**Customer Segmentation for Variable Annuity Products**

Mobing Zhuang

Sha Li

Yawen Li

Interdisciplinary Data Science Masters Program, University of Washington

Submitted to—

Milliman, Seattle Life Practice

1301 Fifth Avenue, Suite 3800  
Seattle, WA 98101 USA

**Contents**

Executive Summary 3

Statement of Problem 4

Objectives 4

Technical Approach 4

Identifying Needs of Customers 4

Identifying Target Specifications 5

Literature Review 6

Design Concept 10

Project Management 13

Deliverables 13

Communication and Coordination with Sponsor 14

Team Qualifications 15

References 15

Appendix A: Résumés of Team Members 18

Sha Li 18

Yawen Li 19

Mobing Zhuang 20

**Executive Summary**

As the growth of total sales of variable annuity (VA) in U.S. market, insurance companies face big challenges in terms of pricing their products due to the great uncertainties of policyholders’ behaviors. This project offers predictive analysis of customers’ behaviors for our project sponsor Milliman, one of the world’s largest providers of actuarial and related products and services, and provide the sponsor reasonable suggestions and guidance to differentiate profitability of policyholders for their client insurers. To fill our sponsor’s need, we plan to perform various clustering algorithms including existing unsupervised learning packages and self-implement optimization algorithms on this customer segmentation problem. The expected outcomes will include justified number of customer clusters along with the corresponding optimal algorithm with detailed explanations.

**Statement of Problem**

A variable annuity is a contract between a customer and an insurance company. With a variable annuity, the insurance company agrees to make periodic payments to the customer in the future. In the U.S., there are $2 trillion in assets in variable annuity (VA) market under management. It is well known that policyholder behaviors have an enormous impact on how much it will cost insurers to provide a product. Nevertheless, the enormous uncertainty in the VA space around policyholder’s behavior poses risk towards insurance companies, and such risk is not hedgeable. Accurate estimation of policyholder’s behavior will be greatly helpful for insurers’ in-force management. The big challenge in front of most insurers is that most of them have relatively little ability to distinguish between policyholders that will behave in different ways.

This project sponsor Milliman has been a leader in actuarial consulting since 1947. They are independent provider of actuarial and risk management services to the insurance industry and self-insured organizations worldwide. Their groundbreaking work of actuarial, analytical, and data management solutions include the Milliman managed Risk Strategy and the Milliman Sustainable Income PlanTM. The goal for this project requested by sponsor Milliman is to use predictive analytics to understand policyholders’ behaviors and to distinguish policyholders’ profitability for their client insurers.

Specifically, the purpose of this research is to 1) collect and generate significant customer data to identify customers’ behaviors by enhancing the original data from insurance companies with data from other sources, e.g., census data, consumer data, mortgage data, and credit data; 2) develop a robust predictive analytics model based on enhanced data for customer segmentation, and differentiate between their profitability. This research will help insurers to be more informed in predicting policyholders’ behaviors and distinguishing their profitability. Eventually, the results of customer segmentation will guide insurers’ product design and marketing, and make better estimation on how much it costs to provide a current product.

**Design Objectives**

This document proposes to improve on the current methodology of customer segmentation from three aspects:

(1) Implement data pipeline on Azure Data Science Virtual Machines (DSVM)

(2) Improve the current clustering technique, K-means

* Convergence and initial partition: implement algorithms that has reasonable initial partition, fast convergence, and a global search of optimum.
* The number of K: implement algorithms to find optimum number of clusters.
* Robustness: implement algorithms to be resilient to outliers and noise.
* Extension of definition of means: replace means with other metrics in the algorithm

(3) Explore and compare K-means with other nonlinear clustering method

We will use perform nonlinear transformation on the data matrix and compare the clustering results with K-means.

**Technical Approach**

Our technical approach will be designed upon customer needs, and build target specification metrics based on customer needs. In order to achieve our target specifications, we propose a series of data analysis and algorithms methods.

