HealthSCOPE: An Interactive Distributed Data Mining Framework for Scalable Prediction of Healthcare Costs

James Marquardt*, Stacey Newman*, Deepa Hattarki*, Rajagopalan Srinivasan*, Shanu Sushmita*, Prabhu Ram†, Viren Prasad†, David Hazel*, Archana Ramesh*, Martine De Cock*‡, Ankur Teredesai*
*Center for Data Science, Institute of Technology
University of Washington Tacoma
1900 Commerce Street, Tacoma WA-98402, USA
Email: {jamarq, newmsc8, hattd, velamurr,sshanu, dhazel, aramesh2, mdecock, ankurt}@uw.edu
†Edifecs
2600 116th Avenue Ne #200, Bellevue WA-98004, USA
Email: {prabhu.ram,VirendraP}@edifecs.com
‡Dept. of Applied Mathematics, Computer Science and Statistics
Ghent University
Krijgslaan 281 (S9), 9000 Gent, Belgium
Email: martine.decock@ugent.be

***Abstract*—In this demonstration proposal we describe HealthSCOPE (Healthcare Scalable COst Prediction Engine), a framework for exploring historical and present day healthcare costs as well as for predicting future costs. HealthSCOPE can be used by individuals to estimate their healthcare costs in the coming year. In addition, HealthSCOPE supports a population based view for actuaries and insurers who want to estimate the future costs of a population based on historical claims data, a typical scenario for accountable care organizations (ACOs).**

Using our interactive data mining framework, users can view claims (sample files will be provided), use HealthSCOPE to predict costs for the upcoming year, interactively select from a set of possible medical conditions, understand the factors that contribute to the cost, and compare costs against historical averages. The back-end system contains cloud based prediction services hosted on the Microsoft Azure infrastructure that allow the easy deployment of models encoded in Predictive Model Markup Language (PMML) and trained using either Spark MLLib or various non-distributed environments.

Keywords*—*Healthcare cost prediction, insurance claims data, distributed data mining, Spark, Microsoft Azure, PMML

I. Introduction

Healthcare cost prediction is of immense importance to improve accountability in care. According to the World Health Organization, healthcare costs for 2012 in the United States were highest across the world; both per-capita ($8,233) as well as percentage of the Gross Domestic Product (17.6%). Yet, amongst comparable nations, the United States rank towards the bottom in terms of quality of care (1). With a goal of changing this, healthcare reform policies are currently underway, promoting initiatives for managing the overall health of a population while keeping costs manageable (2). Important challenges for this reform are to leverage existing large and varied clinical and claims datasets to estimate future healthcare costs, and to take measures in care-management that reduce such costs while improving overall population health. Data used for cost prediction models are often voluminous, diverse, and vary significantly over time. As the amount and complexity of data increases, so does the challenge to store, retrieve and manipulate. A scalable healthcare cost prediction system should aim to provide high-availability, fault-tolerance and fast-processing of a large number of claims so that the overall quality of the services is not affected. Handling such diverse, voluminous, and dynamic data is still a challenge for most of the existing healthcare cost prediction algorithms, and to the best of our knowledge a framework providing distributed algorithms and corresponding web based APIs to support individual as well as population level analytics has not been demonstrated.

In this demonstration proposal, we describe HealthSCOPE (*Health*care *S*calable *CO*st *P*rediction *E*ngine) – an interactive predictive modeling framework for the exploration of healthcare costs. HealthSCOPE allows for fast analysis and estimation of costs, as well as ease of predictive model deployment. It is a generic healthcare cost prediction framework that allows individuals to upload their healthcare history and understand costs. HealthSCOPE can make cost predictions for a population of beneficiaries[1]. End users (typically managers at Accountable Care Organizations – ACOs, or benefit managers at insurance companies) can upload basic demographics, recent diagnoses, and comorbidities of a population for the past year and browse through population characteristics, estimates of the the healthcare costs associated with the next year and the factors that contribute most to high costs. This second use case scenario caters to hospitals and insurance providers who are interested in understanding costs associated with a heterogeneous population, given a history of insurance claims for that population.

