THE LYTX™ ADVANTAGE: USING PREDICTIVE ANALYTICS TO DELIVER DRIVER SAFETY

BRYON COOK
VP, ANALYTICS

www.lytx.com
## CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction</td>
<td>3</td>
</tr>
<tr>
<td>Predictive Analytics Overview</td>
<td>3</td>
</tr>
<tr>
<td>Drivers in Use of Predictive Analytics</td>
<td>4</td>
</tr>
<tr>
<td>Technological Advances</td>
<td>4</td>
</tr>
<tr>
<td>Data Availability</td>
<td>4</td>
</tr>
<tr>
<td>Why Predictive Analytics Projects Fail</td>
<td>6</td>
</tr>
<tr>
<td>Why Predictive Analytics Succeeds at Lytx</td>
<td>6</td>
</tr>
<tr>
<td>Lytx Safety Score™</td>
<td>6</td>
</tr>
<tr>
<td>Target Variable – What are We Solving?</td>
<td>6</td>
</tr>
<tr>
<td>What is The Lytx Safety Score?</td>
<td>6</td>
</tr>
<tr>
<td>Definition and Usage</td>
<td>6</td>
</tr>
<tr>
<td>What Drives the Score</td>
<td>7</td>
</tr>
<tr>
<td>Items not in a Safety score</td>
<td>7</td>
</tr>
<tr>
<td>Analytical Techniques</td>
<td>8</td>
</tr>
<tr>
<td>Measuring Lytx Safety Score Efficacy</td>
<td>9</td>
</tr>
<tr>
<td>Predictive Analytic Modeling Process</td>
<td>12</td>
</tr>
<tr>
<td>Sense: Data Capture</td>
<td>12</td>
</tr>
<tr>
<td>Predict: Model Development</td>
<td>12</td>
</tr>
<tr>
<td>Process</td>
<td>12</td>
</tr>
<tr>
<td>Prevent: Changing Behavior Using Analytics</td>
<td>13</td>
</tr>
<tr>
<td>CONCLUSION</td>
<td>15</td>
</tr>
<tr>
<td>About Lytx</td>
<td>15</td>
</tr>
<tr>
<td>APPENDIX</td>
<td>16</td>
</tr>
<tr>
<td>Predict: Model Development</td>
<td>16</td>
</tr>
<tr>
<td>Analytical Techniques</td>
<td>18</td>
</tr>
<tr>
<td>Advanced Modeling Techniques</td>
<td>20</td>
</tr>
</tbody>
</table>
INTRODUCTION

We are in the midst of a public safety crisis with worldwide implications. According to the Centers for Disease Control, fatal vehicle collisions are one of the leading causes of death in the United States today with one person dying in a collision every 15 minutes. Even more striking is the fact that more than 90 percent of these collisions are preventable. In addition, traffic collisions needlessly cost two percent of the U.S. gross domestic product every year.

At Lytx™ we harness the power of data to change human behavior and make good companies become even better. The Company’s flagship product, DriveCam powered by Lytx™ identifies and addresses the causes of risky driving behavior by capturing data from multiple technologies and sensors and combining it with the Lytx Engine™ technology and Lytx Insights™ program to help clients be safer on the road every day. The DriveCam® Program is dedicated to helping professional drivers perform every day to the best of their ability to help make sure they return safely to their homes each night.

Lytx uses predictive analytics to help prevent future incidents and save lives. Though long used by the Fortune 1000 and other large organizations, predictive analytics historically has been both difficult and expensive to employ due to the need for highly trained analytical staff, massive hardware and processors, and inaccessible data. In the past few years, predictive analytics has gone from an exotic technique practiced in just a few niches, to a competitive asset with a rapidly expanding range of uses. The increasing adoption of predictive analytics is fueled by converging trends: the Big Data phenomenon, ever-improving tools for data analysis, and a steady stream of demonstrated successes in new applications.

Every day, 2.5 quintillion bytes of data (equivalent to 57.5 Billion 32GB IPad) is created. So much that 90% of the data in the world today has been created in the last two years alone. The data comes from everywhere: sensors used to gather climate information, posts to social media sites, digital pictures and videos, purchase transaction records, and cell phone GPS signals to name a few.

The way predictive models produce value is simple in concept; they make it possible to make more educated decisions, more quickly, and with less expense. They can provide support for human decisions, making them more efficient and effective, or in some cases, they can be used to automate an entire decision-making process.

As in-cab and mobile technology has become more prevalent, it is now possible to develop sophisticated predictive models to assess driving risk. This document discusses the Lytx foundations of predictive analytics and its use in the context of identifying risky driving, the drivers of its growth, and some of the technical aspects of determining a viable predictive measure of behaviors that can lead to serious collisions if not addressed.

PREDICTIVE ANALYTICS OVERVIEW

Predictive analytics encompasses a variety of techniques from statistics, modeling, machine learning, and data mining that analyze current and historical facts to make predictions about future, or otherwise unknown, events. The form of these predictive models varies, depending on the behavior or event that they are predicting. Most predictive models generate a score (a credit score, for example), with a higher score indicating a higher likelihood of the given behavior or event occurring.

In business, predictive models exploit patterns found in historical, transactional, and predictive data to identify risks and opportunities. Predictive models seek to explain the relationships between a critical variable (e.g. Collision) and a group of factors (e.g., Driving Patterns and Behaviors) that help predict its outcome.
Most business processes have the potential to benefit from predictive modeling. That said, there are certain situations where predictive models can be especially beneficial in delivering a great deal of value:

- Processes that require a large number of similar decisions
- Where the outcomes have a significant impact, i.e., where there’s a lot at stake in terms of money or lives
- Where there’s abundant information in electronic data form available on which to base decisions and measure outcomes
- Where it’s possible to insert a model calculation into the actual business process, either to automate decisions or to support human decision makers

In the case of risky driving analysis, all of the conditions above are met, which make it an ideal subject for which to develop predictive models.