**Identifying Customer Needs**

Our customer Milliman has been a leader in actuarial consulting since 1947. They provide actuarial and risk management services to the insurance industry and self-insured organizations worldwide. One of the needs from insurance industry is to understand policyholders’ behaviors, specifically: how policyholders behave differently and why do they have such different behaviors, is there a way to distinguish between policyholders that behave in different ways. Our customer Milliman is trying to satisfy this need from insurance industry and provide them with a service/product taking advantage of the data sources of Milliman. Eventually, the results of customer segmentation will guide insurers’ product design and marketing, and make better estimation on how much it costs to provide a current product.

**Identifying Target Specifications**

\*Importance from low to high: 1-5

|  |  |  |  |
| --- | --- | --- | --- |
| Metric No. | Need | Metric | Importance |
| 1 | Enriching the original customer data from insurance companies to collect significant information customers’ behaviors. | Improved and enriched datasets merging the original customer data with census data, customer data, mortgage data with missing value imputed. | 3 |
| 2 | Determination of number of segmentations of policyholders depends on their behavior differences. | Conclude numbers of customer clusters and justify the choice of the cluster numbers. | 5 |
| 3 | Well segmented customers with distinct degree of profitability. | Optimized K-means clustering algorithms with improved performance compared to benchmark. | 4 |
| 4 | Optimized segmentation schema | Compare performance of other clustering algorithms with K-means and justify results | 3 |
| 5 | Justifiable segmentation schema. | Performance test of all clustering algorithms with MG-Hedge | 4 |

**Literature Review**

In order to realize the goal of segmenting customers for variable annuity products with different profitability, our project approach is to address it as an unsupervised clustering machine learning problem. In this section, we will review a number of unsupervised clustering algorithms, including their principles, implementation and justification of choice.

* The K-means Algorithm

1. Algorithm overview

K-means clustering is the most widely used partitional clustering algorithm. It starts by choosing K representative points as the initial centroids. Each point is then assigned to the closest centroid based on a specific proximity measure. After the clusters are formed, the centroids for each cluster are updated. These two steps are then iteratively repeats until the centroids do not change anymore or any other convergence criterion is met.

1. Algorithm implementation

A basic K-means algorithm can be implemented using the scikit-learn module. The sklearn.cluster.Kmeans object in the module is solved using Lloyd’s algorithm. Then fit and predict methods can be used to fit train samples and assign results to test samples.

1. Algorithm justification

Advantages: Running time is relatively short. It is suitable for even high dimensional data. K-means is also easy to interpret and implement.

Disadvantages: K-means assumes the clusters as spherical, therefore it does not work efficiently with complex geometrical shaped data. Moreover, misgrouping is likely take place if hard assign data points.

* Fuzzy K-means clustering

1. Algorithm overview

The fuzzy K-means algorithm is an extension of the K-means algorithm for fuzzy clustering. While hard clustering requires each data point belongs to one and only one cluster, it is not feasible in complex datasets where there are overlapping clusters. Fuzzy clustering extends this notion to associate each data point with every cluster using a membership function vary from 0 to 1. Notice that K-means is a special case of fuzzy K-means in case of the data point is closest to a centroid with probability function equal to 1 and 0 otherwise.

1. Algorithm implementation

Processes of initialization, iteration and termination are the same as the ones used in K-means. The algorithm uses a weighted centroid based on the probabilities from the membership function.

1. Algorithm justification

Advantages: The fuzzy K-means can handle the situations in which the data points are ambiguous in between several clusters. For example, it can make conclusions giving the probabilities of the data point in each relative cluster.

Disadvantages: The huge amount of outputs with growing number of objects results in useless information.

* X-means algorithm

1. Algorithm overview

In the K-means algorithm, the number of clusters is an input parameter specified by the user. X-means can be used to reveal the true number of clusters underlying the distribution.

