TRANSFERABILITY AND ROBUSTNESS OF PREDICTIVE MODELS TO PROACTIVELY ASSESS REAL-TIME FREEWAY CRASH RISK

A Thesis
presented to
the Faculty of California Polytechnic State University,
San Luis Obispo

In Partial Fulfillment
of the Requirements for the Degree
Master of Science in Civil and Environmental Engineering

by
Cameron Hunter Shew
October 2012

COMMITTEE MEMBERSHIP

TITLE:	Transferability and Robustness of Predictive Models to Proactively Assess Real-Time Freeway Crash Risk
AUTHOR:	Cameron Hunter Shew
DATE SUBMITTED:	October 2012
COMMITTEE CHAIR:	Anurag Pande, Ph.D. Assistant Professor Department of Civil and Environmental Engineering
COMMITTEE MEMBER:	Kimberley Mastako, Ph.D. Lecturer Department of Civil and Environmental Engineering
COMMITTEE MEMBER:	Cornelius Nuworsoo, Ph.D., AICP Associate Professor Department of City and Regional Planning

ABSTRACT

Transferability and Robustness of Predictive Models
To Assess Real-Time Freeway Crash Risk
Cameron Hunter Shew

This thesis describes the development and evaluation of real-time crash risk assessment models for four freeway corridors, US-101 NB (northbound) and SB (southbound) as well as I-880 NB and SB. Crash data for these freeway segments for the 16-month period from January 2010 through April 2011 are used to link historical crash occurrences with real-time traffic patterns observed through loop detector data.

The analysis techniques adopted for this study are logistic regression and classification trees, which are one of the most common data mining tools. The crash risk assessment models are developed based on a binary classification approach (crash and non-crash outcomes), with traffic parameters measured at surrounding vehicle detection station (VDS) locations as the independent variables. The classification performance assessment methodology accounts for rarity of crashes compared to non-crash cases in the sample instead of the more common pre-specified threshold-based classification.

Prior to development of the models, some of the data-related issues such as data cleaning and aggregation were addressed. Based on the modeling efforts, it was found that the turbulence in terms of speed variation is significantly associated with crash risk on the US-101 NB corridor. The models estimated with data from US-101 NB were evaluated based on their classification performance, not only on US-101 NB, but also on the other three freeways for transferability assessment. It was found that the predictive model derived from one freeway can be readily applied to other freeways, although the classification performance decreases. The models which transfer best to other roadways were found to be those that use the least number of VDSs–that is, using one upstream and downstream station rather than two or three.

The classification accuracy of the models is discussed in terms of how the models can be used for real-time crash risk assessment, which may be helpful to authorities for freeway segments with newly installed traffic surveillance apparatuses, since the real-time crash risk assessment models from nearby freeways with existing infrastructure would be able to provide a reasonable estimate of crash risk. These models can also be applied for developing and testing variable speed limits (VSLs) and ramp metering strategies that proactively attempt to reduce crash risk.

The robustness of the model output is assessed by location, time of day and day of week. The analysis shows that on some locations the models may require further learning due to higher than expected false positive (e.g., the I-680/I-280 interchange on US-101 NB) or false negative rates. The approach for post-processing the results from the model provides ideas to refine the model prior to or during the implementation.

Keywords: Real-time crash risk, data mining, classification tree, proactive traffic management, loop detector data, transferability, robustness

ACKNOWLEDGMENTS

First and foremost, I would like to thank my advisor, Dr. Anurag Pande, for his dedication to the success of this research. Dr. Pande's expertise in proactive crash risk assessment proved to be an invaluable asset to the project, as well as his tremendous contributions of personal time overseeing my work and helping with the modeling process.

I also thank Dr. Cornelius Nuworsoo of the Cal Poly City and Regional Planning Department for his feedback in editing portions of this thesis, particularly the content submitted to the Mineta Transportation Institute (MTI) and published as *MTI Report 11-15, Proactive Assessment of Accident Risk to Improve Safety on a System of Freeways*. The Mineta Transportation Institute funded this research, and I thank Research Director Dr. Karen Philbrick and Executive Director Rod Diridon for their support. I would also like to thank MTI staff, including Director of Communications and Special Projects Donna Maurillo and Administrative Assistant Jill Carter for their assistance.

I am very grateful for the support and mentorship of Cal Poly Civil Engineering lecturer Dr. Kimberley Mastako, who has personally advocated for me and helped me secure several internships. I thank Drs. Pande, Nuworsoo, and Mastako for volunteering to serve on my thesis committee.