[1]A beneficiary, in the context of this proposal, is an individual (patient) who receives the benefits of a health insurance contract.

Previous efforts and tools are restricted both in their dependence on linear regression (which cannot handle high dimensional and varied data), and rule based approaches (which require a lot of domain knowledge and on-premise compute infrastructure typically expensive to scale), but more recently, data mining approaches are also being explored for this purpose (3). However, to the best of our knowledge regression trees have not been used for the cost prediction problems. HealthSCOPE benefits from regression trees and exhibits better performance when compared to a linear regression model. Further, we are not aware of any claims based cost prediction as a service implementations (either for individuals or for populations).

HealthSCOPE is powered by Spark[2], which provides fast and parallel processing of large scale data through cluster computing. A useful feature of HealthSCOPE is that it allows consumers, health plan portals, or ACOs to simply upload their own claims data and use scalable implementations of supervised machine learning methods for prediction and exploration purposes. Additionally, HealthSCOPE allows further development and deployment of predictive models trained using Spark MLLib. HealthSCOPE provides the capability to export predictive models trained using Spark MLLib to PMML[3].The rest of the paper is structured as follows: In Section II we present the user interface of HealthSCOPE and how it can be used by the end user. The design of the back-end system is described in Section III followed by demonstration logistics. Related work is addressed as relevant throughout the proposal.

II. HealthSCOPE Features and Demonstration Overview

HealthSCOPE provides a framework for exploring historical and present day healthcare costs, as well as for predicting future costs. HealthSCOPE can be used by individuals to estimate their healthcare costs in the coming year. In addition, HealthSCOPE supports a population based view for actuaries and insurers who want to estimate the future costs of a population based on historical claims data, a typical scenario for accountable care organizations (ACOs). In this proposal, due to space limitations, we describe the ACO user scenario.

A. Population Level View

There is growing interest in reducing overall healthcare costs, and one of the best ways to control medical costs is by improving the overall health of the patient population (4). To this effect, various payment models are being explored. One model that has gained traction is called Accountable Care Organizations or ACOs. An ACO is often a group of physicians and other healthcare organizations such as hospitals who form an entity and work together with Medicare or commercial healthcare insurance companies to provide care to a population of patients. ACOs share both financial and medical responsibility for providing coordinated care to patients in hopes of limiting unnecessary spending. ACOs coordinate care for their patients on an individual basis and reach across medical specialties and care settings. By managing the health of the overall patient population and the health of individual high risk patients, ACOs aim to achieve cost savings.

Fig. 1. Screenshot showing combined and average predicted costs for the group of beneficiaries across different age groups. The green arrow indicates the form for uploading healthcare data to be evaluated, the red arrow indicates navigation options for different population visualizations, and the yellow arrow indicates population cost predictions. In the current visualization, it can be seen that for the beneficiaries over age 65, future healthcare cost (yellow bubble) is predicted to be higher that the previous year cost (blue bubble).

In order to cater to the population level responsibilities of ACOs, HealthSCOPE generates a combination of visualizations and cost predictions deemed valuable for analysis. Upon loading health data relevant to a particular population using the upload button (as marked with a green arrow in the Figure 1), the ACO user will see combined and average predicted costs for the group of beneficiaries being evaluated. This information is highly valuable in gauging the overall cost risk of a particular population. Several charts are also generated to convey what factors contribute to the overall predicted costs. These charts detail the breakdown of costs with respect to groupings of beneficiaries in terms of age, race, income level, and residence location. An example breakdown of predicted cost and historic cost across different age groups can be seen in Figure 1. Such a breakdown across different groups can help to identify groups that may cause an increase in overall healthcare cost in future.

B. Beneficiary Level View

As an extension to their population level exploration, analysis at the beneficiary level is also very valuable. By identifying high-risk beneficiaries, ACOs and providers can focus more attention on specific beneficiaries to improve overall health. In order to facilitate this level of analysis, HealthSCOPE identifies beneficiaries within a population who are forecast to have a particularly large increase in cost in future years. Upon selecting a beneficiary to evaluate more closely, the user is presented with the factors that make the beneficiary in question costly, as seen in Figure 2. In particular, actionable items such as comorbidities are highlighted. As a function to demonstrate how certain interventions will affect

[2]https://spark.apache.org

[3]Predictive Model Markup Language (PMML) provides an open standard for representing data mining models. In this way, models can easily be shared between different applications avoiding proprietary issues and incompatibilities. Currently, all major commercial and open source data mining tools already support PMML.