1. Algorithm implementation

The main idea of this algorithm is to optimize the Bayesian information criterion in order to make splitting decisions of the current centroids to better fit the data. In X-means, K value will be assessed by using the Gaussian probabilistic model and the maximum likelihood estimates.

1. Algorithm justification

Advantages: The number of clusters K can be estimated very fast with a fast inner-loop for expensive algorithms dealing with high dimensional data sets.

* The K-harmonic Means Algorithm

1. Algorithm overview

In the K-means algorithm, one of the major problems is its performance depends on the initialization of the centers of the clusters a lot. This algorithm is developed from the K-means algorithm and it is not sensitive to the initialization of centers.

1. Algorithm implementation

The K-harmonic means algorithm replace the minimum function in the error function of the K-means with the harmonic average function.

1. Algorithm justification

Advantages: K-Harmonic Means is very insensitive to the initial centers of clusters and it converges faster than K-Means when the initialization is far from a local optimal of K-Means.

Disadvantages: So far, we cannot justify K-Harmonic Means performs well on a very large datasets. Also, unlike the performance function of K-means, K-Harmonic Means is more intuitive without providing statistical interpretations.

* The GMM Algorithm

1. Algorithm overview

A Gaussian mixture model (GMM) is a model of a distribution as a mixture of K separate multivariate normal distributions, each with individual parameters represented collectively by θ. The probability density function of the model given the parameters is thus:

where xi is a D dimensional point, N denotes the normal probability distribution of an observation, and αk, μk, and ∑k, are the amplitude, mean, and covariance matrix of each Gaussian component. In this model, each point observed has some probability of having been generated by each component. These are called the responsibilities rik of the point xi for each of the K components. Having fit a GMM to a data set and deriving the responsibilities, a hard clustering for each point can be found by assigning that point to the component whose responsibility is highest.

1. Algorithm implementation

GMM can be implemented using the scikit-learn module. The sklearn.mixture. GMM object in the module is an implementation of the expectation-maximization (EM) algorithm for fitting Gaussian Mixture Models. Then fit and predict methods can be used to fit train samples and assign results to test samples.

1. Algorithm justification

Advantages: GMM does not assume clusters to be of any geometry and it works well with non-linear geometric distributions. It also does not bias the cluster sizes to have specific structures as does by K-Means.

Disadvantages: GMM uses all the components it has access to, so initialization of clusters will be difficult when dimensionality of data is high, and it is usually difficult to interpret.

One can think of GMM as generalising K-means clustering to incorporate information about the covariance structure of the data as well as the centers of the latent Gaussians.

* Hierarchical clustering

1. Algorithm overview

Hierarchical clustering algorithms were developed to overcome some of the disadvantages associated with partitional-based clustering (such as K-means) methods. Partitional methods generally require a user predefined parameter K to obtain a clustering solution, while hierarchical algorithms were developed to build a more deterministic and flexible mechanism for clustering the data objects. A hierarchy can be understood as an analogy to a standard binary tree. The root represents all the sets of data objects to be clustered (level 0). The child entries are the subsets of the entire dataset corresponds to some set of clusters. HIerarchical methods can be categorized into agglomerative and divisive clustering methods. Agglomerative methods adopt a bottom-up strategy, start from singleton clusters at the bottom level and continue merging two clusters at a time. Divisive methods, on the other hand, adopts a top-down strategy, start with all the data objects in a huge macro-cluster and split it continuously into two groups.

1. Algorithm implementation

Agglomerative algorithm can be implemented using the scikit-learn module. The sklearn.cluster.Agglomerative Clustering object in the module is an implementation of agglomerative algorithm. Then fit and predict methods can be used to fit train samples and assign results to test samples.

1. Algorithm justification

Advantages: It allows for cutting the hierarchy at any given level and obtaining the clusters correspondingly. Therefore, it does not require a predefined user specified parameter K. It also can use non-Euclidean distances for distance measurements.