I thankfully acknowledge the use of the Caltrans PeMS (Performance Measurement System) in conducting this research. I extend my deepest appreciation to Dr. Alexander Skabardonis of the Institute of Transportation Studies (ITS) at the University of California Berkeley, who was helpful in providing access to the data. Many thanks to Dr. Koohong Chung of Caltrans, who gave valuable comments on the research idea at the proposal stage and then provided important leads for the data used herein. Mr. Joe Yu (a graduate student at Cal Poly) also helped with this initial effort.

On a personal note, I would like to thank my parents, Benjamin and Kirsten Shew, my brother, Colin Shew, and my grandmothers, Jacqueline Sutherland and Lily Lung, for their love and support, including their financial investments in my education. To my wonderful girlfriend, Gina Bagdonas, thank you for being my close confidant, best friend, and so much more. To all of my other best friends, particularly Adam Nash, James Loy, Jenna Korver, Danelle Wacker, Grant Guillen, and the rest of my CLDC peeps, thank you for all of the fond memories, great dances, and friendships which I will cherish forever. I love you all.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

I. INTRODUCTION

Much progress has been made in recent years in shifting from reactive (incident detection) to proactive (real-time crash risk assessment) traffic strategies as traffic safety on freeways continues to be a growing concern. Reliable models that can take in real-time loop detector information, and discern normal flow conditions from crash-prone conditions, are keys to implementing crash preventative measures. This area of research has gained increased attention since the vehicle detector stations (VDS) on freeways have been able to gather real-time traffic data and the capabilities to collect, archive, and analyze these data have grown manifolds in the recent past.

This thesis presents the findings of a study sponsored by the Mineta Transportation Institute (MTI), and carried out jointly by the California Polytechnic State University, San Luis Obispo (Cal Poly) and San Jose State University (SJSU). This research effort aims to not only develop statistical models relating traffic flow variables to crash likelihood, but to test the transferability of these models on other, nearby freeway corridors. A few past studies have already demonstrated that statistical links between real-time traffic flow variables (such as average speed, volume, occupancy, and their respective standard deviations) and crash likelihood can be established. However, all of these previous studies have mostly focused on one particular highway corridor. This research advances the current body of knowledge by exploring whether driver characteristics and behavior are similar enough in close geographic proximity to accurately apply the estimated classification models from one roadway segment onto another.

This thesis also explores model robustness, including patterns in misclassification errors (false positives and negatives), as well as their potential causes. While the safety applications through intelligent transportation systems (ITS) need to be studied further, this study used the following steps towards estimating crash risk estimation models and assess their transferability:

1. Assemble a database of archived loop detector data for four study segments (US-101 NB/SB, I-880 NB/SB), within the milepost range in the vicinity of San Jose metropolitan area for the 16-month study period (January 1, 2010 to April 30, 2011).

2. Assemble a database of observed crash data for the same duration, including information on date, time, and location of crash. The information was obtained from the Performance Measurement System (PeMS) database for the study period.

3. Create a database of "normal" conditions, so that there are 10 "normal" observations for each crash observation. The date, time, and location of these *non*-crashes were randomly chosen from the range of all possible dates, times, and locations combinations for the 16-month period identified above. These were times/locations that did not observe any crash and using these data along with the crash information the database for binary classification was setup.

4. Extract loop detector data for all crash and non-crash events, given the date, time, and milepost information from Performance Measure-

ment System (PeMS) database.

5. Perform statistical (logistic regression) and data mining (classification tree) based analysis to fit the most appropriate classification model that explains the effects of traffic flow variables on crash-risk. These variables are measured at different locations upstream and downstream of the crash, from different time durations prior to the crash, to gain an understanding of spatiotemporal impact these variables have on crash risk.

6. Select the best models estimated from the US-101 NB crash and non-crash data, and use them to score the datasets (which include both crash and non-crash observations) for US-101 SB and I-880 NB and SB.

7. Examine the classification performance of the models on these datasets (transferability) and discuss the results in the context of a real-time application.

8. Assess the robustness of the models by analyzing false positive and false negative classifications of crash/non-crash by location, time of day, and day of week.

