THE ONLY PREDICTIVE ANALYTICS GUIDE YOU NEED

Table of Contents

Introduction

Every day more than 2.5 quintillion bytes of data is created!

Somewhere in this data lie patterns. Uncover them and you stand at predicting future outcomes.

This is exactly what the branch of data science, predictive analytics focuses on. With a combination of statistical algorithms, machine learning techniques, and historical data, predictive analytics empowers people to better plan their future.
Predictive analytics sees application in every business vertical imaginable—from sales and marketing to healthcare, retail, finance, media and entertainment, travel, and even law enforcement and defense.

Thriving in the Era of Advanced Analytics

Traditional analytics tools have been serving businesses well for many years. By explaining why certain events in the market were taking place, businesses could learn from their past behaviors. But as the ability to collect and store data improved, so did the interest in using that data to look forward, rather than backward.

The rise of big data meant that more powerful algorithms were created to interrogate databases. This gave us the uncanny ability to understand the future, and thrive.

Why is Predictive Analytics Important?

Predictive analytics has been around for years. But now more businesses are turning to it to gain a competitive advantage. But why now?

- The proliferation of the internet and mobile phones has resulted in increased amounts and types of data, that can produce valuable insights
- Computers have now become faster and better
- The cost of collecting data has drastically reduced
- The competition in most business verticals has increased

All of this has made predictive analytics more relevant than before. It is now shaping the world, with application in every vertical imaginable.

Various Predictive Model and their Applications

Naïve Bayes Classifier

Naive Bayes Classifier is a machine learning algorithm. It takes advantage of Bayes Theorem and probability theory to predict the category of a sample. It sees application in many different areas, some of which are:

Facial Recognition

Email Spam Detection

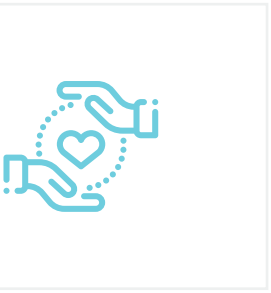
News Categorization

Weather Prediction

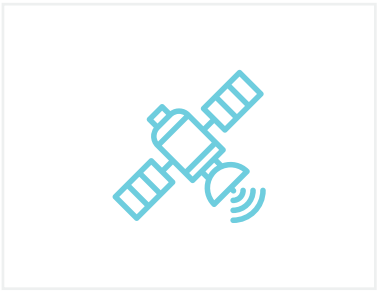
Medical Diagonsis

Decision Trees

Decision tree builds classification or regression models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. Some of the real-world applications of decision trees are:



Healthcare Management



Remote Sensing

Energy Consumption

Fraudulent Financial Statements Detection

Fault Diagnosis in Machinery

Logistic Regression

Logistic Regression is a classification algorithm. It is used to predict a binary outcome (1 / 0, Yes / No, True / False) given a set of independent variables. To represent binary / categorical outcome, we use dummy variables. It is a special case of linear regression when the outcome variable is categorical, where we are using log of odds as dependent variable. It predicts the probability of occurrence of an event by fitting data to a logit function.

Logistic Regression sees application in:

Handwriting Recognition

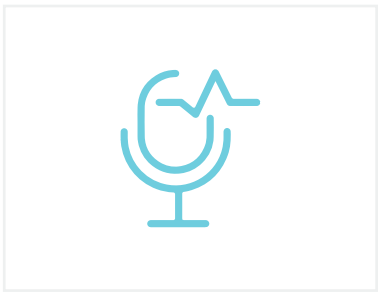
Budget Restructuring

Software Cost Prediction

Credit Scoring

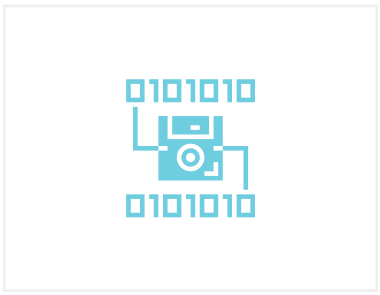
Artificial Neural Networks

Artificial Neural Networks are commonly applied machine learning algorithms that are based on biological neural networks. The goal of these networks is to solve problems the same way as a human brain would. ANNs see applications in:



Speech Recognition

Medical Diagnosis



Machine Translation

Image Processing

Forecasting

K-Means Clustering

K-means clustering is a type of unsupervised learning, which is used when a user has unlabeled data (i.e., data without defined categories or groups). The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable K. The algorithm works iteratively to assign each data point to one of K groups based on the features that are provided. Data points are clustered based on feature similarity.