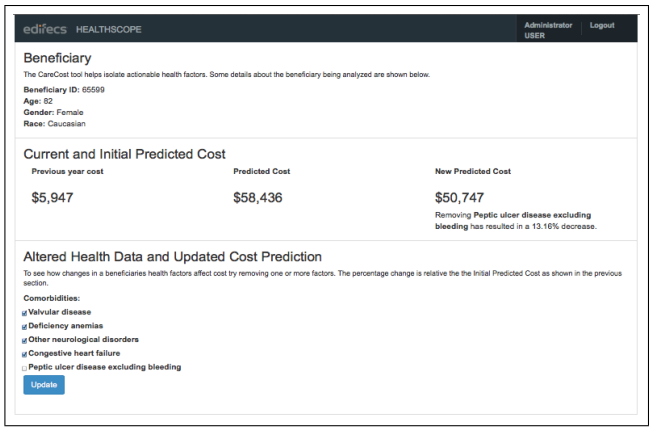

Fig. 2. Screenshot showing the individual level view reached from the population level evaluation.

the beneficiary's forecasted cost, the user is able to remove particular comorbidities from the beneficiary's record. Upon completing these modifications to the record, an updated cost prediction is generated for the beneficiary. For instance, in Figure 2 it can been be seen that the comorbidity "peptic ulcer" is one of the contributing factors in the overall cost increase for the selected beneficiary. The user, can uncheck this comorbidity and get an estimate on how this intervention will affect the beneficiary's forecasted cost. In this case, the new prediction shows a drop of 13% in the overall forecasted cost.

III. HealthSCOPE Design and Internals

The overall architecture of HealthSCOPE is shown in Figure 3. The framework consists of a user interface (UI), a cost prediction API (which is hosted on Microsoft Azure) and a Cost Prediction Engine (CPE). Next, we describe these modules.

a) User Interface (UI): The user interface shown in Figure 1 allows the user to upload data (in csv) with details like age, race, gender, location (through zip code), chronic conditions, and so on. Sample files for the demonstration are provided. After upload the UI provides a predicted overall cost and basic gender breakdown on the right. It also indicates how much of an increase or decrease from total cost for the population one can expect. In addition there are several visualization options for exploring the underlying uploaded data and the cost predictions (see Section II, Figure 1 and 2).

b) Cost Prediction API: The entire service is accessed through Representational State Transfer(REST) APIs using HTTP calls. Data entered by the user is parsed into comma separated values by this module, and is then sent to the model selector. In some user scenarios (e.g, ACO user) this layer also computes the population aggregations and visualization statistics. The Cost Prediction API collects the predicted costs from the cost prediction engine and sends the predicted values back to the user interface.

c) Cost Prediction Engine: We utilize ADAPA (Adaptive Decision and Predictive Analytics)[4] to perform beneficiary scoring. The Cost prediction Engine consists of the following additional sub-modules:

1) *Model Selector:* This sub-module is configurable to send the incoming input variables from the Cost Prediction API to one of the appropriate prediction models available in the Model Bank.
2) *Model Bank:* ADAPA is used to load and deploy predictive models in PMML format for scoring. Currently, there are two prediction models available in the Model Bank: Linear Regression and Regression Trees. Both models are trained using the R statistical computing language, and using MLlib in Spark. For evaluation purposes, the performance and error of each model is compared to a mean baseline, and linear regression model. We use linear regression as an additional baseline due to its extensive use in the literature of healthcare cost prediction. Our models have shown improvement (lower prediction errors) over linear regression as well as average baseline models. In our ongoing research, we continue to explore additional algorithms like weighted KNN, SVM, Naive Bayes, etc. Our future goal is to add more trained models to the model bank collection.
3) *Big Data Stack:* The Big Data stack is powered by Spark, which provides fast and distributed processing of large scale data on a cluster of commodity hardware.