Disadvantages: The merge or split decisions once made at any given level in the hierarchy cannot be undone and therefore it is not that flexible. The runtime of hierarchical algorithm is also much longer than K-means.

All above mentioned clustering techniques have their advantages and disadvantages in solving customer segmentation problems. K-means method is the most popular for its interpretability, especially in financial and insurance industry. Extensions of K-means algorithm, e.g. fuzzy K-means clustering, X-mean, and K-harmonic Means are also widely studied to enhance the segmentation performance of K-means methodology. Besides K-means algorithms and its extensions, other less interpretable methods are also applied to customer segmentations for their unique advantage over K-means, especially when handling complex data. GMM has greater flexibility due to clusters having unconstrained covariances. And it allows soft assignments, meaning that a customer can belong to two clusters with varying degree (probability). Hierarchical clustering has the advantage of not predefining the number of clusters. And it helps us visualize how the data may be clustering together by generating a plot called the dendrogram. However, it was reported to not performing very well with large data set. There are a few studies on two-stage clustering that combine hierarchical and K-means method. Essentially, initial clusters are determined by hierarchical analysis and then K-means is employed to discover the final clusters. Such integration may give powerful clustering results, especially with large data sets.

**Design Concept**

* Data Pipeline

The raw data sets include policy holder’s information from insurance company, their census data, consumer data, mortgage data, and credit data will be copied to DSVM by data engineers at Milliman. We will perform the following data processing and analysis on this great platform, a pre-configured environment in the cloud for data science. The plentiful options of languages, tools of development, data ingestion, data visualization, and machine learning will provide great flexibility for the project. In addition, data security issue will not be of major concern since the data is hosted on Azure.

We will build a data pipeline by writing highly reproducible code to process the raw data sets and ingest the final data set to platforms like Spark, Hadoop, or SQL server for storage and subsequent analysis.

In the process of building data pipeline, we will also take care of the missing data. The typical approaches in data analysis are to simply remove any rows with missing entries from the matrix or impute a mean or median value of the corresponding variable. We plan to use generalized low rank model (GLRM) to perform missing value imputation. Essentially, this method “borrow strength” from the entries that are not missing to improve our global understanding of the data matrix. GLRM has been implemented on many platforms/languages that are available in DSVM, e.g. H2O, Spark, Python, and Julia.

* Data Source

1. Census Data

Census tract and block group level data will be used in this study. We will geocode the addresses associated with each VA contract to identify the corresponding census tract and block group.  
 We will focus on the following variables:  
 Population density  
 Demographic - distribution by gender, age, education, and household type  
 Income - distribution and percentiles, per capita average, distribution by type(wage, self-employment, social security, etc.), percent below poverty level  
 Housing - percent owner occupied, distribution by value of owner occupied units, mortgage status, and mortgage amounts by age  
 Workforce - distribution and by occupation and industry

1. Mortgage Data

The mortgage data using in this study is originally data and analytics for the mortgage and real estate industries. The data is compiled from county property tax and deed records, as well as mortgage lenders. Milliman provided addresses to the mortgage data provider and receive the matching property data.   
 We will focus on the following variables:  
 Loans - the number of active loans, loan types, loan terms, loan ages, amount borrowed, estimated payments and outstanding loan amounts, estimated loan to value  
 Property - estimated value, change in value, value per square foot, age of home, years owned, lot size  
 Neighborhood - average loan to value, property value, year built, etc. for the ZIP code

1. Consumer Data

Consumer data includes various marketing scores and household lifestyle/lifestage clusters. Consumer data provider used names and addresses from the policy data to match data from their database.