This model sees application in:

Market Segmentation

Computer Vision

Geostatistics

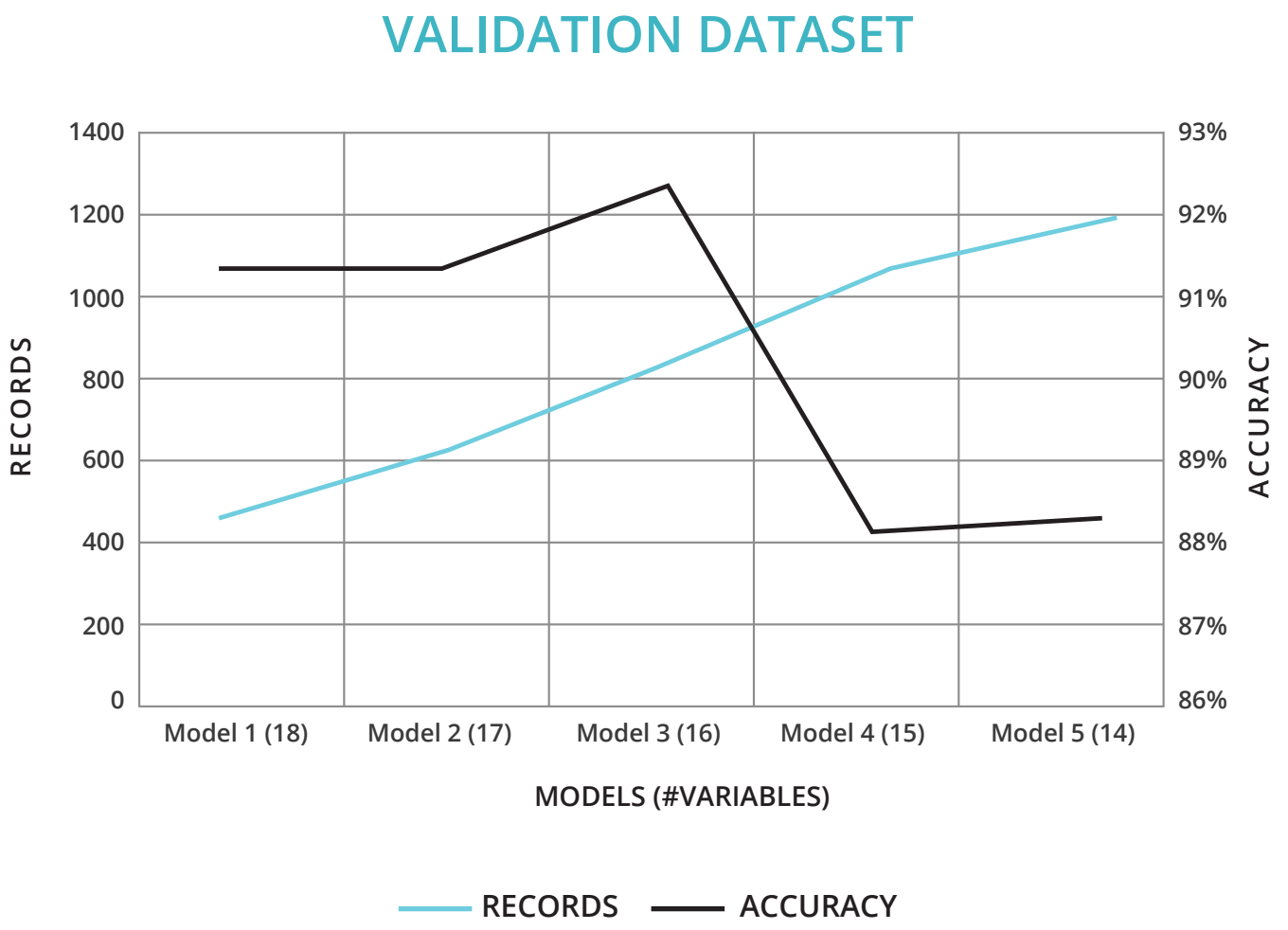
Astronomy

Agriculture

Model Optimization

To run a predictive model, you need a large chunk of data. So, we need to select a dataset that presents a fine balance between the number of records and variables.

