Data Analytics Life Cycle – Phase 1: Discovery | Phase 2: Data Preparation | Phase 3: Model Planning | Phase 4: Model Building | Phase 5: Communicate Results | Phase 6: Operationalize

Data Preparation ELT/ETL (Extract Load Transform) vs (Extract Transform Load) - The more typical approach is transforming data before loading into database. However, with an analytic sandbox, ELT is recommended as having the original raw data can provide significant value. Data can still be transformed to a new state or stay in its original form. ie credit cards, outliers are trimmed, fraudulent charges are usually the outlier, ELT works better for this kind of analysis.

Model Planning – Problem : Technique : Method | I want to group items by similarity, find structure/commonalities in data : Clustering : K-means clustering | I want to discover relationships between actions or items : Association Rule : Apriori | I want to determine the relationship between the outcome and the input variables : Regressing : Linear/Logistic Regression | I want to assign (known) labels to objects : Classification : Naïve Bayes/Decision Trees

K-means Clustering – Unsupervised learning method: finding hidden structure within unlabeled data | form homogenous groups with data set based on internal structure | results tight group that are far apart from data points in other groups | two measures of Euclidian distance, between data points, and between clusters | Steps, 1. Choose k and generate k random centroids, 2. Assign records closest to centroid, 3. Recalculate centroid (mean of all records), 4. Repeat step 2 and 3 until assignments no longer change. | K can be picked by finding elbow of within-sum-of-squares, aka how tight average each cluster is. | Pros, easy to assign new data to existing clusters, concise output | Cons, only numerical data not category, sensitive to first guess, variables must be on same scale, wrong k guess = poor results, tends to produce round clusters which is sometimes not desired

Association Rule/Apriori – unsupervised | itemsets = discrete set of items linked together someway | Support = percent of occurrences that contain item | Apriori Property = any subset of frequent itemset is also frequent. | Confidence percent of transactions that contain X that also contain Y “X->Y” | Support count = number of occurrences | support = support count / total count | frequent itemset = greater than threshold | association rule Support of X + Y / Total | confidence X + Y / X | Lower threshold, whole Apriori tree can be pruned. | Lift (X->Y) = Support (X^Y)/(Support(X)\*Support(Y)) How many times more often X and Y occur together than if they were statistically independent. | Leverage (X->Y) = Support(X^Y) – Support(X)\*Support(Y) | Pro – uses clever observation to prune search space, easy to parallelize | Con, requires many database scans, exponential time complexity, can mistakenly fund spurious relationship addressed with lift and leverage. | Hold out data can be used to test if results of above can predict what items are missing. | Use cases: Market basket analysis, recommendation system, web usage patterns.

Linear Regression – Regression focus on relationship between outcome and how each input variable affects outcome | We want to fit linear model to observed data f(x) = w1x + w0 | The goal is to minimize the sum of squared error SSE = Sum((yi – f(xi))^2) = Sum((yi – w1x – w0)^2) and end up with f(x) = mean(y) + sum((xi – mean(x))(yi – mean(y))) / sum((xi – mean(x))^2) \* (x – mean(x)) | y = b0 + b1x1 + b2x2 + ….. | Express categorical variable as binary ie gender 0/1, or state is 49 binary numbers. Zip code is extremely complex. | b0 is reference situation ie female, Alabama, 0 years old, with 0 years experience. | bi is change in y as a xi changes assuming all else equal. | significance is probably bi = 0, bi is significant if P(bi=0) is small. | N-fold cross validation done by splitting data into N sets and analyzing N-1 folds. Compare analysis among all folds to see if analysis was accurate. | goodness of fit metric R^2 = 1 – Sserr/SStot = Sum((y-ypred)^2) / Sum((y-ymean)^2) [We want this as close as 1 as possible]. | Sanity check – Do the sign make sense? Any coefficients excessively large? | Look at prediction vs true outcome and see if symmetric, evenly over/under, outliers shown, and if not rework coefficients removing correlated ones and using correct ranges. | Pro Concise representation with coefficients, robust to redundant/correlated variables, explanatory value, easy to score data | Con Dos not handle missing values well, assume variable affects the outcome linearly and additively, can’t handle outcome in discontinuos way like step function, hard to handle discrete drivers with many distinct values like zip code. | Use cases Home value, customer lifetime value, loss given default on loan.

Logistic Regression – Used to estimate the probability an event will occur as a function of other variables | example default = f(creditScore, income, loanAmt, existingDebt). Standard threshold 0.5 for yes/no result. | ln(P(y=1)/(1-P(y=1)) = b0 + b1x1 + b2x2 + … | Categorical is same as liners | Computing coefficients same as linear | exp(bj) tells us how the odds-ration of y = 1 changes for every change in xj. Bcreditscore = -0.69, exp (-0.69) = 0.5 = ½. For everything same, one point increase leads to halved odds-ration. Negative number indication negative relation. | significance is same as linear. | regression preserves summary statistics because probability mass = counts. | pseudo-R2 = 1 – deviance\null deviance | null deviance is error that would be if probability of true was the global probability. How well model explains data| sanity checks like linear regression and can indicate strongly correlated inputs | to use logistic regression to find classifiesrs , you have to set a threshold. | False Positive Rate fraction of negative instances that were misclassified FPR = # False Positives / All negatives. False Negative Rate fraction of positive instance that were misclassified. True Positive Rate TPR = 1 – FNR = # true positives/all positives. | Received Operative Characteristics (ROC) curve plots FPR,TPR as threshold is varied from 0 to 1. Area under curve tells how well model predicts. Can help set classifier threshold | Pro Explanatory value, robust with redundant/correlated variables, concise representation, easy to score data, return good probability estimates, preserve summary stats of training data | Cons does not handle missing values well, assume variables affects the log-odds linearly and additively, cannot handle outcome in discontinuous way, not good with big categorical values, correlated values make insane sanity checks. | logit is inverted by sigmoid function and is logitP(y=1). | Binary classification = probability of event ie borrower will default, customer will churn.

Naïve Bayesian Classifiers