This thesis is organized into seven chapters, including the Introduction. The next chapter provides a thorough review of relevant past research efforts, including those aimed at real-time identification of crash prone conditions. Chapter 3 presents background information about the study area, as well as the data prepara-

tion process. Chapter 4 presents the results of the logistic regression and data mining models and how well these models performed on nearby freeways. Chapter 5 discusses the conclusions from these results and other relevant issues with regards to application of these results. Chapter 6 discusses the robustness of the models and presents ideas on refinements prior to or during implementation. Chapter 7 draws conclusions on the results and suggests topics of future work in the area of proactive crash risk assessment.

II. LITERATURE REVIEW

This chapter reviews previous studies from the literature relevant to this research. The literature review is divided into two sections. The first section is a summary of traffic safety studies with real-time identification of crash prone conditions on the freeway as their objective. All of these studies are fairly recent; indicating that the idea of using loop detector data for traffic safety applications is still in its early stages. These safety studies are further categorized into two groups: a) the exploratory studies and b) studies establishing statistical links. The second section of the review is the summary of data mining applications in the areas of incident detection and crash analysis.

SAFETY APPLICATIONS OF ITS-ARCHIVED DATA

Golob, Recker, and Alvarez (2004b) categorized traffic safety related studies into two groups. First, the aggregate studies, in which units of analysis represent counts of crashes or crash rates for specific time periods (typically months or years) and locations (specific roads or networks). The traffic flow in these studies is represented by the parameters of statistical distributions of traffic (e.g., Annual Average Daily Traffic (AADT)) for similar time and location (e.g., Zhou and Sisiopiku 1997).The second group of studies consist of disaggregate analysis, in which the units of analysis are the crashes themselves and traffic flow is represented by parameters of traffic flow at the time and location of each crash.

While determination of freeway crash patterns has been the stated focus of traffic safety literature, most of the studies belong to the former group. Disaggregate

studies are relatively new, and are made possible by the recent enhancements in capabilities to collect, store and analyze real-time traffic data through intelligent transportation system (ITS) applications. In this section such previous studies are summarized and critically reviewed since this research falls in the group of disaggregate studies.

Exploratory Studies

Hughes and Council (1999) were one of the first authors to explore the relationship between freeway safety and peak period operations using loop detector data. They concluded that macroscopic measures, such as AADT and even hourly volume, in fact, correlate poorly to real time system performance. Their work mostly relied upon the data coming from a single milepost location during the peak periods of the day, on which they tried to overlay the crash time at that particular location to infer about the changes in system performance as it approaches the time of the crash. The changes in the performance were also examined from “snapshots” provided by cameras installed on the freeway.

One of their most important observations was that “design inconsistency,” that is the non-uniform application of geometric design standards, is a key factor of crash causation. Future research should consider “traffic flow consistency,” that is, the variability in traffic parameters (such as speed, volume, and occupancy) as an important variable from a human-factor standpoint. They also expressed a need for determining the exact time of the crash to avoid “cause and effect” fallacy.

Studies Establishing Statistical Links

Madanat and Liu (1995) came up with an incident likelihood prediction model using loop data as input. The focus of their research was to enhance existing incident detection algorithms with likelihood of incidents. They actually considered two types of incidents *a)* crashes and *b)* overheating vehicles. Binary logit was the methodology used for analysis. They concluded that merging section, visibility and rain are statistically the most significant factors for crash likelihood prediction.

Lee, Saccomanno, and Hellinga (2002) introduced the concept of “crash precursors” and hypothesized that the likelihood of crash occurrence is significantly affected by short-term turbulence of traffic flow. They came up with factors such as speed variation along the length of the roadway (i.e. the difference between the speeds upstream and downstream of the crash location) and also across the three lanes at the crash location. Another important factor identified by them was traffic density at the instant of the crash. Weather, road geometry and time of the day were used as external controls. With these variables, a crash prediction model was developed using log-linear analysis. According to the authors the log-linear model was chosen so that the exposure can be easily determined, which would have been difficult, if instead a logit model was used. In order to test the goodness of fit for the model, Pearson chi-square test was performed. The test measured how close the expected frequencies are to the observed frequencies for any combination of crash precursors and control factors. At 95% confidence level the model yielded a good fit.

In a subsequent study, Lee, Hellinga, and Saccomanno (2003) continued their work along the same lines and modified the aforementioned model. They incorporated an algorithm to get a better estimate of time of the crash and the length of time slice (prior to the crash), that is, duration to be examined. They concluded that variation of speed has a relatively longer term effect on crash potential than density and average speed difference between upstream and downstream ends of roadway sections. It was also observed that the average variation of speed difference across adjacent lanes doesn't have direct impact on crashes and hence was eliminated from the model.