**DRIVERS IN USE OF PREDICTIVE ANALYTICS**

Although Lytx has leveraged predictive analytics techniques for many years, several key drivers have enhanced our ability to perform complex predictive modeling.

**TECHNOLOGICAL ADVANCES**

Big data is a collection of data sets that are so large and complex that they become awkward to work with using traditional database management tools. The volume, variety, velocity, and heavy computational requirements of big data have introduced challenges across the board to capture, store, search, share, analyze, and visualize data. Thanks to technological advances in computer hardware including faster CPUs, cheaper memory, and Massive Parallel Processing (MPP) architectures; and new software technologies such as Hadoop, MapReduce, and in-database and text analytics for processing big data, it is now feasible to collect, analyze, and mine massive amounts of structured and unstructured data for new insights.

Advances in computer hardware and software design have also yielded software packages that quickly perform such calculations, allowing Lytx to efficiently analyze the data they produce and validate their predictive models.

**DATA AVAILABILITY**

The validity of any predictive model depends on the quality and quantity of data available to develop it. Over the past decade, standard vehicle sensors have grown to more than 100 sensors per vehicle providing more in-vehicle and driving information than ever before. In conjunction with video and observational analysis, Lytx is now capable of delivering very accurate predictive models to determine how drivers drive and the risk that is associated with their driving patterns through what is called the Lytx Engine.

The Lytx Engine applies advanced predictive models to prioritize, select and route data to review centers where teams of expert reviewers identify and verify behaviors from video events, adding structured labels to the data. These labels provide the basis for the scoring algorithms. They also allow for the continuous improvement of the predictive models that power real-time decision algorithms in the Lytx sensors and servers. The scoring algorithms and statistical models enable the creation of driver safety and coaching effectiveness models, both of which are examples of the tools Lytx uses that predict the likelihood of future collisions. These important predictors help safety managers and coaches understand and prioritize their areas of focus. The constantly growing database enables Lytx developers to refine and improve our ability to predict risky behaviors so that even more value can be delivered to clients. In short, Lytx harnesses the vast amounts of sensory data and distills it into useful, actionable insights for our clients.
There are Six Big V’s that drive the Company’s ability to derive insights in driver risk:

**Three Big V’s**

**Volume:** Big data implies enormous volumes of data (terabytes, petabytes and more). When people were the only source of data, it stayed relatively small. More data aids in predictive modeling by providing higher accuracy and confidence in the estimations.

**Variety:** Variety refers to the many sources and types of data; both structured and unstructured. Variety allows Lytx to test more attributes and combinations of attributes to deliver more accurate insights.

**Velocity:** Data is streaming in at unprecedented speed and must be dealt with in a timely manner. All types are now being captured (structured, semi-structured, and unstructured). Through real-time data and self-learning algorithms, businesses can take advantage of real-time decision making to optimize their business.

**Often Used 4th V**

**Veracity:** Big Data Veracity refers to the truthfulness in context of the data. It defines the biases, noise, uncertainty, imprecision, or abnormality in data. In order to build accurate predictive models, it is imperative that the data being mined is understood (how clean are each of the data sets and which data can be trusted) and meaningful to the problem being analyzed. It is critical that the team and partners work to help keep the data clean and implement processes to keep ‘dirty data’ from accumulating in the systems. For in-vehicle technologies, video can provide an invaluable resource to validate the data and ensure the understanding and veracity of the data being captured.

**Upcoming Big Data V’s**

**Validity:** Validity describes the correctness and accuracy of the data for the intended use and the method by which the data is captured (e.g. are you using the correct sensors to capture the data you need and are they calibrated correctly?). Clearly, valid data is key to making the right decisions.

**Volatility:** Big data volatility refers to how long data is valid and how long should it be stored. In today’s world of real-time data you need to determine at what point is data no longer relevant to the current analysis.
WHY PREDICTIVE ANALYTICS PROJECTS FAIL

Unfortunately, even while all of the ‘V’s’ of Big Data exist, most Big Data projects have tended to fail to date. There are five primary reasons that Big Data projects fail:

1. Focus on technology rather than business opportunities.
2. Inability to provide data access to subject matter experts.
3. Failure to achieve alignment and full enterprise adoption.
4. Lack of knowledge and sophistication to understand that the project’s total cost of ownership includes people as well as information technology systems.
5. Lack of access to necessary data.

Many of the big data projects and proof-of-concepts now underway are more about testing technology than uncovering business value. In order to develop accurate and valuable predictive models, it is important to understand the business problem first to ensure the models are solving a real-world problem.

WHY PREDICTIVE ANALYTICS SUCCEEDS AT LYTX

While many modeling projects fail, Lytx continues to show success in its models by leveraging predictive modeling to make practical use of real data to solve real problems. Real data and real problems are the key. Lytx works very closely with its customers, academic institutions, and strategic partners to get a keen understanding of the space we operate in and the problems that exist within that space. Once those problems are identified, the company utilizes its analytic expertise to capture and process the data to solve those specific real-world problems and leverage the Lytx Insights™ program to ensure that real-change is occurring.

LYTX SAFETY SCORE

TARGET VARIABLE – WHAT ARE WE SOLVING?

The Lytx Safety Score™ is a predictive model to assess a driver’s probability of being involved in a future collision.

WHAT IS THE LYTX SAFETY SCORE?

