**PROJECT DOCUMENT**

**PROJECT 1: SOFT PA**

Most health plan Drug Formularies have procedures to limit or restrict certain drugs. This is done to encourage doctors and patients to use these drugs appropriately, for the safety and best interest of all members. Most often, these drugs have safety issues, a high potential for inappropriate use, or have lower-priced alternatives on the formulary. These drugs must meet specific criteria for use before they will be considered a covered benefit. During the pre-authorization process, your doctor will obtain approval from Arise Health Plan for you to receive coverage for a drug on the formulary. The process usually involves these steps:

1. Your practitioner may be required to send us certain medical information to help us make a decision.
2. Your practitioner’s office and you are notified as to whether or not the drug is approved.
3. If approved, the drug is prescribed to you.
4. If a drug pre-authorization has been denied, or not submitted, your pharmacy will not be able to file the drug claim under your prescription benefit, so you will be responsible for the entire cost of the prescription.

If a member would like to initiate the pre-authorization process with their provider, he/she may print this form and give it to the provider.

**List**: https://www.wpsic.com/files/drugpreauth.pdf

|  |  |
| --- | --- |
| **Input** | CSV with Pipe Separated |
| **Country** | USA |
| **Nrows** | 2 Years (4 million) |
| **Ncols** | 15 |
| **System given** | 16 GB RAM |
| **Types** | Classification (Binary) |
| **Columns** | Carrier ID  Account ID  Employer Group  Plan Code  Drug Code  Gender  Total Drug Package Qty  DOB  Brand Name  Manufacturer  Units of use |
| **Target** | Approve Flag |
| **Cleaning** | Treated missing values in Gender (by mode), Units of use (using domain), Brand Name and Manufacturer (by mode) |
| **Feature engineering** | Age  Formulary Status  [A drug formulary is a list of prescription drugs, both generic and brand name, used by practitioners to identify drugs that offer the greatest overall value. A committee of physicians, nurse practitioners, and pharmacists maintain the formulary.]  Drug Tier  [Drugs on a formulary are typically grouped into tiers. The tier that your medication is in determines your portion of the drug cost   * Tier 1 usually includes generic medications. * Tier 2 usually includes preferred brand name medications. * Tier 3 usually includes non-preferred brand name medications. * Tier 4 usually includes specialty medications (3-Tier programs do not have a unique tier for specialty medications) ]   A medication may be placed in tier 3 or 4 if it is new and not yet proven to be safe or effective; or there is a similar drug on a lower tier of the formulary that may provide you with the same benefit at a lower cost.  Drug Strength  Based on the Drug meta data like Drug Brand name, Drug Generic name, Drug code, Drug benefit type, Drug units, Drug manufacturer and Drug Tier the drugs are grouped into 4 clusters (KMEANS + Domain based clustering).   1. High strength 2. medium 3. low strength 4. very low strength |
| **Important features** | Age  Drug Strength  Gender  Units of Use |
| **Data split** | Train test (70:30) |
| **Models** | * Logistic * Decision tree * ADABoost * Random Forest * XGB |
| **Best model** | XGB |
| **Metric** | Fscore = 0.85 |
| **Parameters** | * 800 trees * Learning rate = 0.01 * Alpha = 2 * Depth = 5 * Colsample = 0.6 * Row sample = 0.6 |
| **Insights** | If Drug Tier falls in Tier - 3/4 then high chance of rejection  If Gender = Female and Drug strength is high then chance of rejection  If Age = Teen/Child and Drug Strength is high then chance of rejection |
| **Impact** | $2 million (one year)  2% Reduction in number of PA |
| **Data time range** | 2014-2015 (Predicted on 2016) |

**PROJECT 2: Vehicle Insurance Fraud Detection Solution**

Insurance fraud detection is a challenging problem, given the variety of fraud patterns and relatively small ratio of known frauds in typical samples. While building detection models, the savings from loss prevention needs to be balanced with cost of false alerts. Machine learning techniques allow for improving predictive accuracy, enabling loss control units to achieve higher coverage with low false positive rates.

Insurance frauds cover the range of improper activities which an individual may commit in order to achieve a favourable outcome from the insurance company. This could range from staging the incident, misrepresenting the situation including the relevant actors and the cause of incident and finally the extent of damage caused.