The prediction models in both studies relied upon the log-linear models developed in the past to estimate crash frequencies on freeways using the aggregate measures of traffic flow variables. The main difference being that they determined the crash precursors included in the model in an objective manner and not based on their subjective categorization. In one of their most recent related studies, Lee, Hellinga, and Saccomanno (2004) proposed the application of these models and estimated real-time crash potential. The main focus of this study was to reduce the crash potential obtained from the model through different control strategies of variable speed limits (*VSL*). To mimic responses from the drivers to changes in speed limits, the microscopic simulation tool, *PARAMICS,* was used. At least on the simulated data the VSL showed significant safety benefits measured in terms of reduction in crash potential estimated from their model.

A later study (Gayah et al. 2006) similarly used *PARAMICS* to assess the effectiveness of various ITS strategies in mitigating crash-prone conditions on the

previously-studied Interstate-4 corridor in Orlando. The authors also concluded that VSL had significant benefits in crash reduction in high-speed conditions preceding crashes, but that such a benefit could only be achieved by ramp metering in the congested regime.

Continuing this trend of investigating advanced traffic management (ATM) strategies, Nezamuddin et al. (2011) used VISSIM to model VSL, peak-period shoulder lane use, and both strategies together. Their study assessed the effects of these strategies on speed, throughput, and safety on a section of the Missouri-Pacific Expressway in Austin, Texas. Speed harmonization and a reduction in number of stops per vehicle and vehicle conflicts were achieved with VSL; however, this came at the expense of operating speed. Shoulder use increased operating speed and decreased traffic density, but had the opposite effect of increasing speed variability and has many other safety considerations that must be addressed. Ramp metering was not addressed in this study.

Similar to the aforementioned studies, weather, environmental, and loop detector data were analyzed for association with different incident types (Songchitruksa and Balke 2006). It was found that 5-min average occupancy and coefficient of variation in speed had the strongest association with crash risk, and other factors such as visibility, time of day, and lighting condition strongly affected the type of incident that occurred.

A study by Pande, Mohamed Abdel-Aty, and Hsia (2005) utilized within-stratum one-covariate logistic regression models to determine the relative risk of crash occurrence, measured by the hazard ratio. This ratio represents the increase in

risk of crash occurrence (in log odds) by changing the covariate by one unit. The study found that the log of coefficient of variation in speed and average occupancy (expressed as percentage), and standard deviation of volume, most significantly affected the likelihood of crash occurrence. Additionally, it was determined that computing these parameters at a 5 minute time interval was more closely associated with crash risk than at 3 minute intervals. Contour plots of spatiotemporal variation of crash risk were created, and the one representing the log of the coefficient of variation in speed most clearly demonstrated increasing crash risk as the time and location of the crash were approached. The authors also proposed a methodology to identify crash-prone conditions in real time, for potential use in proactive traffic management.

Oh et al. (2001) showed that five minutes standard deviation of 30-second speed measurements was the best indicator of “disruptive” traffic flow leading to a crash as opposed to “normal” traffic flow. They used the Bayesian classifier to categorize the two possible traffic flow conditions. Since Bayesian classifier requires a probability distribution function for each competing class, the standard deviations of speed over crash and non-crash cases were used to fit non-parametric distribution functions using Kernel smoothing techniques. The potential application of the model in real-time was also demonstrated.

A more detailed analysis of patterns in crash characteristics as a function of real-time traffic flow was done by Golob and Recker (2003). The methodology used was non-linear (nonparametric) canonical correlation analysis (NLCCA) with three sets of variables. The first set comprised a seven-category segmentation

variable defining lighting and weather conditions; the second set was made up of crash characteristics (collision type, location and severity); and the third set consisted of real-time traffic flow variables. Since NLCCA requires reducing collinearity in the data, principal component analysis (PCA) was performed to identify relatively independent measurements of traffic flow conditions. The results of the PCA are shown below.