DEFINITION AND USAGE

The Lytx Safety Score represents a driver’s probability of being involved in a future collision based on their past DriveCam events. (A risky driving behavior recorded by sensors is classified as an event.) Within the DriveCam Online® web interface (DOL), scores are ranked within three tiers: RED / YELLOW / GREEN

- **RED** = Low “Safety” Score
- **YELLOW** = Medium “Safety” Score
- **GREEN** = High “Safety” Score

**RED** drivers have a higher probability of being involved in a collision (and are more likely the riskiest drivers) based on DriveCam event history. This probability, however, is not an indication of whether or not a future collision is likely to be the fault of the driver.
The tiered Lytx Safety Score Ranking ‘widget’ can be seen to the right (this is on the DOL3.0 coach home page). Each driver receives a numerical value (for which a ‘color’ is assigned) based on how close the driver’s profile matches the profile of a collision driver.

**WHAT DRIVES THE SCORE**

Due to the complexity of the statistical analysis used in a safety score determination, and the fact that our proprietary scoring algorithms constitute trade secrets and are not made publicly available, you cannot precisely figure your own Safety Score. However, some of the general criteria used in calculating the Lytx Safety Score are described here to provide a general idea of what can positively or negatively impact the score.

The Lytx Safety Score evaluates a driver’s information against various attributes of drivers who have experienced one or more collisions. Through segmenting collision drivers (drivers who have had at least one collision during a given time period) and non-collision drivers (drivers who have not had a collision during the given time period), Lytx has been able to isolate the attributes that are more highly predictive of a future collision.

Some of the information used to come up with the Lytx Safety Score includes:

- **General Driving Criteria**
  - General Driving Aggressiveness
  - Risky ‘Event’ Frequency
  - Behavioral Severity and Behavioral Combinations
  - Posted Speed Violations
  - Distance Traveled

- **Contextual Adjustments**
  - Duration on the Program
  - Vehicle Type Adjustments
  - Geospatial Adjustments
  - Driving Adjustments for Environmental Conditions
  - Behavioral Change After Intervention

Using all of this information and more, the predictive models allow Lytx to segment drivers into categories (Red, Yellow, Green) of risk which predict the probability that a driver will be in a collision (and the severity of the collision) in the future. Based on our model, a driver in the green category is the less likely to be in a collision in the near future, and a driver in the red category is more likely to be in a collision. By making these probabilities easy to assess, the Lytx Safety Score can help fleet managers make better decisions on how to prioritize risk within their fleets.

**ITEMS NOT IN A SAFETY SCORE**

The Lytx Score is an unbiased view of fleet behavior. Factors such as age, sex, income, and length of employment are not currently factored into the safety score, and are not considered in the calculation of the score. For most companies, the Lytx Safety Score is only one aspect, albeit an important one, of the overall safety assessment of a driver.
In order to compare ‘collision’ and ‘non-collision’ drivers, Lytx leverages several basic and advanced data mining, statistical, and predictive modeling techniques (see ‘Predictive Analytic Modeling Process: Predict Model Development’ and ‘Appendix ‘within this white paper for sample techniques).

The Company has made strides in each of the six ‘V’ categories over the past 16 years resulting in an improvement of the efficacy of its predictive models:

**Volume:** As of January 1, 2014, Lytx had observed over 27 billion miles of driving. That is equivalent to driving around the circumference of the Earth nearly 1.1 million times or equivalent to nearly 57,000 round trips to the moon. With the current customer base, approximately 10 billion miles driven annually are being added to the Lytx dataset (the circumference of the Earth is rounded nearly 400,000 times a year or nearly 21,000 lunar round trips per annum). Hundreds of millions of data points are being analyzed daily and this number continues to grow with each new client brought on to our service.

**Variety:** Currently, dozens of sensors are being leveraged within the DriveCam Program to ensure a variety of information that allows analysts to cross-correlate attributes for the highest possible efficacy in models.

**Velocity:** Lytx receives, at a minimum, daily feedback from multiple sensors. Many of provide near-real time information. This frequency within the datasets allows changes to be quickly detected in driving patterns, allowing the predictive models to adjust and enabling Lytx to notify customers of anomalies of concern.

**Veracity:** One of the extraordinary advantages Lytx has in its analytics is the ability to have a continuous flow of human reviewed observational behavioral analysis. Beyond the obvious advantage of being able to isolate event triggers, perform root cause analysis, and receive video for exonerations, the continuous feedback that video analysis provides allows greatly enhanced analytics. It provides a non-replicable answer key that allows continuous learning every day and enables the application of that knowledge to future video, as well as the extrapolation to non-video data. These human reviewed and validated datasets get fed into the Lytx models daily for continuous learning and improvement in the predictive models.

**Validity:** As part of the data mining efforts, Lytx takes great pride in our ability to analyze data efficiently and to ensure the cleanliness, completeness, and validity of the datasets. As an analytics organization, the Company aims to provide the highest quality models, and considerable effort is made to help ensure effective and quality procedures are in place.

**Volatility:** Lytx has been in business for 16 years and has been collecting data from its sensors for the last 9 years. This long history has provided rich data sets that allow the Company to factor in seasonality of the data, while also providing understanding about the data time frames that each of the models require to be effective.

This myriad of high quality data gives Lytx the ability to rapidly predict changes in driving that may indicate a higher probability of a collision by leveraging transaction-based analytics to detect changes in risk as they occur, and generate fresh scores daily. Transaction scoring is the key to making accurate real-time risk decisions.

This chart is an oversimplified graphic to describe how the safety score model evaluates transactional information and is for illustrative purposes only:
Data attributes such as these are inputs into the predictive modeling algorithms that allow the program to distinguish safer versus riskier drivers.