The prediction models in the model bank of HealthSCOPE's Cost Prediction Engine are trained on historical insurance claims data. While healthcare datasets are often used for predicting future healthcare cost, the goal varies from predicting individual cost, to estimating total population healthcare costs (5; 6; 3). The underlying data for building these predictive models often come from claims data, clinical data and/or self-reported data (i.e., questionnaires). These datasets are sometimes used separately (e.g., CDPS[5] which uses claims data, or PRA[6] which uses self-reported data) or as a combination of one or more datasets (e.g., Dorr's algorithm from Care Management Plus[7] uses claim and questionnaire data). Although the predictive power of claims data is often challenged, its utility has been established through several dedicated studies (3; 7). Furthermore, in the context of building a healthcare cost prediction system for individuals, being able to leverage claims data is beneficial since often the privacy concerns associated with clinical data (such as lab results, vitals, etc.) are far more constrained than those associated with claims data.

HealthSCOPE is currently trained on the State Inpatient Database (SID) of the state Washington. This database is part of the family of databases developed for the Healthcare Cost and Utilization Project (HCUP)[8]. The dataset consists of inpatient discharge records from community hospitals in the State of Washington with all-payer, encounter-level information from 2011 to 2012. The data consists of 650,000 records

[4]http://zementis.com/products/adapa/

[5]http://cdps.ucsd.edu

[6]https://www.highmarkblueshield.com/pdf_file/ger_binder/pra-overview.pdf

[7]http://caremanagementplus.org

[8]http://www.hcup-us.ahrq.gov/db/state/siddbdocumentation.jsp

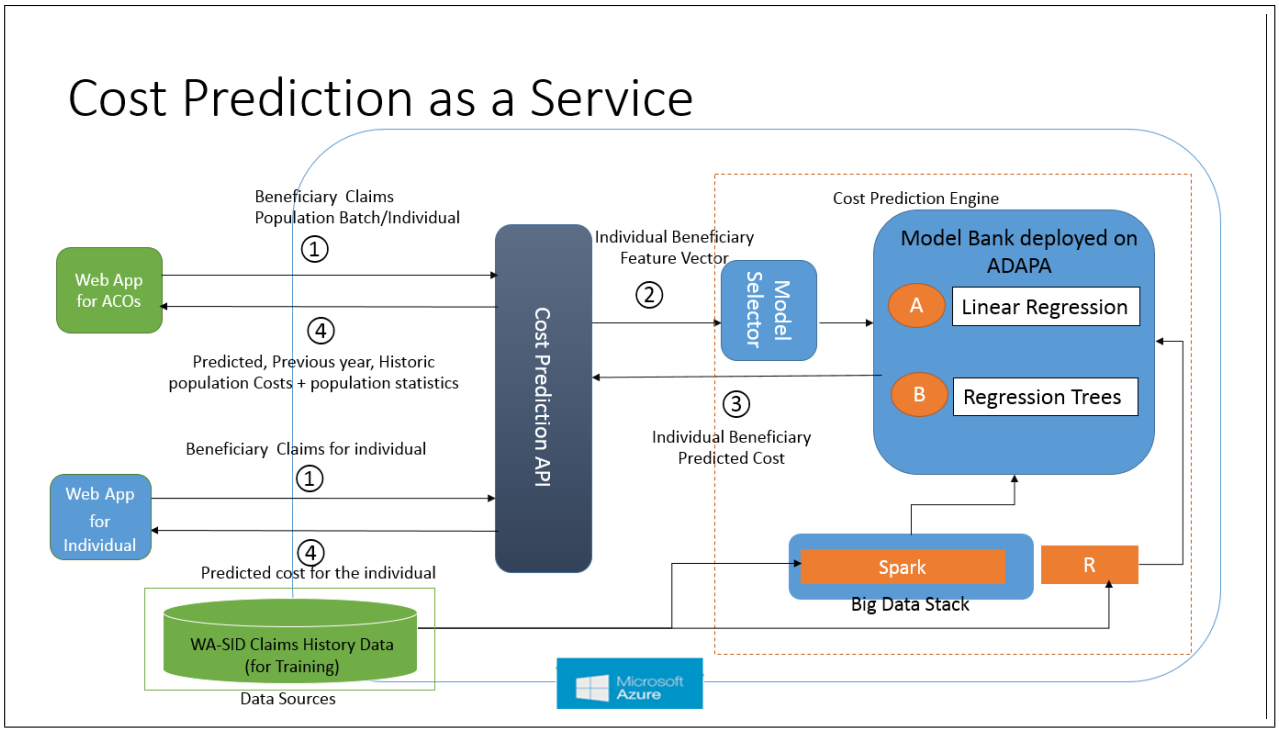

Fig. 3. Overall architecture of HealthSCOPE. A user can enter (upload) data with details like age, race, gender, location (through zip code), and chronic conditions through a Web UI. The data submitted by the user is then sent to the Cost Prediction API. The API parses the data into comma separated values and sends it to the Model Selector. Based on the configuration in the Model Selector the parsed values are sent to one of the prediction models in the Model Bank. The predicted values are sent back through the Cost Prediction API to the user. The whole API is deployed in Microsoft Azure.

per year corresponding to 480,000 beneficiaries with over 1000 attributes. Some of the example attributes are shown in Table I. More detailed information about the data and its features is available on the HCUP-SID website[9].