Table 1. Interpretation of Principal Components and Variable Selection

Factor	Interpretation	Represented by
1	Central tendency of speed	Median volume/occupancy interior lane
2	Central tendency of volume	Mean volume left lane
3	Temporal variation in volume—Left and interior lanes	Variation in volume for left lane
4	Temporal variation in speed—Left and interior lanes	Variation in volume/occupancy interior lane
5	Temporal variation in speed—Right lane	Variation in volume/occupancy right lane
6	Temporal variation in volume—Right lane	Variation in volume right lane

Source: Golob and Recker (2003)

It was concluded that the collision type is the best-explained characteristic and is related to the median speed, and to left-lane and interior lane variations in speed. Moreover the severity of the crash tracks the inverse of the traffic volume, and is influenced more by volume than the speed.

Based on these results, one of their later studies (Golob, Recker, and Alvarez 2004a) used data for more than 1000 crashes over six major freeways in Orange

County, California and developed a software tool *FITS* (Flow Impacts on Traffic Safety) to forecast type of crashes that are most likely to occur for the flow conditions being monitored. A case study application of this tool on a section of *SR 55* was also demonstrated.

Golob and Recker (2004) also showed that certain traffic flow regimes are more conducive to traffic crashes than the others. Of the eight traffic flow regimes found to exist on the six freeways in Orange County (California), the study found that nearly 76% of all crashes occurred in the four traffic regimes that represent flow nearing or at congestion. This displays a correlation between the types of flow and crashes and indicates that understanding the patterns in real-time traffic flow might be the key to 'predict' crashes on urban freeways. It should be noted that none of the studies by Golob et al. included non-crash loop data as a measure of 'normal' traffic conditions.

This link between traffic congestion and freeway crashes was also noted by Zhang et al. (2005) in a study that explored the relationship between crashes, weather conditions, and traffic congestion. The study showed that the relationship between the "Relative Risk Ratio" (a measure of crash probability) resembles an inverted U-shaped curve with a peak value during moderate congestion and low points at free flow and heavy congestion.

Park and Ritchie (2004) showed that the lane-changing behavior and presence of long vehicles within a freeway section has significant impact on section speed variability. The section speed variance rather than the point speed variance was used to demonstrate the traffic changes more efficiently. The traffic data for their

study were not obtained from more conventional single or dual loop detectors. Instead, a state-of-the-art vehicle-signature based traffic monitoring technology providing individual vehicle trajectories as well as accurate vehicle classification was used.

Pande & Abdel-Aty (2006) further correlated lane-changing maneuvers with both sideswipe and angle crashes on the inner lanes of a freeway. Classification trees using data collected from loop detectors on the Interstate-4 corridor identified average speed upstream and downstream of the crash location, and difference in occupancy of adjacent lanes, as having significant association with the crash/non-crash binary variable. Satisfactory classification accuracy indicated the potential for real-time application in identifying risk for lane change-related crashes.

Another study by Pande and Mohamed Abdel-Aty (2006a) analyzed rear-end crashes occurring under two flow regimes, extended congestion and near free-flow 5-10 minutes prior to a crash. It was observed that, in the first case, coefficient of variation in speed and average occupancy distinguished crash from randomly selected non-crash cases. In the second case of nearly free-flow conditions preceding a crash, average speed and occupancy at downstream of the crash location were identified as significant factors. The authors proposed a strategy for real-time identification of crash-prone conditions using neural network-based classifiers.

While almost all studies have indicated a relationship between crash occurrence and speed variability, a recent study by Kockelman and Ma (2004) found no evi-

dence to the fact that speeds or their variations trigger crashes. The study was conducted for the same area as Golob, Recker, and Alvarez (2004b). Their sample size was limited to 55 severe crashes that occurred during January 1998 and with such a small sample their conclusions remain suspect. Similarly, Ishak and Alecsandru (2005) were unable to separate pre-incident, post-incident, and non-incident traffic regimes from each other and it was indicated that conditions before a crash might not be discernible in real-time. The study was performed using part of the ITS-archived data from Interstate 4 in Orlando, Florida that was used in the research by Pande (2003). However, data for only 116 crashes were used which raises concerns about the validity of the findings from this research.

Various modeling methodologies have previously been explored by the researchers, including Probabilistic neural network (PNN) (Mohamed Abdel-Aty and Pande 2005), matched case-control Logistic Regression(Mohamed Abdel-Aty et al. 2004), split models (Mohamed Abdel-Aty, Uddin, and Pande 2005), multi-layer perceptron (MLP)/radial basis function (RBF) neural network architectures (Pande 2003) and Generalized Estimation Equation (Abdel-Aty and Abdalla 2004). The data for these studies were collected from a 13.2-mile central corridor of Interstate 4 in Orlando. All these studies made significant contributions towards enriching the literature. However, as explained later in this chapter, it must be acknowledged that there remains sufficient scope for improvement.