**MEASURING LYTX SAFETY SCORE EFFICACY**

The most important stage of predictive modeling is validating the accuracy of Lytx’s models and determining the actual efficacy in their predictions. Each month, Lytx data shows about 0.97% of the driver population utilizing the DriveCam Program will be involved in some type of collision. Based on this data, the Company could take the easy route and simply classify all drivers as non-collision drivers and achieve an accuracy rate of over 99% each month. However, this approach would deliver little value to clients. Therefore, Lytx holds itself to a higher standard.

Lytx’s measurement criteria isolates a specific target variable (in this case, a collision event), and employs a much stricter standard of measurement. The primary objectives in developing the Lytx Safety Score are to identify collision drivers PRIOR to their collisions to provide customers an opportunity to modify those drivers’ behaviors and prevent the collisions from occurring. This is balanced with trying to minimize a customer’s operational burden since they have limited time to coach drivers and need to focus on the highest priority drivers each week.

Based on those objectives, the Safety Score measurement focuses on the 0.97% of the drivers that get in a collision each month and we measure success as follows:

> What percentage of the drivers that were involved in a collision were predicted accurately in the 90 day period prior to a collision?

Specifically, was the collision driver in the ‘RED’ at least one-third of the time in the 90-day period prior to their collision? Lytx will classify those collision drivers as ‘RED’ nearly 70% of the time using this measurement methodology, while generally limiting the ‘RED’ drivers to 10-20% of the driver population. This measurement is intended to ensure companies have more time to change the drivers’ behaviors and focus on the higher priority drivers.

The Lytx predictive models are continually tuned to improve the Safety Score metric. The chart below shows a synopsis of Year-Over-Year performance and improvement on the Safety Score metric.
Lytx’s team of experts is continually refining its methodologies to allow for even greater accuracy and expects to continue to deliver industry-leading efficacy.

Over the past 16 years the number of sensors has increasingly improved the algorithms. By capturing ‘Black-Box’ telematics data, third-party data, as well as video observational analysis, Lytx is able to simulate varying Safety Score Models utilizing varying sets of the data to determine the importance of the variety and volume of information in relation to improvements to the model efficacy. The table below shows the predictive capabilities of varying datasets on the same drivers (and same timeframes) on the Lytx Collision Prediction Model Efficacy metric:
As is evident from the table above, when looking at the same drivers and time periods, telematics-only solutions are limited to < 40% efficacy on the Lytx Score Metric. By fusing Video (‘Our Non-Replicable Answer Key’ to Improve Analytics) with other data, Lytx is able to achieve double the predictive power of a ‘Black Box’ Only telematics solution.

An odds ratio (OR) is a measure of association between an exposure and an outcome. In the case of the Lytx Safety Score, the OR represents the odds that a driver will get into a collision in the next 90 days given a particular score.

As can be seen from the odds ratio chart, the odds of a driver getting into a collision are significantly increased when we attribute a low Safety Score.
PREDICTIVE ANALYTIC MODELING PROCESS

SENSE: DATA CAPTURE

In order to develop accurate predictive models and leverage the ‘Big V’s’ of data, Lytx uses a multi-sensor fusion technology. **Sensor fusion** is the combining of sensory data (or data derived from sensor data) from disparate sources such that the resulting information is in some sense better than would be possible when these sources were used individually. The term better in this case can mean more accurate, more complete, or more dependable, or refer to the result of an emerging view.

In the context of assessing risky driving and operational analysis, Lytx leverages several sources of data including, but not limited to, accelerometers, geospatial sensors, on-board sensors, posted speed, observational analysis (review of human behavior from video), and third-party environmental information. This data collection allows several types of sensors to be merged in the predictive model development that allows us to improve our Volume, Variety, Velocity, and Veracity of the data that enters into our algorithms for the most effective results.

PREDICT: MODEL DEVELOPMENT

In order to develop the analytic models, Lytx applies a standardized process and leverages several analytical techniques. In order to get the highest efficacy models, the technique and approach are varied slightly to provide the best results for the clients.

PROCESS

In order to ensure that the analytics projects are successful and don’t fall into the ‘Big Data Failure’ traps, strict processes are followed for the development of new algorithms. This process ensures the ability to iterate quickly and develop new and improved models efficiently for customers. Specifically six primary processes are used (timing within each process may vary based on the availability, cleanliness, and complexity of the algorithms that are built). The six key process steps are described below (a more detailed description of each process is listed in the appendix):
PREVENT: CHANGING BEHAVIOR USING ANALYTICS

There are many theories on how to change human behavior. However, all of the theories share some basic components:

1. Identify the target behavior(s)
2. Understand the target behavior(s) in context
3. Consider the possible intervention methods and implement and measure the behavior changes

In conjunction with the Lytx Insights™ Program, the Company leverages analytics in all of these areas to effectively prevent risky behaviors from reoccurring and change the way drivers interact in their work environment. Prevention requires analytics in all three areas:

<table>
<thead>
<tr>
<th>Behavioral Change Category</th>
<th>Analytics Category</th>
<th>Types of Analytics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identify Behaviors</td>
<td>Sense</td>
<td>Sensor Fusion (e.g. Machine, Observational, &amp; Third Party Data)</td>
</tr>
<tr>
<td>Understand Behaviors in Context</td>
<td>Predict</td>
<td>Predictive Models (e.g. Lytx Safety Score)</td>
</tr>
<tr>
<td>Intervention</td>
<td>Prevent</td>
<td>Predictive Models (e.g. Key Performance Indicators, Quarterly Performance Reviews, Industry Benchmarking)</td>
</tr>
</tbody>
</table>

The **Sense** and **Predict** models were discussed in the prior sections. They are prerequisites to the **Prevent** models, which are a blend of the Lytx Engine and the Lytx Insights Program.