We will focus on the following variables:  
 Lifestyle - health conditions (e.g. anxiety, arthritis), interests (e.g. golf, travel, fitness)  
 Financial - estimated income, net worth, liabilities, types of credit cards and loans  
Property - dwelling type, mortgage amount, home sale price, home equity,home market value, refinance activity, loan to value  
 Marketing scores - general, automotive finance, credit card  
 Census - percent owner occupied, percent families, median school years completed, percent single family homes, median home value

1. Credit Data

Credit data is compiled from lenders such as credit card issuers, auto finance companies, and mortgage lenders as well as collection agencies and courts. Credit data provider used policyholder name, addresses, date of birth, and last 4 digit of SSN to match on the credit records.  
 We will focus on the following variables:  
 Length of credit history - number, age, types of accounts  
 Recent credit seeking - number of inquiries, number of recently opened accounts  
 Utilization - balance of limits, changes in balances  
 Payment history - number of times past due, number of charge-offs, number of collections  
 Public records - bankruptcies, liens  
 Credit scores

* K-means and Improvement on K-means

K-means is chosen as the main clustering method to be researched because it is the most popular, efficient, and interpretable method for customer segmentation problems. While traditional K-means has many desirable properties as a clustering technique, it also suffers several drawbacks, particularly the inherent limitations of its optimization logic of convergence. We plan to modify and build more efficient algorithms on top of the K-means algorithm that typically found in a standard R or Python K-means package from the following aspects:

1. Convergence and initial partition

The first problem with K-means is that iteratively optimal procedure cannot guarantee the global convergence to a global optimum. This means, the initial assignment of centroids may affect the result significantly, thus the assignment of final cluster could be highly unstable for some data points. Reasonable initial partition, fast convergence, and a global search will be implemented.

1. The number of K

The problem of estimating the correct number of clusters is one of the major challenges for K-means, it is not ideal to have too many clusters in this analysis. We plan to research on this topic and write algorithms to help identify optimized number of clusters.

1. Robustness

K-means is sensitive to outliers and noise. The calculation of centroids is based on all data points in the cluster, including the outlier. Therefore, forcing all points into a cluster distorts the shape of clusters and misrepresent the prototype. Possible solutions to this problem is to use K-medoids which is more resilient to outliers.

1. Extension of definition of means

The definition of means limits the application of K-means only to numerical variables. Even if categorical variables are encoded to be numerical, the obtained means may not have a physical meaning or may be difficult to interpret. Possible solutions to this problem would be replacing means with medians or modes, the overall outline of the algorithms are similar, but medians or modes will be used as centroids.

* Other Clustering

Apart from K-means clustering, we’ll also test two other nonlinear clustering methods: the GMM( Gaussian mixture model) algorithm and hierarchical algorithm on the customer segmentation problem. Compared to K-means, GMM does not bias the cluster sizes to have specific structures and hierarchical algorithm does not predefine the number of clusters (K). One of our deliverables is determine the number of clusters and justify the choices. Using GMM and hierarchical algorithms will help us test our hypothesis on customer cluster numbers and sizes. They will also serve as comparable tests to gauge our K-means algorithms performance.

**Project Management**



**Figure 1:** Gantt chart for the project. The bars indicate the timeline of the tasks. (see attached project\_plan.xlsx).

In order to ensure the project is able to be delivered on time, we specify our desired deliverables and weekly plan in the following sections. Note that both deliverables and schedule may change to one degree or another upon sponsor's feedbacks and time remaining.

**Deliverables**

1. Improved and enriched datasets with missing value imputed using generalized low rank model (GLRM) algorithm.
2. Conclude numbers of customer clusters and justify the choice of the cluster numbers.
3. Optimized clustering algorithms with reasonable justifications.
4. Clustering analysis code files with detailed documentation as well as test results on clustered customer profitability using MG-Hedge.
5. Well documented project report including data processing, data analysis, algorithm optimization and machine learning processes, as well as clustering results.
6. Well organized final project presentation/poster summarizing the project work and results.

