W205 Course Project Final Report

DriveSafe

Commuter Risk Application

David Jablonski, Jennifer Middleton, Steve Pelkey, Kari Ross

Dec 15th, 2016

W205 Course Project Final Report

Commuter Risk Application

David Jablonski, Jennifer Middleton, Steve Pelkey, Kari Ross

Dec 15th, 2016

Table of Contents

List of Figures

Introduction

Businesses can save millions of dollars by helping their employees reduce the number and severity of motor vehicle accidents they are involved in. DriveSafe, our innovative accident risk assessment and alerting product is a critical tool for helping businesses improve bottom line performance while providing a key distinguishing corporate benefit to their employees.

Accidents that result in injury cost billions of dollars every year. In 2010 the Centers for Disease Control and Prevention placed the cost of medical care and productivity losses associated with motor vehicle crash injuries at over $99 billion. This translates into nearly $500, for each licensed driver in the United States. Businesses large and small bear the brunt of these costs, not to mention the impact on individual employees and their families when these incidents and tragedies occur.

Numerous studies have shown that crashes are more likely in inclement weather despite the fact that motorists adjust their driving habits. For example, motorists adjust their road behavior during rain. They overtake less, driver slower, and increase their following distance (Hogema, 1996; Agarwal et al., 2005). However, the risk of a crash during rain is still greater than in dry weather. The changes in driving behavior are, apparently, insufficient to compensate for the greater risk during inclement weather (Thoma, 1993). Another study showed that there was an increase in the number of crashes of between 35% and 182% when it rained. Ice forming on the road surface even led to an increase of between 77% and 245%. However, ice forming is far less frequent than rain, and thus has a smaller impact on the total number of crashes (Stiers, 2005).

DriveSafe alerts companies and their employees to the increased risk of injury and the likelihood of incurring a more serious injury based on the weather conditions. This alone can directly impact the bottom line. Businesses that go further and authorize remote work during adverse weather can potentially save even more. All while providing a unique and valuable resource to their workforce at an excellent price point. DriveSafe is an employee benefit that can't be beat.

Product Capabilities

Risk Score Calculations

DriveSafe evaluates the impact of weather on motor vehicle accidents. The DriveSafe team has built a model calculating relative risk scores for accident severity based on historical accidents in the area. Weather conditions are updated hourly, and based on current conditions, users are notified when there is increased driving risk in their area based inclement weather conditions.

Data Engine

The DriveSafe Data Engineering team has amassed an initial base of weather and accident data. Our elite data engineers are constantly adding to our core repository of accident and weather data. Currently focused on Maryland, DriveSafe is planning nation-wide expansion, tapping Federal, State, and Local data.

The initial repository is loaded with raw accident data from the State of Maryland, including vehicles and persons involved. This data is enhanced with information about the closest weather station based on the latitude and longitude of the accident, using the spherical law of cosines. These methods are generally applicable and will be used in the nation-wide buildout of the product. Additional data quality control and improvements will be added validating latitude and longitude as well as validating police reports of road conditions against official weather sources.

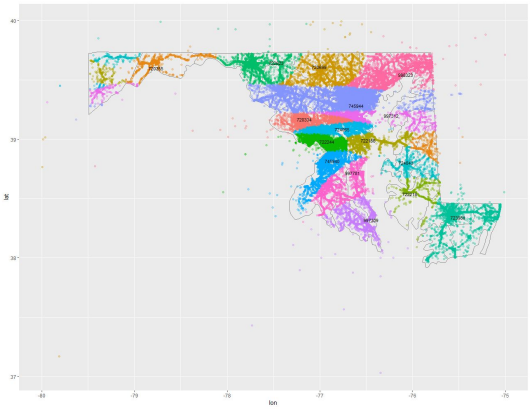

Figure 1: DriveSafe Weather Station and Crash Analysis

In addition to the accident data, DriveSafe has hourly data from NOAA. This comples and variable length data has been parsed and loaded into our distributed repository. The data is partitioned to support national scalability and enhanced with hourly counts of nearby accidents. Upcoming releases will extend our accumulated weather data to include hourly precipitation data.

