITED - Ghaziabad, India Employee Management System Project **Oct 2012 - Dec 2012**
Performed statistical analysis for the core HR process of performance management, based on interdependent and dependent variables.

Rajiv Veeraraghavan



RAJIV VEERARAGHAVAN

Email: rajiv92@uw.edu Ph: (206) 557-2433

EDUCATION

- **University of Washington** **Mar 2018**
  Degree: Master's in Data Science GPA: 3.89/4
  Courses: Applied Statistics, Machine Learning, NLP, Scalable Systems
- **National Institute of Technology, Karnataka** **May 2014**
  Degree: B.Tech in Computer Science GPA: 3.85/4

PROFESSIONAL EXPERIENCE

- **Quora** **June-Sept 2017**
  Data Science Intern, Mountain View
  - Devised a technique to identify novelty effects in experiments. This helps data scientists better understand the true effect and account for novelty bias while analyzing experiments.
  - Analyzed controlled experiments to drive product decisions around Ads growth.
  - Proposed and implemented an ensemble technique using topic embeddings from a neural network to remove incorrectly tagged topics to questions.
- **Goldman Sachs** **2014-2016**
  Quantitative Strategist, Bangalore
  - Developed a predictive model to enable real time auto-matching of incoming receipts with firm's books and records. Designed a custom parser for extracting features from unstructured textual data and a sophisticated pruner for rules generated by Apriori.
  - Revamped a quantitative risk model used for prioritizing outstanding collateral disputes.
  - Deployed an anomaly detection algorithm as a REST service to detect erroneous values in the collateral posted by Goldman Sachs to clearing houses.

PROJECTS

- **Karnataka Crime Prediction**
  Built a Gaussian process regressor(GPR) model to predict the crime measure of a district in Karnataka. Used PCA as a dimensionality reduction step before GPR to achieve an accuracy of 82%.
- **Machine vs Human Translation**
  Evaluated different machine learning techniques such as logistic regression, SVM, gradient boosted decision trees to predict if a translation from French to English was performed a human or a machine.
- **An Improvised SVD based approach to recommender systems [research intern]**
  Advisor: Prof. Bhiksha Raj, Carnegie Mellon University
  Developed a singular value decomposition based algorithm that can model nonlinear distributions to predict missing user-item ratings.
- **Satisfiability Problem using modular Hopfield Networks [research intern]**
  Advisor: Prof. Sitabhra Sinha, Institute of Mathematical Sciences
  Formulated Hopfield Networks to solve the Satisfiability (SAT) problem and explored ways to leverage their optimal modular structure to solve SAT efficiently.

TEACHING EXPERIENCE

- **Database Management Systems**, University of Washington **Mar 2017-present**
  Teaching Assistant for 2 semesters

HONORS

- Recipient of **Indian Academy of Sciences** (IAS) summer research fellowship, May-July 2012. (Awarded to **top 200** engineering students across the country)
- Received **Best Poster** award at the Internship program in technology supported education conducted by Carnegie Mellon University in Bangalore, Dec 2012.
- Secured **1st place** in **Karnataka Crime Prediction** contest conducted on Kaggle.

TECHNICAL SKILLS

- **Programming**: Python, R, C, C++, SQL, d3.js
  Tools & Frameworks: Hadoop, Hive, Spark, Redshift, AsterixDB, Redis