Critical Review

It is evident that the idea of exploring the loop data in traffic safety research is still in its preliminary stages. Some of the aforementioned studies do have a potential

application in the field of real-time proactive traffic management, but they have not fully analyzed the “recipe” of crashes. This is besides the fact that the statistical analysis in some cases isn’t really sound from a theoretical point of view.

The research conducted in Canada (Lee, Hellinga, and Saccomanno 2003) has an advantage over other research groups with dual loops placed close to each other (38 loops on a 10-km stretch of the freeway). Their analysis is based on a log-linear crash frequency model. As this is not based on classification, it cannot decipher whether or not conditions are risky in real-time. It is therefore unsuitable for real-time classification of the loop data patterns.

Golob and Recker (2003) have established sound statistical links between environmental factors, traffic flow as obtained from loop data, and crash occurrence but their findings are limited by the fact that the traffic data is obtained from single loop detectors and speed has to be estimated using a proportional variable (volume/occupancy). The *FITS* tool developed by Golob, Recker, and Alvarez (2004a) has limited application, due to a systematic pattern of missing values within the data used for development of this tool.

The classification model developed by Oh et al. (2001) seems to have the most promising online application, but due to limited crash data (only 52 crashes) their model remains far from being implemented in the field. The only factor used for classification is the 5-minute standard deviation of speed; other significant factors such as geometry, weather and other traffic flow variables were not considered. It is also to be understood that if a crash prediction model has to be useful one must classify the data much ahead of the crash occurrence time and not just 5-

minutes prior to the crash so that the Regional Transportation Management Center (*RTMC*) has some time for analysis, prediction and dissemination of the information.

The use of limited crash and traffic data is what causes concerns about the findings by Ishak and Alecsandru (2005) as well. In the study pre-incident, post-incident, and non-incident traffic flow regimes were described by 30-second average speed and its variation depicted through spatio-temporal contour charts. Using second-order statistical analyses, the charts were measured for smoothness, homogeneity, and randomness. No consistent pattern for any of the statistical measures was found within three different categories of traffic regimes (i.e., the pre-incident, post-incident, and non-incident). Therefore, it was concluded that conditions belonging to these regimes could not be differentiated from each other based on loop data. However, only 116 crashes were used in the analysis with speed and its variation as the only independent parameters. It is likely that more crash and non-crash data along with different flow parameters from a range of stations located around crash locations would have yielded better results towards separating these three distinct traffic regimes. The findings from some of the previous studies by Abdel-Aty et al. (differentiating pre-crash from non-crash) and Al-Deek et al. (separating post-incident from non-incident) that used the loop data from the same corridor make this postulation all the more plausible.

In this regard, the investigators deem that the most critical issue not addressed by past research is the issue of transferability. Since gathering data from different sources and combining them is a significant effort, it would be worthwhile to know

whether models developed from one freeway can be applied to the data from other freeways. While it may be unreasonable for models developed with data from a dense urban freeway environment to perform well on a rural freeway corridor; no studies have even tested models from the same geographical area to other freeways in close proximity. This study makes an effort in that direction.

APPLICATIONS OF DATA MINING IN TRANSPORTATION

Data mining is defined as the process of extracting valid, previously unknown and ultimately comprehensive information from large databases (Hand, Mannila, and Smyth 2001). Over the years data mining has emerged as a powerful new instrument offering value across a broad spectrum of information intensive industries involving huge amounts of data including banking, logistics, etc. The potential of various data mining techniques in the field of transportation engineering, however, remains underutilized with the exception of neural network applications for incident detection.

Of all data mining applications in transportation engineering, the “incident detection” algorithms are the most relevant to this research problem, since detecting an incident also involves classification of traffic flow patterns emanating from loop detectors. The critical distinction being that while we are interested in ‘pre-crash’ data, detection algorithms involve analysis of ‘post-incident’ loop data. In the following section data mining based incident detection algorithms are reviewed.

Incident Detection Algorithms

Cheu and Ritchie (1995) developed three types of neural network models, name-

ly multi-layer feed forward (MLF), the self-organizing feature map (SOFM) and adaptive resonance theory 2 (ART2) to classify traffic data obtained from loop detectors with the objective of using the classified output to detect lane-blocking freeway incidents.