Potential situations could include:

* Covering-up for a situation that wasn’t covered under insurance (e.g. drunk driving, performing risky acts, illegal activities etc.)
* Misrepresenting the context of the incident: This could include transferring the blame to incidents where the insured party is to blame, failure to take agreed upon safety measures
* Inflating the impact of the incident: Increasing the estimate of loss incurred either through addition of unrelated losses (faking losses) or attributing increased cost to the losses

|  |  |
| --- | --- |
| **Input** | Multiple CSV files   * Vehicle data * Claim data * Policy data * Party data * Repair data   Created a master data by joining over id |
| **Country** | AUS (Client Name: SunCorp) |
| **Nrows** | 2.5 million |
| **Ncols** | 35 |
| **System given** | 8 GB RAM |
| **Types** | Classification (Binary) |
| **Columns** | Vehicle   * id * Name * Manufactured year * bought date * price   Claim   * id * loss date * lodge date * police inform date * loss state * sum insured * drunk flag * loss cause (accident/theft) * Number of Third parties involved   Policy   * id * policy start date * end date * amount   Party   * id * gender * dob * occupation * injury type * licence type * state   Repair   * id * amount * type |
| **Target** | Fraud (Yes or No) |
| **Cleaning** | Removed   * duplicates * blank rows * incorrect ids * dropped missing rows (1% rows) |
| **Feature engineering** | party age  vehicle age  police lodge time (same day/next day/others)  claim lodge time (same day/next day/others)  claim age  days to expiry  policy expired or not  holiday flag (from calendar)  area of loss (high risk/medium risk/low risk)  Grouped vehicles into types |
| **Important features** | Drunk flag  Holiday flag  Licence type  Claim amount (if amount is high - high chance of fraud)  Police lodge time  Area of loss (In high risk areas we have more frauds)  Vehicle type (high level vehicles claim more frauds) |
| **Data split** | Train test (70:30) |
| **Models** | * Logistic * Decision Tree * XGBoost * RUS Boost * SMOTE Boost |
| **Best model** | Weighted Ensemble of RUS + SMOTE + XGB |
| **Metric** | Fscore = 0.93 |
| **Insights** | * Nearly 86% of the claims that are marked as fraud were not reported to police, whereas most of the non-fraudulent claims were reported to the police. * In 82% of the frauds the vehicle age was 6 to 8 years (i.e. old vehicles tend to be involved more in frauds), whereas in the case of non-fraudulent claims most of the vehicle age is less than 4 years. * When an accident happens on holiday week it is 80% more likely to be a fraud |
| **Impact** | Over $3 million |
| **Data time range** | 2014 - 2015 |

**PROJECT 3: Health Care Cost Predictive Modelling**

**Model**: Predicts an individual’s cost to an insurer

**Goal**: Determine future health care costs from prior costs, demographics, and diagnoses

**Aim**:

* Allow more accurate health insurance rate-setting
* Identify individuals for medical management
* Transfer funds between insurers to offset risk on new health insurance exchange post 2014

**Challenge**:

* Data is highly nonlinear: costs range from $0 to $300,000
* Costs fluctuate from year to year
* many variables;

|  |  |
| --- | --- |
| **Input** | CSV file |
| **Country** | USA |
| **Nrows** | 30000 (Per Year, 2 Year data)  (30000 individuals data) |
| **Ncols** | 56 |
| **System given** | 16 GB RAM |
| **Types** | Regression |
| **Columns** | DOB  previous year costs  category wise costs (pharmacy, medical, in patient)  gender  medical condition flags  diabetes  pregnant  hypertension  endocrine  Renal failure (kidney failure)  Number of PCP visits(Primary care physician)  Number of ER visits(emergency room)  State |
| **Target** | Cost |
| **Cleaning** | NONE (as of now) |
| **Feature engineering** | Divided the data into two datasets (one for male and one for female)   * because different features are dependent on gender and we wanted it to be the first split criteria   For now:   * Age from DOB * total costs (sum of all category wise costs) |
| **Important features** | Feature selection done using BORUTA package in R  (Read about BORUTA)   * Age * State * Diabetes * In Women - Pregnancy * Number of PCP visits * previous year costs * total costs |
| **Data split** | 70-30 (in both male and female data sets) |
| **Models** | * Linear regression * Decision Trees * Random Forest * XGBoost |
| **Best model** | * For now: (XGB + Linear regression with regularization)   + **Why**: XGB captured values and regression captured trend * Yet to be fine tuned |
| **Metric** | MAPE: 17% error (For now) |
| **Insights** | * Year 1 cost accounts for 51% of year 2 cost. It is the single most important predictor. * Age 15-44 males are more than 40% less expensive than females in this age group. Women of childbearing age have a much higher expected cost. * Each primary care physician (PCP) visits in year 1 raises expected cost by 10 %. Individuals who visit their PCP often are expected to have a much higher cost. * chronic conditions could indicate high probability of expenses * Diabetes and other metabolic/endocrine disorders are the most important diagnoses to note for insurers. |
| **Impact** | UNKNOWN |
| **Data time range** | 2012-2013 |