Within the Prevent models, Lytx focuses on determining whether the intervention and change management methods are effective, allowing the team to recommend various intervention models that would be beneficial to
customers. In particular, key performance indicators (KPIs) are leveraged to determine if the data being captured is complete, timely and accurate (improves the Sense and Predict stages) as well as evaluating the efficacy of the management teams on changing behavior. The Prevent Analytics are also predictive models that isolate specific target variables (e.g. How many of this coach’s drivers will get in a collision in the next ‘n’ months?) so the coach’s behavior as well as the drivers’ behavior can be changed.

This type of information aids in the intervention process to effect lasting behavioral change. There are several types of interventions that may be appropriate depending on the behaviors that need to be changed, the severities of those behaviors, the personality of the people involved, and the environmental factors impacting the behaviors. The analytic assessments and Insights program can recommend the best intervention techniques.

**Intervention function** | **Definition** | **Driver Safety Examples**
--- | --- | ---
**Education** | Increasing knowledge or understanding | Providing information on safe driving awareness
**Persuasion** | Using communication to induce positive or negative feelings or stimulate action | Ensure awareness of safe driving around the workplace
**Incentives** | Creating expectation of reward | Financial bonus and recognition programs
**Coercion** | Creating expectation of discipline or cost | Verbal and written warnings
**Training** | Imparting skills | Advanced driver training to increase

### Intervention function | Definition | Driver Safety Examples
--- | --- | ---
**Restriction** | Using rules that limit engagement in the target behavior or competing or supporting behavior | Eliminate cell phones in vehicles when drivers are on jobs
**Environmental Restructuring** | Changing the physical or social context | Creating group bonuses to incent peer pressure amongst drivers
**Modelling** | Providing an example for people to aspire to or imitate | Create a ‘driver of the year’ program to showcase a good driver
**Enablement** | Increasing means/reducing barriers to increase capability or opportunity | Peer review risky driving videos to see what types of risky driving is occurring and how create awareness of the drivers’ own driving.

### CONCLUSION

The Lytx Safety Score helps save lives. The combination of Lytx technology and active driver education represents an innovative solution -- predictive analytics – which can benefit fleets, municipalities, delivery services and individual drivers who are committed to safety. Lytx stands for the entire process of Sense, Predict, Prevent™ empowering its customer to be safer, better companies, bringing drivers home safely at night and making the roads safer, for everybody.

### ABOUT LYTX

**About Lytx**

At Lytx (formerly DriveCam, Inc.), we harness the power of data to change human behavior and help good companies become even better. Our flagship product, DriveCam powered by Lytx™, is a leading solution for driver safety in the industries we serve, and our RAIR™ Compliance Services help DOT-regulated fleets comply with safety regulations, complementing the DriveCam program. Our solutions protect drivers from more than 950 commercial and government fleet clients worldwide who drive billions of miles annually. Our clients realize significant ROI by lowering operating and insurance costs, while achieving greater efficiency and compliance. Most of all, we strive to help save lives – on our roads and in our communities, every day. Lytx is privately held and headquartered in San Diego. For more information, visit [www.lytx.com](http://www.lytx.com).

# # #

Lytx; DriveCam powered by Lytx; Lytx Engine; Lytx Insights; Sense, Predict, Prevent; Delivering Insights. Driving Results; Lytx Safety Score; and RAIR are trademarks of Lytx, Inc. DriveCam and DriveCam Online are registered trademarks of Lytx, Inc.
This section amplifies the description on page 12 of the six processes Lytx uses to develop its predictive model.

**IDENTIFY TARGET VARIABLE**

The most critical step in our process is understanding the business problems that we are trying to solve. Leveraging our Product Management and Product Marketing teams, we are able to get insight into the problems that our customers are trying to solve at a granular level. Most big data projects historically have tried to capture more data and a variety of data without really understanding the problem that they are trying to solve.

This is like looking at a giant haystack with the task of “find something interesting”. This task is futile and usually just wastes time and money as the task is too broad. If we take the same haystack and create a task of “finding the needle in the haystack,” we now have a viable problem that we can solve. We can determine the attributes associated to the needle (e.g. how its size differs from the hay, how its composition is different, where it is most often found in the haystack, etc.) and build predictive models to determine where the next needle will be found.

Once a business problem has been identified, we determine a target variable (the variable that we are trying to predict) that will allow us to address the question we are trying to answer. Often the target variable will be a calculated or derivative variable (e.g. what % of my drivers got in a collision) and it could also be a simple classification (e.g. did this driver get in a collision?). The following steps allow us to build attributes that may be correlative and predictive of the target variable.

**DATA CLEANSING AND ORGANIZING**

The data must be cleansed of missing values and outliers. Missing values are, quite literally, missing data. For example, there may be a number of records in a database for which the driver of a vehicle has not been identified. Handling missing data may be tricky; since the information may be unavailable for a number of different reasons (e.g. Data not provided, bad data was entered, data was excluded as an outlier). The analyst must decide whether to include, exclude, or replace them with a more typical value for that variable to ensure the models are not skewed by bad records. Data is never completely clean, and typically minor variances in data cleanliness will not have a material impact on the predictive models. However, if data accuracy and cleanliness is a widespread problem, it is important that the data quality is factored into the final predictive model. We often utilize weighting mechanisms to adjust for the quality of the data if certain data is suspect.

**DATA EXPLORATION AND VISUALIZATION**

The analyst examines the data, computing and analyzing various summary statistics and derivatives regarding the different variables contained in the data set. Categorical variables are variables such as gender and driver’s license status that possess no natural numerical order. These variables must be identified and handled separately from continuous numerical variables such as length of time with a commercial driver’s license and driver’s age.