The Artificial neural network models (ANNs) were designed to classify the input data into one of the two states, an incident or incident-free condition. ANN models were trained using post-incident loop detector data generated from INTRAS, a microscopic traffic simulation model as, according to the authors, it would have been impractical to put extensive effort in collecting real life data. INTRAS initially generated the incident and incident free input vectors in a ratio of 1:4. The incident input vectors were later replicated to make the number of state 1 and state 2 vectors equal in the training data set. The input vectors used were 16-dimensional, consisting of upstream and downstream detectors' volume and occupancy at 30-second slices after the time of the incident. Based on the performance of these networks on field evaluation data, they reported that multi-layer perceptron (MLP) neural networks always produced consistently better results than the other two networks and that these results were also better than the traditional detection algorithms.

Abdulhai and Ritchie (1999) tried to identify the requirements of a successful detection framework and found that inability to address the issues of predicted probability of incident occurrence is one of the major shortcomings of detection algorithms. They proposed the concept of statistical distance and a modified probabilistic neural network model (PNN2) in addition to Bayesian based tradi-

tional probabilistic neural network (PNN) model to detect the patterns in the loop data. They also reported that these two models were competitive with the more frequently used MLP neural networks for incident detection.

Ishak and Al-Deek (1999) conducted a study which did not use simulation data and training and testing of the neural network models for incident detection but rather real-life loop data only. In this regard some more studies by Al-Deek, Ishak, and Khan (1996) and Al-Deek, Garib, and Radwan (1998) on incident detection are remarkable. The data used by Ishak and Al-Deek (1999) were collected from the same Interstate 4 corridor for which the initial crash prediction models were developed by Pande and Abdel-Aty (2008). Input patterns of various dimensions were attempted and the network size was changed accordingly in order to achieve better performances. One of their interesting findings was that while using the MLF neural network, the incidents might be detected better with the speed patterns alone rather than using occupancy patterns or a combination of speed-occupancy patterns.

Data Mining Applications in Traffic Safety

A comparison between the fuzzy K-nearest neighbor algorithm and MLP neural network to identify crash-prone locations was made by Sayed and Abdelwahab (1998). Results showed that MLP produced slightly more accurate results and achieved higher computational efficiency than fuzzy classification.

Awad and Janson (1998) applied an MLP to model truck crashes at interchanges in Washington State. Results of the neural network were compared with a linear

regression model. Comparison was based on the root mean squared error (RMSE). The trained neural network showed a better fit when the training data is presented. However, the ability of the trained ANN to predict “unseen” test data was unsatisfactory.

Mussone, Ferrari, and Oneta (1999) adopted an MLP approach to analyze traffic crashes that occurred at intersections in Milan, Italy. Results showed that the neural network models could extract information, such as factors explaining crashes and contributing to a higher degree of danger.

Through a sequential review of literature, it was observed that the only neural network architecture explored for traffic safety analysis was the MLP until Abdelwahab and Abdel-Aty (2001) developed Fuzzy ART neural networks to predict driver injury severity in traffic crashes at signalized intersections. These models were compared with the MLP architecture and it was concluded that MLP models were superior tools compared to the ordered logit model and Fuzzy ART. In a later work by the same authors (Abdelwahab and Abdel-Aty 2002), ANN models were used for traffic safety analysis of toll plazas. Driver injury severity (no injury, possible injury, evident injury, severe injury/fatal crashes) and location of the crash (before plaza, at the plaza and after the plaza) were analyzed using MLP as well as radial basis function (RBF) neural network. They reported that for analyzing crash location the nested logit model was the best, while RBF neural network was the best model for driver injury severity analysis.

Probabilistic neural networks (PNN), an implementation of the Bayesian classifier, were explored (Pande and Abdel-Aty 2008) on the Interstate-4 corridor in Or-

lando to identify rear end crash-prone conditions. These crashes were divided into those occurring under congested and relatively free-flow conditions preceding the crash, and decision tree-based classification determined that while their frequencies are comparable, the first condition is much rarer and can hence be described as a “crash-prone” condition. PNN-based classification models were also developed for the free-flow regime.

In the recent past, data mining techniques other than neural networks have also appeared in the traffic safety literature. Vorko and Jovic (2000) used multiple attribute entropy models to classify school-age injuries. Sohn and Shin (2001) employed neural networks and decision tree algorithms to develop classification models for road traffic crash severity (bodily injury or property damage) as a function of potentially correlated categorical factors. It was noticed that classification accuracy of the individual models from both algorithms was relatively low. It was noticed that the use of data fusion or ensemble algorithms were able to increase the classification accuracy. Data fusion techniques try to combine classification results obtained from several individual classifiers and are known to improve the classification accuracy when some results of relatively uncorrelated classifiers are combined. The resulting performance is usually more stable than that of a single classifier.