Serving Layer

The DriveSafe infrastructure serves up summary data to S3, which is then accessed by our serving layer. Our Website, https://w205accidents.shinyapps.io/test/, then delivers up the Relative Risk of Injury or Death based on your nearest Weather Station and the current conditions. Upcoming releases will offer API based access as well as SMS alerts. 2018 plans include a mobile app which can be customized to include corporate weather announcements and remote approvals.

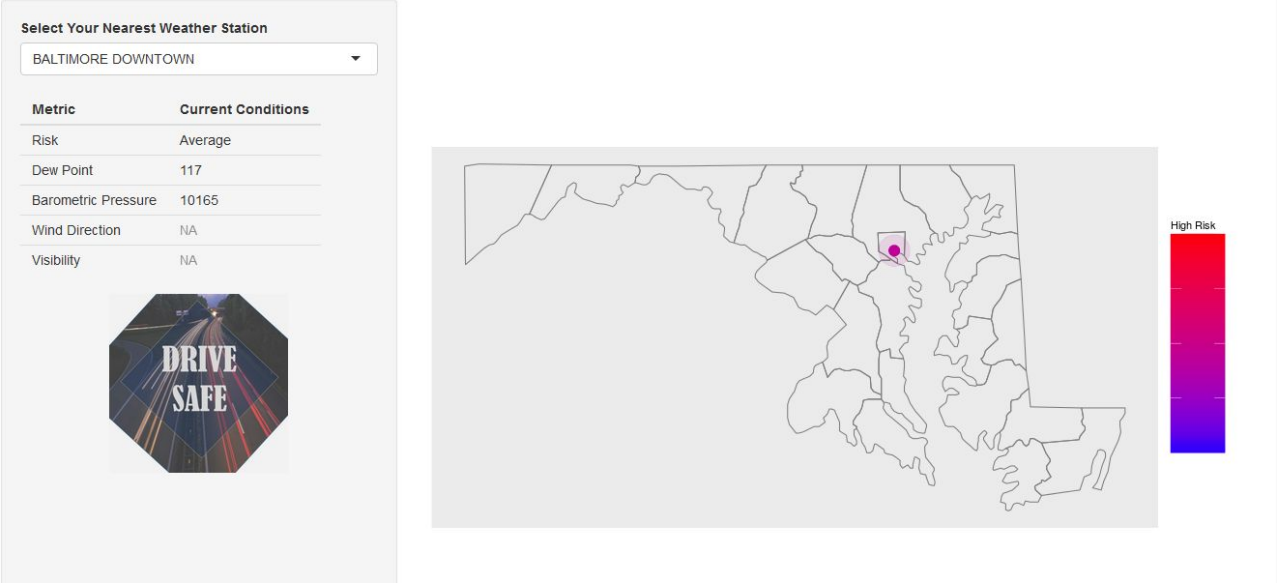

Figure 2: DriveSafe Website Application

Product Architecture

The bulk of the Information Services were done through Amazon Web Services (AWS). This allowed the flexibility of the team to utilize cloud computing for the project. The NOAA data was downloaded to AWS and entered into Hadoop/Parquet covering the last 40 years of data. The team created a Master Table of Weather Stations, including ID, Lat, Lon, and downloaded Maryland crash data to AWS and entered it into Hadoop/Parquet. The accident observations (Maryland) were matched to the nearest weather station and added to corresponding column to the accident master dataset. Weather and datetime variables were then used to compute the likelihood of an accident using a logistic regression with R. NOAA data is subsequently updated in real time, copied to an S3 bucket retrieved by a Shiny instance where risk scores are calculated and displayed. The Architecture Diagram of the end-to-end product starting from the weather and crash data to the users making their commuting decisions is located in Figure 1. [Steve please add some additional for this.]