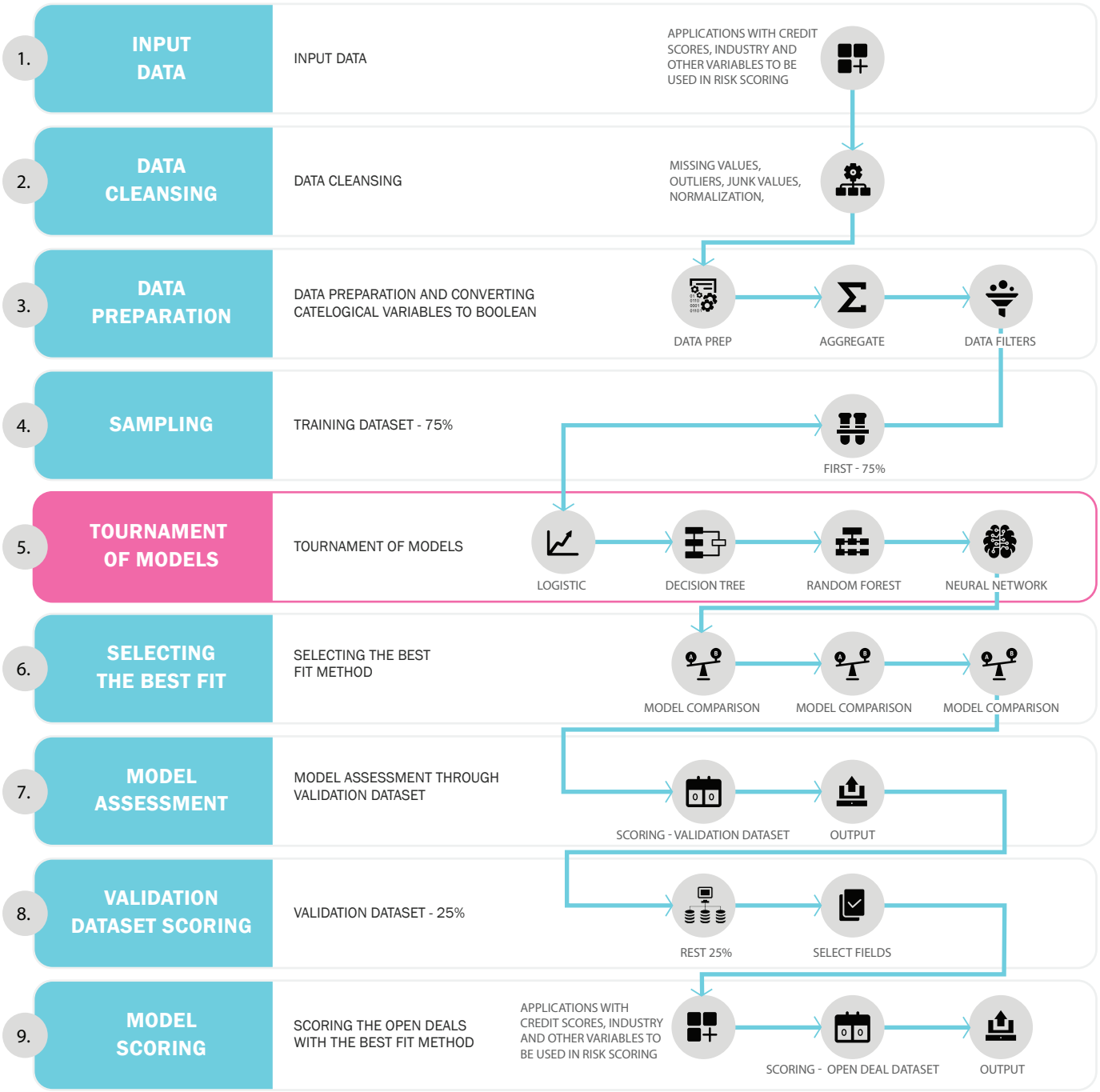
The figure below shows a comparison of the number of records and accuracy for five different datasets. With each dataset, we have reduced the number of variables, which increased the number of records contained in the dataset.



As you can see, when compared to the number of records, there isn't much difference in the accuracy. So we can select the dataset with the highest number of records to be the primary model.

Tournament of Models

In order to select the best fit model, it is necessary to run a tournament of models. As the name suggests, a tournament of models involves creating various models like logistic regression, decision tree, random forest etc. and running them on the same dataset.



This enables us to find the best model for our current dataset. We can then compare the results for all the different models to find the best fit model.

Model Comparison Report

Fit and Error Measures

Model	Accuracy	F1	AUC	Accuracy_0	Accuracy_1
Logit_model	0.9452	0.9650	0.9347	0.9776	0.8354
Decision Tree	0.9251	0.9532	0.9480	0.9431	0.8485
Gradient_Boosting	0.9366	0.9613	0.9882	0.9317	0.8485
Forest_model	0.9798	0.9874	0.9997	0.9752	1.0000

Model: model name is the current comparison.