In this step, we often chart and display the data in various manners to aid us in visualizing correlative relationships and seeing the variance and outliers in each attribute.

The analyst will also randomly split the data into two sets. One of which will be used to build, or train, the model; the other of which will be used to test the quality of the model once it is built.
DATA MINING AND VARIABLE REDUCTION

Data mining is the analysis of data to identify underlying trends, patterns, or relationships. It is a necessary first step in predictive analytics, because the data that the mining process identifies as relevant can then be used to develop the predictive model. One can think of data mining as gathering knowledge about relationships, and the resulting predictive analytics model as applying that knowledge. One distinct advantage to data mining is that it catalogs all relationships (correlations) that may be found among the data, regardless of the cause.

The variables in the data set are examined in terms of their correlation with the target variable. Those that are redundant or highly correlated with each other are removed. For example, a data set may include ‘hours per week driving’ and ‘hours per week non-driving duty statuses. As these variables are most likely the inverse of one another, they may move together and actually hinder the model. In some situations, including both could be problematic.

The remaining variables are often re-scaled so that they can be more efficiently analyzed.

PREDICTIVE MODEL DEVELOPMENT

There are several modeling techniques (see partial list below) that we employ to get the best performing algorithms. After the data mining stage, we conduct peer reviews to analyze the data and recommend potential modeling techniques that would be applicable to the problem set. We also review whether supervised, unsupervised or hybrid (supervised and unsupervised combined) modeling techniques are appropriate based on the early data findings and the frequency in which we expect the data to change.

We also leverage several modeling tools that allow us to rapidly analyze data utilizing multiple techniques and it allows us to quickly identify the modeling techniques that are performing better with the data set.

Based on the outcome of the early models, we also look at ensemble modeling, stacking and/or blending methods, and ‘bucket of models’ which allow us to layer models on top of each other, run multiple models and use voting methods to select the correct model, and embed predictive model results in other predictive models.

Ensembles tend to yield better results as they use multiple models to obtain better predictive performance than could be obtained from any of the single constituent model.

MODEL SELECTION

Once several models have been developed, we analyze the efficacy (how well each model is predicting the target variable) and isolate the better models for further validation. We look at the efficacy of the models as well as the weighting of each of the attributes to determine if the models are working appropriately and often we will find counter-intuitive insights on how attributes react with the target variable.

MODEL VALIDATION

In the validation stage, we run our new predictive models using the test data taken from the original data set (Data Exploration and Visualization Step). To ensure that a predictive model is as accurate as possible, it must be validated through out-of-sample testing. Out-of-sample testing divides data into in-sample data (the data used to develop the model) and out-of-sample data (the model testing window), which includes only data not used.

Sound statistical practices dictate that multiple in-sample and out-of-sample data windows be used to develop and validate a predictive model. Using multiple out-of-sample testing windows minimizes the influence of specific single events.
Once we’ve determined how the models performed in the validation steps versus the training data sets, we are able to select the best predictive model and prepare it for deployment to our customers.

### ANALYTICAL TECHNIQUES

We leverage a multitude of analytical techniques to develop predictive models. The approaches and techniques that we use to conduct predictive analytics can broadly be grouped into regression techniques and machine learning (advanced modeling) techniques.

Predictive models can assume many shapes and sizes, depending on their complexity and the application for which they are designed. This section introduces some of the statistical methods that we use to develop a predictive model. Unlike data mining where we are looking at multiple correlations and analyzing the data, many of the statistical procedures that are employed in predictive models aim to predict a specific target variable.

### REGRESSION TECHNIQUES

Regression models are the mainstay of predictive analytics. The focus lies on establishing a mathematical equation as a model to represent the interactions between the different variables in consideration. Depending on the situation, there is a wide variety of models that can be applied while performing predictive analytics. Some of them are briefly discussed below.

Regression modeling mathematically describes the relationship between the dependent variable (The Target Variable, or the variable to be predicted in predictive models) and independent, explanatory variables given sample data. Regression implies some causation, unlike correlations since modelers must specify a relationship beforehand and then test how well the regression model fits for models the specified relationship.

**LINEAR REGRESSION MODEL (MULTIVARIATE MODELS FOR CONTINUOUS DEPENDENT VARIABLE VALUES)**

The linear regression model analyzes the relationship between the target dependent variable and a set of independent or predictor variables. This type of model is typically used when the target variable is no a discrete value (meaning it has a range of values such as claims dollars per incident). This relationship is expressed as an equation that predicts the response variable as a linear function of the parameters. These parameters are adjusted so that a measure of fit is optimized. Much of the effort in model fitting is focused on controlling the variance and minimizing the size of the residual. The goal of regression is to select the parameters of the model so as to minimize the sum of the squared residuals. This is referred to as ordinary least squares (OLS) estimation and results in best linear unbiased estimates (BLUE) of the parameters (assuming Gauss-Markov assumptions are satisfied). How well the model predicts the dependent variable based on the value of the independent variables can be assessed by using the $R^2$ statistic. It measures the predictive power of the model (i.e. the proportion of the total variation in the dependent variable that is "explained" by variation in the independent variables). Typically an $R^2$ value of 0.7 or higher indicates a strong predictive model.