**Communication and Coordination with Sponsor**

Communication schedule, objectives and format with the sponsor.

|  |  |  |
| --- | --- | --- |
| Date | Communication Objective | Format |
| Week 1 | Initial sponsor meet with all group members; Problem statement by the sponsor; Project plan discussion and agreement on both sides | Email  In-person meeting  Conference calls |
| Week 2-3 | Team accomplish project proposal and seek for sponsor’s feedback | Email  Conference calls |
| Week 4-6 | Team accomplish data preparation, merge and clean and provide updates to sponsor | Email |
| Week 7-10 | Team accomplish algorithm development and seek for sponsor’s support for algorithm performance test as well as feedback | Email  Conference call |
| Week 11 | Mid-term meetup and updates with the sponsor | Office visits |
| Week 12-14 | Team accomplish algorithm optimization and seek for sponsor’s support for algorithm performance test as well as feedback | Office visits  Email  Conference call |
| Week 15 | Team accomplish project result summary and presentation preparation | Email |
| Week 16 | Team accomplish final project touch on presentation and report | Email |
| Week 17 | Team give final project presentation and report to the sponsor and seek for feedback | Office visits  Email |

**Team Qualifications**

Sha Li: Master of Data science student at UW, with a prior Ph.D in biophysical chemistry. Her data science core competencies are data analysis with python and SQL, regression and classification machine learning, as well as data visualization with Tableau. She will also bring business analytic insights to the team from her intern experience at Zillow.

Yawen Li: Master of Data science student at University of Washington - Seattle. She earns her Bachelor’s Degree in Mathematics with a minor in Applied Mathematics. She is familiar with doing data analysis with python and R. In particular, her prior experience in developing feature selection algorithms will bring team more ideas in the process of data imputation and modeling.

Mobing Zhuang: Master of Data science student at UW, with a prior Ph.D in environmental engineering. She is skilled in analyzing data using R and Python, applying machine learning algorithms to solve business problems, and developing web applications using R Shiny. Her internship experience at Milliman, especially her familiarity with VA product and data set, will contribute a lot to the team.