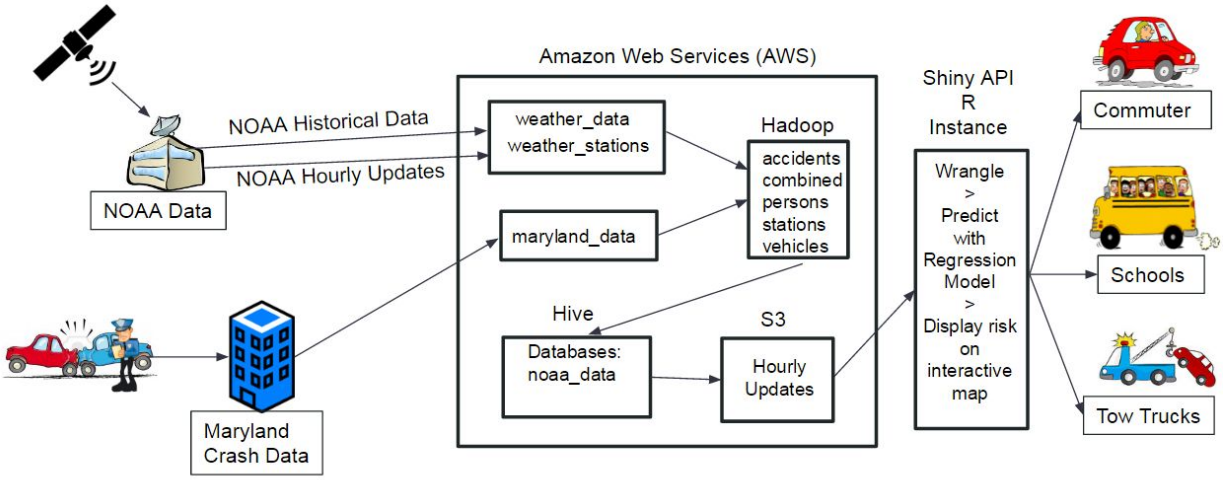

Figure 3: DriveSafe Application Architecture

A number of technologies were pulled together to build this architecture. Hadoop and Hive were used extensively for data storage and scalability. bash, cron, python and Hiveql were used to perform the ETL processes and to build the serving layer. python, R and, of course, HTML and css were used to build the front-end.

Product Futures

Data

First priority will be expanding the data engine to include precipitation data since the expanded fields associated with the NOAA hourly data are not uniformly populated. The substantial amount of missing data required us to do a more limited analysis than originally intended. Hourly precipitation data is more reliably available in a separate dataset and we will pull that data in to enhance our engine.

In addition to incorporating precipitation data, the DriveSafe team will be enhancing the streaming data. NOAA updates are slow and we have been identifying other sites, including the Farmers Almanac as possible sources of data that updates more in real-time.

Finally, we are committed to expanding our offering nationwide. There are data challenges with this, as Maryland has more complete accident records than many of the other states, but we believe that by supplementing with local sources and building out our rigorous data quality program we will have sufficient data to support the required analyses.

Analysis

Armed with a further enhanced data engine, our proprietary algorithm will be able to reliably incorporate more variables. Currently, dewpoint was the only variable with statistical significance, however the actual weather observations were so sparse they could not be used as the basis for analysis. Precipitation data will help with that tremendously as other studies have shown precipitation to be one of the largest predictors of accidents and accident severity.

Moving forward improvements will be made to the risk calculations. Now that we have learned more about the structure and relationships in the data, a time series model will be developed for DriveSafe to be able to respond to new data as regions driving and weather change. This will also help provide better results for any seasonality in the driving and weather patterns. This will especially help in areas where the weather change in more dramatic.

Scaling

The DriveSafe data engine has been architected for scale, so as we build out our data sources, the data engine will scale accordingly. Both S3 and Shiny.io are designed to scale out. However, multiple different web applications might eventually be needed as our analyses become more complicated and features are added.

Customization

Our original vision included customizations for users based on risk preferences, location information and corporate affiliations. These will be incorporated with the mobile release.

Figure 4: Future DriveSafe iPhone Application

Product Risks

Data availability might be impaired as US Government Data Policy changes.

Data quality and completeness could continue to limit the scope of analysis.