**LOGISTIC REGRESSION (DISCRETE LOGIT MODELS WHEN THE DEPENDENT VARIABLE IS BINARY)**

In a classification setting, assigning outcome probabilities to observations can be achieved through the use of a logistic model, which is basically a method which transforms information about the binary dependent variable into an unbounded continuous variable (e.g. probability of a collision) and estimates a regular multivariate model. The Wald and likelihood-ratio test are used to test the statistical significance of each coefficient $b$ in the model (analogous to the t tests used in OLS regression). A test assessing the goodness-of-fit of a classification model is the "percentage correctly predicted".
PROBIT REGRESSION (DISCRETE MODELS WHEN THE DEPENDENT VARIABLE IS BINARY)

A probit model is a similar regression-type model like the logit model, but it is a type of regression where the dependent variable can only take two values, for example ‘Drivers with a Collision’ or ‘Drivers without a Collision’. The name is from \textit{probability} + \textit{unit}. The purpose of the model is to estimate the probability that an observation with particular characteristics will fall into a specific one of the categories. The suitability of an estimated binary model can be evaluated by counting the number of true observations equaling 1, and the number equaling zero, for which the model assigns a correct predicted classification. The coefficients obtained from the logit and probit models are fairly close. However, the odds ratio is easier to interpret in the logit model which makes it a preferred method for predictive algorithms. However, there are two reasons why the probit model is sometimes preferred:

- There is a strong belief that the underlying distribution is normal (logit typically assumes a linear relation).
- The actual event is not a binary outcome (\textit{e.g.}, collision) but a proportion (\textit{e.g.}, proportion of total collisions that are severe).

TIME SERIES MODELS

Time series models are used for predicting or forecasting the future behavior of variables. These models account for the fact that data points taken over time may have an internal structure (such as autocorrelation, trend or seasonal variation) that should be accounted for. For instance, risky driving behaviors may be more prevalent during winter months over summer months (seasonality of data) or certain types of vehicles may be driven more during certain months of the year. As a result, standard regression techniques cannot be applied to time series data and a methodology has been developed to decompose the trend, seasonal and cyclical component of the series. Time series models estimate difference equations containing stochastic (variables having a random probability distribution or pattern that may be analyzed statistically but may not be predicted precisely; such as weather) components. These types of models allow us to derive meaningful statistics and other characteristics of data. The results of these models are typically entered into our advanced predictive models for further analysis.

SURVIVAL OR DURATION ANALYSIS

Survival analysis is another name for time to event analysis. These techniques were primarily developed in the medical and biological sciences, but they are also widely used in the social sciences like ‘Risky Driving’, as well as in engineering (reliability and failure time analysis).

Conventional statistical models, such as multiple linear regression, have difficulty with survival data which typically has data that is censored (\textit{e.g.} a non-collision driver that we have history on, but will eventually get in a collision) or a non-normal distribution. Normally distributed data has a symmetric distribution, meaning it could take positive as well as negative values, but duration (Time to Incident) cannot be negative and therefore normality cannot be assumed when dealing with duration/survival data.

We collect data on collision and non-collision drivers frequently, but we know that a percentage of the non-collision drivers will eventually be classified as collision drivers. Survival analysis allows us to predict the time to an event (\textit{e.g.}, a collision) as well as the probability of not having the event (\textit{e.g.} never have a collision).

CLASSIFICATION AND REGRESSION TREES (DECISION TREE ANALYSIS)

Hierarchical Optimal Discriminant Analysis (HODA), (also called classification tree analysis) is a generalization of optimal discriminant analysis that may be used to identify the statistical model that has maximum accuracy for predicting the value of a categorical dependent variable for a dataset consisting of categorical and continuous variables. The output of HODA is a non-orthogonal tree that combines categorical variables and cut points for
continuous variables that yields maximum predictive accuracy, an assessment of the exact Type I error rate, and an evaluation of potential cross-generalizability of the statistical model.

Classification and Regression Trees (CART) is a non-parametric decision tree learning technique that produces either classification or regression trees, depending on whether the dependent variable is categorical or numeric, respectively.

Decision trees are formed by a collection of rules based on variables in the modeling data set:

- Rules based on variables' values are selected to get the best split to differentiate observations based on the dependent variable (e.g., Drivers that have had a ‘Drowsy Driving’ behavior may head down one path of the tree, and drivers without will head down another path).
- Once a rule is selected and splits a node into two, the same process is applied to each "child" node (i.e. it is a recursive procedure)
- Splitting stops when CART detects no further gain can be made, or some pre-set stopping rules are met. (Alternatively, the data are split as much as possible and then the tree is later pruned.)

Each branch of the tree ends in a terminal node (this defines the Target Variable – e.g., driver will have a collision or not). Each observation falls into one and exactly one terminal node, and each terminal node is uniquely defined by a set of rules.

ADVANCED MODELING TECHNIQUES

In addition to regression models, we leverage advanced modeling techniques such as neural networks to create predictive models. Neural networks are nonlinear statistical modeling tools. In general, neural networks can handle many more variables than the regression techniques and they also address some of the other limitations associated with regression techniques such as statistical concerns regarding dimensionality (meaning the data may have non-linear or non-normalized causation patterns and there may be other correlations that are not causal). In particular, we leverage Machine Learning Techniques, Unsupervised Machine Learning Techniques, and Ensemble Modeling.

MACHINE LEARNING TECHNIQUES

Machine learning is a branch of artificial intelligence that was originally employed to develop techniques to enable computers to learn. Today, since it includes a number of advanced statistical methods for regression and classification, it finds application in a wide variety of fields including medical diagnostics, credit card fraud detection, face and speech recognition and analysis of the stock market. In certain applications it is sufficient to directly predict the dependent variable without focusing on the underlying relationships between variables. In other cases, the underlying relationships can be very complex and the mathematical form of the dependencies unknown. For such cases, machine learning techniques emulate human cognition and learn from training examples to predict future events.