Accuracy: overall accuracy; number of correct predictions of all classes divided by total sample number.

Accuracy_[class name]: accuracy of class [class name], number of samples that are correctly predicted to be a class [class name] divided by number of samples predicted to be a class [class name]

AUC: are under the ROC curve, only available for two-class classification.

F1: F1 score, precision * recall / (precision + recall)

The Challenge Associated with Predictive Analytics

Predictive analytics aims at making the world a better place, with its application in healthcare, defense and law enforcement, urban transport, waste management etc. Though there is one major challenge associated with effectively implementing predictive analytics.

Data Quality and Volume

Data is the fuel that powers predictive analytics, and the predictive engine needs a lot of fuel. While 90% of the data in the world has been created in the last few years, storing all this data is not a challenge in itself.

About 80-90% of this data is unstructured. This is the data which comes from emails, video feeds, images etc. Structured data, on the other hand, is data that is neatly organized in databases, like financial and banking data.

Predictive analytics relies on historical data that should meet exceptionally high standards and should be comprehensive and diverse to be effective. Despite having access to a large chunk of data, only 3% of companies' data meets basic quality standard.

This creates a peculiar problem where data scientists have to spend up to 80% of their time cleansing data before training the predictive model. Even then, not all the data can be cleansed, which might lead to distorted results.

This is where the challenge lies.

Case Study

The Customer

The customer is a SaaS marketing platform company that empowers marketers to create lasting relationships and grow revenue. A trusted platform for thousands of CMOs and marketers, the client helps master digital marketing to engage customers and prospects.

The Objective

The customer wanted a predictive model to predict the closed won probability for their open renewal opportunities.

Predictive Analysis Process

- Accounts
- Opportunities
- Users
- Activities

- Data Cleansing
- Data Structuring
- Data Pre-processing
- Variable Creation

- Association (correlation) Analysis
- Determined Factors/Effects
- Account Scoring based on Sample Dataset
- Applied Scores to Test Dataset

- Reports of Findings

Modeling Techniques Used

Statistical Techniques

- Association Analysis
- Logistic Regression

Machine Learning Techniques

- Decision Tree
- Boosted Model
- Forest Model

Factors Considered for the Model

- Opportunity Characteristics
- Account Characteristics
- User Characteristics

Model Optimization

5 different data bases were used to perform the modelling

Dimensions	Model 1	Model 2	Model 3	Model 4	Model 5
Total Records	3393	4586	5952	7090	7839
#Variables	18	17	16	15	14
Training Set Opp	1086	1478	1925	2538	2799
Validation Set Opp	466	634	825	1087	1200
Open Opp	1841	2474	3202	3465	3840

Observations:

- All models have a balanced accuracy across the target variable factors
- The models provide us with 90% prediction accuracy enabling us to predict the final outcome for renewal opportunities
- By taking in the balance between the # records and # variables used to train a model, we see there is not much difference between the accuracy for each model when compared to the increase of records. So, we can take the model with highest number of record to be our primary model.

Results

Model Technique Comparison

Model Comparison Report					
Fit and Error Measures					
Model	Accuracy	F1	AUC	Accuracy_0	Accuracy_1
Logit_model	0.9452	0.9650	0.9347	0.9776	0.8354
Decision Tree	0.9251	0.9532	0.9480	0.9431	0.8485
Gradient_Boosting	0.9366	0.9613	0.9882	0.9317	0.8485
Forest_model	0.9798	0.9874	0.9997	0.9752	1.0000

Model: model name is the current comparison.

Accuracy: overall accuracy; number of correct predictions of all classes divided by total sample number.

Accuracy_[class name]: accuracy of class [class name], number of samples that are correctly predicted to be a class [class name] divided by number of samples predicted to be a class [class name]

AUC: are under the ROC curve, only available for two-class classification.