Customization framework may require the collection of personalized data, introducing more security concerns.

Return on Investment

Customers

Based on the 2010 CDC estimate of $500 per licensed driver in the US, the savings for large corporations of more than 20,000 employees could be substantial. The losses cannot be completely ameliorated, but even a 20% drop in accidents, which can be achieved by reducing the number of drivers in adverse conditions would result in savings of approximately $1,000,000. Since the annual license fee for the DriveSafe corporate offering is $150,000, results could be upwards of 5x ROI. This type of return could translate into significant sales. Per use pricing for smaller companies can also generate a recurring stream of revenue. Individual and family use provides a wide market, with an advertising opportunity to market driving and safety products to a very targeted audience.

Exit Strategy

DriveSafe fits well within the corporate benefit solution space, and DriveSafe has every expectation that one of the larger corporate benefits providers could be a potential buyer.

Alternatively, if randomized user testing does not demonstrate sufficiently high adoption rates and behavioral change rates, DriveSafe will not obtain the intended levels of market penetration and it is possible that it will go bankrupt.

Product Challenges

While the weather data supplied by NOAA covers the entire U.S., the crash dataset is provided by the state and each state might not make their police information public for use by DriveSafe. This would hinder development into these markets.

Also, some accidents go unreported, so this would affect the risk scores but apps like Waze have drivers report accidents they see which gives a time and location of the accident. This information could be leveraged along with the police information to give a more precise calculation.

Lastly public data sources often have data quality issues. In the case of the weather

data, the amount of missing data impacts the scope of the possible analysis.

MVP Award

Every great team has a lot of great players. But usually there is one at any particular point in time -- whether a game or a season -- that stands out more than the others. For our team, on this effort, the MVP Award goes to Kari Ross for building the most complex portion of the DriveSafe Data Engine. Thanks Kari!

Kari Ross
SSD Engineer @ Seagate, MIDS Student @ Berkeley

Appendix

Summary of NOAA Dataset

Starting from the Data.gov website which has many open source data sets, the team focused in on the NOAA weather data https://www.ncdc.noaa.gov/cdo-web/datasets. Is has many advantages going to for it. It's open source and free. It has decades of history and it is complete and update regularly using the Historical NOAA data and the NOAA API for hourly updates. The weather stations provide coverage over the entire United States. This allows the flexibility to expand the project as the capability is developed. Now that the team has reliable weather data source.

Summary of Maryland Crash Dataset

Originally Crash fatality data from New York was the data set that was selected but upon further review, the team wondered to open it up to all crashes to assess the risk of the commute. An extensive data set for the vehicle crash data, Maryland Statewide Vehicle Crash Data, https://data.maryland.gov/Public-Safety/Maryland-Statewide-Vehicle-Crash-Data-Dictionary/7xpx-5fte/data.

ETL Methods

Data was pulled into Hadoop using Python scripts that parsed, cleaned and transformed the data, merging data into a structure that allowed the combining of the two datasets.

Amazon Web Services

By utilizing Amazon Web Services (AWS), DriveSafe can rapidly and effectively expand to fit the storage and computing power needed to scale to new markets. This has already been proven by big companies like Airbnb and Netflix that have built their data infrastructure using AWS. Maryland is the test market for this new and powerful application. The security inherent in using AWS means DriveSafe's code will be protected from any malicious attempts.

HIVE Database

All data was first brought into Hadoop and then into HIVE for ease of use and scalability. The data was partitioned to support scaling as more and more data is incorporated into the DriveSafe data engine. Hourly updates are pushed into HIVE for availability to the serving layer.

S3 / Shiny

Shiny is a web application framework for R that creates data dashboards. Users can manipulate the web interface which causes the server to update the interface using R. DriveSafe uses shinyapps.io, a scalable platform as a service, to host our Shiny application. An R script hosted by shinyapps.io pulls the most recent weather data from an Amazon S3 bucket, and manipulates, analyzes, and displays these data through a public interactive display at https://w205accidents.shinyapps.io/test/.