**References**

1. Ryan Maas, Jeremy Hyrkas, Olivia Grace Telford, Magdalena Balazinska, Andrew Connolly, and Bill Howe. 2015. Gaussian Mixture Models Use-Case: In-Memory Analysis with Myria. In Proceedings of the 3rd VLDB Workshop on In-Memory Data Mangement and Analytics (IMDM '15). ACM, New York, NY, USA, , Article 3 , 8 pages. DOI=<http://dx.doi.org/10.1145/2803140.2803143>.
2. F. Pedregosa et al. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12:2825–2830, 2011.
3. <http://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html>
4. <http://scikit-learn.org/stable/modules/mixture.html>
5. Guojun Gan, Chaoqun Ma, and Jianhong Wu, Data Clustering: Theory, Algorithms, and applications, ASA-SIAM Series on Statistics and Applied Probability, SIAM, Philadelphia, ASA, Alexandria, VA, 2007, pp. 151-173
6. Franck, Dernoncourt, and Vaibhav, Agrawal, editors. What is the difference between K-Means and Fuzzy-C Means Clustering. Quora, <https://www.quora.com/What-is-the-difference-between-K-Means-and-Fuzzy-C-Means-Clustering>. Accessed 09 Nov. 2017.
7. Boldeanu, Silviu.Fuzzy Clustering, Babes-Bolyai University[internet], Volume 1: 6-7. Web. 09 Nov. 2017. Available From: <https://pdfs.semanticscholar.org/13e4/3b4710620fb3e6e303d612f541e8e74777d6.pdf>
8. P. Lloyd, Stuart. (1982). Least Squares Quantization in PCM's. IEEE Transactions on Information Theory. 28. 129-136. 10.1109/TIT.1982.1056489.
9. MacQueen, J. Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics, 281--297, University of California Press, Berkeley, Calif., 1967.
10. <https://azure.microsoft.com/en-us/services/virtual-machines/data-science-virtual-machines/>
11. Berkhin P. (2006) A Survey of Clustering Data Mining Techniques. In: Kogan J., Nicholas C., Teboulle M. (eds) Grouping Multidimensional Data. Springer, Berlin, Heidelberg.
12. A.K.Jain, M.N.Murty, and P.J.Flynn. Data clustering: A review. ACM Computing Surveys (CSUR), 31(3):264-323, 1999.
13. Kuo, R. J., An, Y. L., Wang, H. S., & Chung, W. J. (2006). Integration of self-organizing feature maps neural network and genetic K-means algorithm for market segmentation. Expert Systems with Applications, 30, 313–324.
14. Abdulkadir Hiziroglu, Soft computing applications in customer segmentation: State-of-art review and critique, In Expert Systems with Applications, Volume 40, Issue 16, 2013, Pages 6491-6507, ISSN 0957-4174, <https://doi.org/10.1016/j.eswa.2013.05.052>.
15. Jandaghi, G. & Moradpour, Z. (2015). Segmentation of Life Insurance Customers Based on their Profile Using Fuzzy Clustering. International Letters of Social and Humanistic Sciences. 61, pp. 17-24. SciPress.
16. Fadly Hamka, Harry Bouwman, Mark de Reuver, Maarten Kroesen, Mobile customer segmentation based on smartphone measurement, In Telematics and Informatics, Volume 31, Issue 2, 2014, Pages 220-227, ISSN 0736-5853, <https://doi.org/10.1016/j.tele.2013.08.006>.
17. Hosseini, Monireh and Shabani, Mostafa, New approach to customer segmentation based on changes in customer value, Journal of Marketing Analytics, Volume 3, 2015, Pages 110--121, ISSN 2050-3326, <https://doi.org/10.1057/jma.2015.10>.
18. Huang Zhexue, Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values, Data Mining and Knowledge Discovery, Volume 2, Issue 3, 1998, Pages 283--304, ISSN 1573-756X, <https://doi.org/10.1023/A:1009769707641>.
19. Balaji, S., & Srivatsa S. K. (2012). Customer segmentation for decision support using clustering and association rule based approaches. International Journal of Computer Science and Engineering Technology, 3(11), 525-529.

|  |  |
| --- | --- |
|  |  |

**Appendix A:**

**Résumés of Team Members**

|  |  |
| --- | --- |
|  | Sha Li |

|  |  |
| --- | --- |
|  | **Core competencies**   * Machine Learning * Data Analysis with SQL, Python * Data Visualization with Tableau   **Unique experience**   * Internship at Zillow Pricing Analytics * Ph.D Biophysical Chemistry   **Areas of Interest**   * Business Analytics * Deep Learning Applications * Recommendation Engine * Corgi |

Yawen Li

## Core competencies

* R
* Python
* SQL

## Secondary Skills

* Machine Learning
* Statistical/Mathematical Modeling
* Java
* Data Visualization

## ABOUT MYSELF

* Detailed Oriented Person
* Mathematics Background
* I gained great experience in machine learning, feature selection and its application to the biological sciences during the summer internship.
* The study abroad experience strengthens my communication skills with peers as well as adaptability.

## aREA OF INTEREST

* Machine Learning Applications
* Medical Records Analysis
* Urban Life Problems

## Expectation

* Improve data visualization skills by dealing with more complex datasets and try to use other ways such as d3 to visualize data if possible
* Do more coding in object-oriented programming language, such as Java



***Mobing Zhuang***

My Skills:

* Proficient: R, Python, SQL, Tableau
* Prior experience: HTML, CSS, JavaScript
* Data wrangling and visualization
* Descriptive and inferential statistical analysis

My Unique Aspects:

* Shiny app developer
* Background in environmental engineering
* Agile project planning

My Interests:

* Advanced clustering methods applied to customer segmentation problem
* Solving problem using machine learning techniques
* Environment related topics