F1: F1 score, precision * recall / (precision + recall)

Note: We used AUC as a criteria to select the best fit model
* These visuals are for demonstration purpose only - no real data

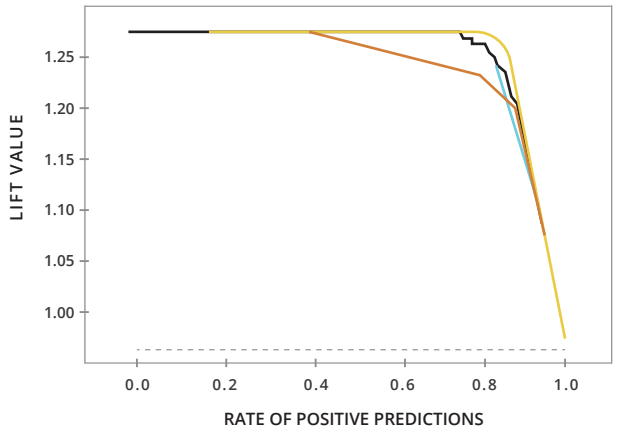
.90 - 1.00	=	Excellent (A)
.80 - .90	=	Good (B)
.70 - .80	=	Fair (C)
.60 - .70	=	Poor (D)
.50 - .60	=	Fail (F)

The model comparison report presents information—accuracy, precision—that enable us to find the best technique.

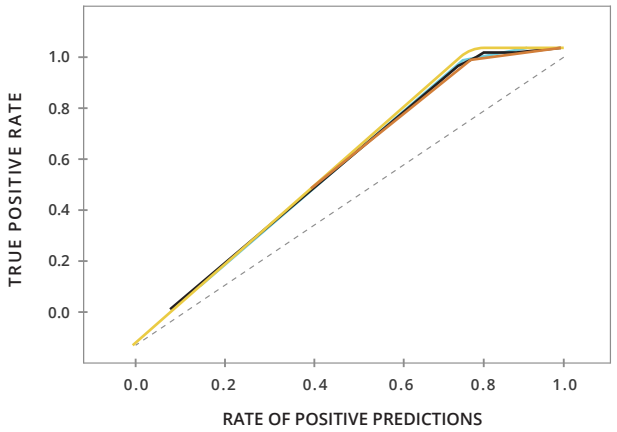
Results

Performance Diagnostic Plots

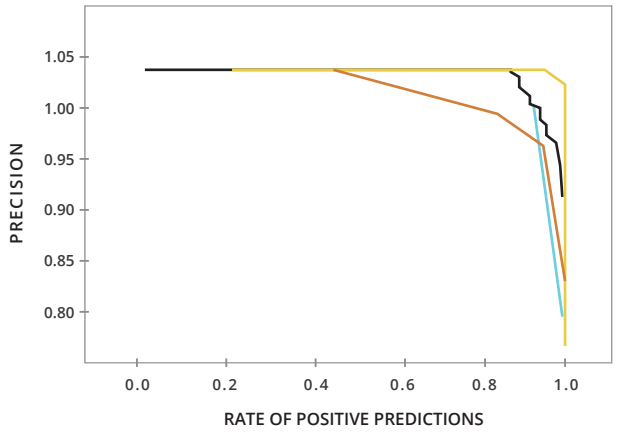
LIFT CURVE



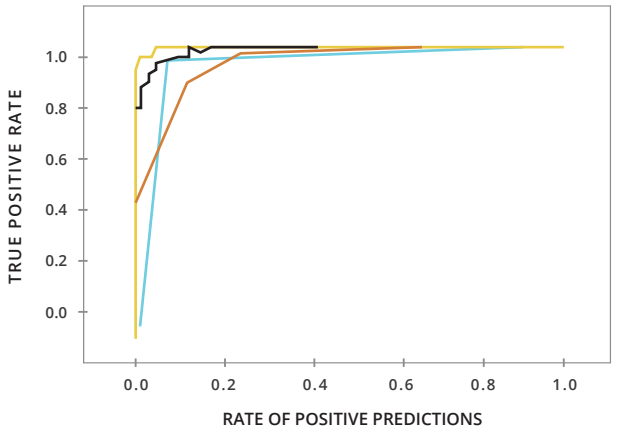
GAIN CHART



PRECISION AND RECALL CURVE



ROC CURVE



DECISION TREE
LOGIT_MODEL
FOREST_MODEL
GRADIENT_BOOSTING

Next, to evaluate each model, we used performance diagnostic plots—Lift Curve, Gain Chart, Precision-Recall Curve, and ROC Curve.

About Grazitti Interactive

Grazitti Interactive is a global digital services provider leveraging cloud, mobile and social media technologies to reinvent the way you do business. Since 2008, Grazitti has been helping companies power their business with its data analytics and business intelligence service.

As a global consultancy, we have strategic partnerships with technology pioneers like Alteryx, Marketo, Salesforce.com, Adobe, Optimizely and Jive. We combine these new platforms with our innovative approaches to provide effective solutions to our clients. Doing this has allowed us to help hundreds of companies to transform their business and save millions.

For more info about marketing attribution, drop us an email at info@grazitti.com.