HEALTHCARE
ANALYTICS
USE CASE

Predicting Appointment Cancellations with Accuracy

Analyzing healthcare data with Analance can help your organization plan its staffing and day-to-day operations efficiently. Analance gives you insights on who is mostly likely to cancel an appointment and when cancellations are most likely to occur – all displayed on easy to use dashboards and reports. Explore the insights Analance found for a large chain of pain management clinics, which helped predict appointment cancellations, streamline scheduling and save money.

BUSINESS CHALLENGE

Unpredictable patient appointment cancellations led to a large decrease in profits loss of practitioner time, and scheduling ineffectiveness.

SOLUTIONING PROCESS AT A GLANCE

The process of statistical consulting and solutioning starts with a thorough understanding of the business challenge, its impact, and the data available for analysis. With this information, we arrive at a solution to mitigate or control the challenge, offer continued client support, and adjust models over time.
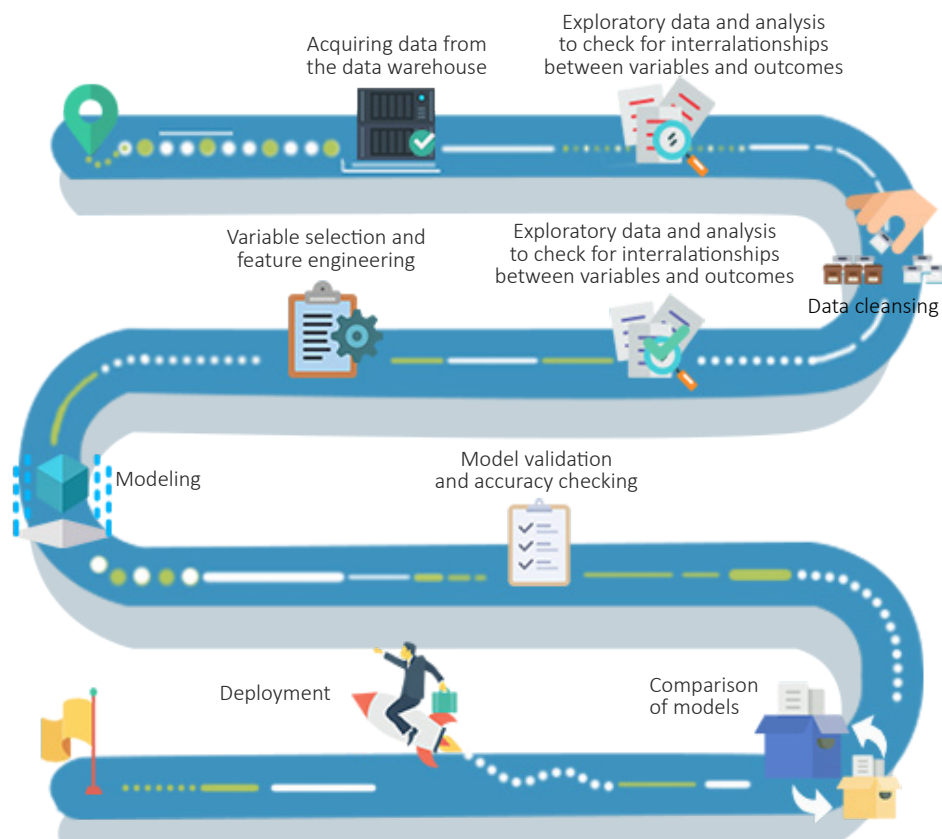
OUR PROCESS

A data dump was acquired and put through a stringent exploratory process before trying to correlate what information was available to solve the challenge at hand.

A record's outcome was classified as "Cancelled" if it matched one of the following criteria:

- Portal Cancellation by patient
- Cancelled by the patient with no reason
- Cancelled by the patient for financial reasons
- Cancelled in less than 24 hours of the appointment start time

- Cancelled due to No OON (Out Of Network) referral available
- Cancelled due to other pending referral
- Cancelled due to illness

PREDICTIVE MODELING STAGES

Acquiring data from the data warehouse

Exploratory data and analysis to check for interralationships between variables and outcomes

Variable selection and feature engineering

Exploratory data and analysis to check for interralationships between variables and outcomes

Data cleansing

Modeling

Model validation and accuracy checking

Deployment

Comparison of models

Some variables used as predictors were:

- **City** – City patient comes from
- **State** – State patient comes from
- **County** – County patient comes from
- **MaritalStatus** – Marital Status of the patient
- **EmploymentStatus** – Employment status of the patient
- **ReferSource** – Source of referral of patient
- **DefaultDoctorId** – ID number of the default provider for the patient

- **PrimaryCareDoctorId** – ID number of the primary care physician for the patient
- **RefDoctorId** – ID number of the referring doctor
- **PrimaryInsuranceCarrierId** – ID number of the primary insurance carrier
- **SecondaryInsuranceCarrierId** – ID number of the secondary insurance carrier
- **TertiaryInsuranceCarrierId** – ID number of the tertiary insurance carrier

THE MODELING PROCESS

A Mathematical/Statistical/Econometric predictive model is a method of representing a variable of interest (outcome) as a function of other variables (predictors) with an assumed relationship between the outcome and predictors.

All variables available were studied to understand distributions. Data was cleaned by the means of handling outlying values, missing values, and looking for inter-relationships between predictors before looking to see if any data had a significant relationship with the outcome.

Outlying values were removed from the data after which missing values were replaced with column means (average value) for numeric data and column modes (most frequent label) for categorical or ordered data. The industry standard Box Plot was used to check for outliers and a tabulation or summary of data was used to check for missing values. If two or more predictors were inter-related (for instance height in cm, height in inches, height in feet), only one was used.

The industry standard metric VIF (Variance Inflation Factor) was used as a deciding factor. Predictors with VIF values above 6, indicating a significant inter-relationship with another predictor/s, were removed from the analysis.

The industry standard statistical test used to check for relationships between predictors and the outcome of interest is the Chi-Squared test of association. A p-value less than .05 indicates that the probability of association occurring in the population is less than 5% and is of no significance. Using this test for all predictor-outcome combinations helped in restricting the analysis for only those predictors that majorly influenced readmissions.

A further exhaustive exploratory analysis of data using graphs and advanced visualizations found that:

- Minimum order quantity is significantly related to selling price.
- Lot multiplier is significantly related to selling price.
- Supplier variety is significantly related to selling price.
- Product categorization is significantly related to selling price.
- Stock on hand is significantly related to selling price.
- Back order quantity is significantly related to selling price.
- Buying price is the strongest predictor of selling price.

MODEL USED

The modeling process began with using the industry preferred choice (namely, Binomial Logistic Regression), designed specifically for modeling binary outcomes (i.e. Yes / No). Other models, such as Decision Trees, Neural Networks, and Support Vector Machines, were also considered and results from them were analyzed to find the top performing model. A total of 81 different models were built to find the winning model in terms of findings from the confusion matrix. A confusion matrix tabulates the predicted values from the model with the actual values in the data to see how close we are at capturing the true relationship between the outcome and predictors by means of using derived metrics such as the following:

1. **Accuracy**: The model's ability to predict correct values and incorrect values correctly.

2. **Kappa value:** A comparison of observed accuracy with expected accuracy.

3. **Sensitivity:**The sensitivity of a test (also called the true positive rate) is defined as the proportion of cancellations that have been correctly classified as cancellations.

4. **Specificity:**The specificity of a test (also called true negative rate) is defined as the proportion of non-cancellations that have been correctly classified as non-cancellations.

5. **Positive predictive value:** The probability that those values predicted as cancellations are truly cancellations.

6. **Negative predictive value:** The probability that those values predicted as non-cancellations are truly non-cancellations

CONCLUSION AND RECOMMENDATIONS

The ability to capture patterns in the cancellations and learn which appointments are likely to be cancelled ahead of the scheduled appointment helped the clinic chain to solve their business challenge and transform themselves into a proactive and agile organization. By profiling appointments to find patterns in data that led to a high probability of cancellation, the clinic was able to add many value added services, including reminder services to patients to curb the risk of cancellation, and the ability to hold "back up" patients for those appointments at a risk of being cancelled.

ABOUT DUCEN

Ducen IT helps Business and IT users of Fortune 1000 companies with advanced analytics, business intelligence and data management through its unique end-to-end data science platform called Analance. Analance is an enterprise-class, state of the art integrated platform that delivers power and ease of use to business users and data scientists with a seamless experience and platform scalability to support business growth and strategy.

DUCEN
Intelligence to Decision Making