Attributes
Principal and secondary diagnoses and procedures
Hospital utilization codes and charges
Patient demographics characteristics (e.g., sex, age, and, for some states, race)
Total charges
Length of stay

TABLE I. EXAMPLE ATTRIBUTES PRESENT IN THE SID DATASET.

DEMONSTRATION LOGISTICS

The population and individual level interfaces for our framework can be found at http://tinyurl.com/healthscope-aco and http://tinyurl.com/healthscope-indiv respectively. For the purposes of this conference, we will provide a laptop on which to access the application as well as a projector. We will require access to a table large enough to accommodate both the laptop and the projector, as well as access to power outlets for both.

ACKNOWLEDGEMENTS

We would like to thank Dwaine Trummert for the UI Design, as well as Ila Nejati for their helpful suggestions. We would also like to thank the Azure For Research[10] program from Microsoft Research Connections for the compute resources grant enabling us to use Azure infrastructure for this research, Edifecs Inc., for their generous support to the Center for Data Science, and ADAPA for the donation of licenses for their software.

[9]http://www.hcup-us.ahrq.gov/sidoverview.jsp
[10]http://azure4research.com/

REFERENCES

[1] S. Woolf and L. Aron, Eds., *U.S. health in international perspective: Shorter lives, poorer health*. National Academies Press (US), 2013.
[2] "Key features of the affordable care act," http://www.hhs.gov/healthcare/facts/timeline, accessed on June 25, 2014.
[3] D. Bertsimas, M. V. Bjarnadóttir, M. A. Kane, J. C. Kryder, R. Pandey, S. Vempala, and G. Wang, "Algorithmic prediction of health-care costs," *Operations Research*, vol. 56, no. 6, pp. 1382–1392, 2008.
[4] "Return on investments in health," http://www.rwjf.org/content/dam/farm/reports/issue_briefs/2013/rwjf72446, accessed on June 27, 2014.
[5] J. L. Meyers, S. Parasuraman, K. F. Bell, J. P. Graham, and S. D. Candrilli, "The high-cost, type 2 diabetes mellitus patient: an analysis of managed care administrative data," *Archives of Public Health*, vol. 72, no. 1, p. 6, 2014.
[6] M. Seid, J. W. Varni, D. Segall, and P. S. Kurtin, "Health-related quality of life as a predictor of pediatric healthcare costs: a two-year prospective cohort analysis," *Health and quality of life outcomes*, vol. 2, no. 1, p. 48, 2004.
[7] Y. Zhao, A. Ash, R. Ellis, J. Ayanian, G. Pope, B. Bowen, and L. Weyuker, "Predicting pharmacy costs and other medical costs using diagnoses and drug claims." *Medical Care*, vol. 43, no. 1, pp. 34–43, 2005.