NEURAL NETWORKS

Neural networks are nonlinear sophisticated modeling techniques that are able to model complex functions. They can be applied to problems of prediction, classification or control. Neural networks are used when the exact nature of the relationship between inputs and output is not known. A key feature of neural networks is that they learn the relationship between inputs and output through training. There are three types of training in neural
Neural networks is a powerful Artificial intelligence technique that mimics the operation of the human brain (nerves and neurons), and comprises of densely interconnected computer processors working simultaneously. A key feature of neural networks is that they are programmed to 'learn' by sifting data repeatedly, looking for relationships to build mathematical models, and automatically correcting these models to refine them continuously adapting and learning from past patterns.

This self-learning and development of non-linear understanding of the independent and target variables makes this an ideal technique for us to process the variety of data that we capture on driving.

**SUPPORT VECTOR MACHINES**

Support Vector Machines (SVM) are used to detect and exploit complex patterns in data by clustering, classifying and ranking the data. They are supervised learning machines that are used to perform binary classifications and regression estimations. They commonly use methods to apply linear classification techniques to non-linear classification problems. Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other, making it a non-probabilistic binary linear classifier. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible (basically, they take inputs and are able to generate a two dimensional chart that clearly shows separations in the groups – e.g. collision vs non-collision drivers). There are a number of types of SVM such as linear, polynomial, sigmoid etc.

**GEOSPATIAL PREDICTIVE MODELING**

Conceptually, geospatial predictive modeling is rooted in the principle that the occurrences of events being modeled are limited in distribution. Occurrences of events are neither uniform nor random in distribution – there are spatial environment factors (infrastructure, sociocultural, topographic, etc.) that constrain and influence where the locations of events occur. Geospatial predictive modeling attempts to describe those constraints and influences by spatially correlating occurrences of historical geospatial locations with environmental factors that represent those constraints and influences. Geospatial predictive modeling is a process for analyzing events through a geographic filter in order to make statements of likelihood for event occurrence or emergence. This is particular prevalent when we look at seasonality based on geography (winter in Minnesota vs San Diego) and the types of driving (urban vs rural) to understand risky driving.

**UNSUPERVISED LEARNING**

In machine learning, the problem of unsupervised learning is that of trying to find hidden structure in unlabeled data. Since the examples given to the learner are unlabeled, there is no error or reward signal to evaluate a potential solution. This distinguishes unsupervised learning from supervised learning and reinforcement learning. For instance, if we provide an unsupervised model with different driving information about vehicles (without labeling the type of vehicle), the model should be able to develop a classification algorithm on its own to determine the type of vehicle being driven. With enough data, the model can distinguish a passenger car from a waste truck simply by analyzing the driving profile. Unsupervised learning encompasses many other techniques that seek to summarize and explain key features of the data. Many methods employed in unsupervised learning are based on data mining methods used to preprocess data.
ENSEMBLE MODELING

In statistics and machine learning, ensemble methods use multiple models to obtain better predictive performance than could be obtained from any of the constituent models. An ensemble is a technique for combining many weak models in an attempt to produce a strong model. An ensemble is itself a supervised learning algorithm, because it can be trained and then used to make predictions. The trained ensemble, therefore, represents a single prediction even though it is a combination of models (some attempting the same prediction and some predicting other attributes that aid in the final prediction).

Empirically, ensembles tend to yield better results when there is a significant diversity among the models. Many ensemble methods, therefore, seek to promote diversity among the models they combine. Although perhaps non-intuitive, more random algorithms (like random decision trees) can be used to produce a stronger ensemble than very deliberate algorithms. Using a variety of strong learning algorithms, however, has been shown to be more effective than using techniques that attempt to dumb-down the models in order to promote diversity.

COMMON TYPES OF ENSEMBLES

BAYES OPTIMAL CLASSIFIER

The Bayes Optimal Classifier (BOC) is a classification technique. It is an ensemble of all the hypotheses in the hypothesis space (A hypothesis space is a predefined space of the potential hypothesis or outcomes). On average, no other ensemble can outperform it, so it is the ideal ensemble. Each hypothesis is given a vote proportional to the likelihood that the training dataset would be sampled from a system if that hypothesis were true (this generates a weighted average voting system to optimize the predictive power). To facilitate training data of finite size, the vote of each hypothesis is also multiplied by the prior probability of that hypothesis. There are also sophisticated methods of approximating the BOC (Bayesian Model Averaging and Bayesian Model Combination) that are more practically implemented than the BOC.

BOOTSTRAP AGGREGATING (BAGGING)

Bootstrap aggregating, often abbreviated as bagging, and involves having each model in the ensemble vote with equal weight. In order to promote model variance, bagging trains each model in the ensemble using a randomly drawn subset of the training set.

BOOSTING

Boosting involves incrementally building an ensemble by training each new model instance to emphasize the training instances that previous models misclassified. This type of modeling allows us to take prior models that may have misclassified some of the data and reprocess that portion of the data for more predictive accuracy.

BUCKET OF MODELS

A "bucket of models" is an ensemble in which a model selection algorithm is used to choose the best model for each problem. When tested with only one problem, a bucket of models can produce no better results than the best model in the set, but when evaluated across many problems, it will typically produce much better results, on average, than any model in the set. In essence, this algorithm creates a ‘bake-off’ between models to determine which model is best for the particular problem.
**STACKING**

Stacking involves training a learning algorithm to combine the predictions of several other learning algorithms. First, all of the other algorithms are trained using the available data, then a combiner algorithm is trained to make a final prediction using all the predictions of the other algorithms as additional inputs. Stacking typically yields performance better than any single one of the trained models. It has been successfully used on both supervised learning tasks and unsupervised learning.