A multiple model framework (Pande and Abdel-Aty 2007) was fairly recently proposed, incorporating the findings of earlier studies on rear-end and lane-change-related crashes on the Interstate-4 corridor in Orlando. The developed models satisfactorily identified both of these cases, as well as related single-vehicle

crashes. This work elaborates on a previous doctoral dissertation (Pande 2005), which was aimed at identifying the unique precursors to each crash type, and developing models which can be hybridized and applied in real time as part of a proactive traffic management strategy.

Another study was conducted by Xu et al. (2011) on a 9.2 mile stretch of the I-880 corridor in Hayward, California. Using loop detector data gathered by researchers at the University of California, Berkeley, the researchers classified traffic into 5 homogeneous flow states using K-means clustering analysis. The case-control study compared occupancy data for 1 crash case with four non-crash cases, all occurring at the same time and location between loop detectors. The authors developed four logistic regression models, indicating odds ratios 4 to 5 times higher for the "risky" scenarios of free flow upstream to a congested downstream regime and congested upstream flow to free flow downstream, and an odds ratio 2 times higher for flow in the transition region between uncongested and congested flow, when compared to the base case of free flow. The case of congested, homogeneous flow was not statistically different in crash risk than the case of free flow. The authors also developed discriminant functions using linear combinations of the lane occupancy variables; these were able to correctly categorize the type of flow with 97.2% accuracy, and can be deployed in real-time.

Researchers Pham, El Faouzi and Dumont (2011) considered not only the speed and variability in speed as explanatory variables to crash risk, but also meteorological conditions (namely precipitation). Focusing on a 10 km stretch of the A1

motorway near Bern, Switzerland between 2002 and 2007, the authors analyzed 120 rear-end and sideswipe crashes. Data was collected for 30 minutes before each crash (in five-minute intervals), as well as for non-crash cases. Principal component analysis (PCA) was used to normalize and transform traffic situations to self-organizing maps (SOMs), which partition the data points into clusters. Random Forests were then used to develop risk identification models for each of 8 defined flow regimes. 6 of the 8 performed with acceptable accuracy (70% of crash and non-crash cases correctly identified). The two that performed poorly did not have enough data to develop a good statistical model. It was found that rain had a much stronger influence in medium-flow regimes than in either congested or free-flow conditions. For most of the traffic regimes, lane speed and lane variation in speed were the most significant factors in determining crash risk.

CONCLUSIONS FROM THE LITERATURE REVIEW

An extensive review of relevant literature is conducted in this chapter. Findings demonstrate the applications, albeit limited so far, of ITS archived data and/or data mining techniques in the field of traffic safety.

The issues not addressed adequately by studies using real-time loop detector data for 'predicting' crashes, are referred to by Golob, Recker, and Alvarez (2004b) as disaggregate studies, (which was discussed in detail in section on *Safety Applications of ITS-Archived Data)*. The most significant of these issues to be addressed in this research is that of transferability. Therefore, a sufficiently large database with crash and non-crash data is assembled for this study from a subset of the major freeways/expressways in the city of San Jose. Then the

models developed from US-101 NB data are applied to other three corridors for which data are assembled. Freeway Performance Measurement System (PeMS) managed by Caltrans was the source for the archived ITS data (collected and stored on a continuous basis) as well as for the incident data. In the next chapter these data sources and the details of the four corridors are provided in the context of the present research problem.

III. STUDY AREA

This study covers four freeway segments: US-101 NB, SB and I-880 NB, SB in the San Jose area of Santa Clara County, California. These freeway corridors in the city of San Jose run through dense urban development, and are among the busiest in the South Bay Area. The logistic regression and data mining models are estimated using the US-101 NB data and then these models are applied on the three segments; US-101 SB, I-880 NB, and I-880 SB to evaluate transferability of the models. This chapter provides details of these segments along with details of data collection and preparation.

FREEWAY CORRIDORS

US-101 Freeway

US-101 (also known as the “Bayshore Freeway”) is the primary north-south corridor through the City of San Jose. The route runs through southern Santa Clara County as a 6-lane freeway through the suburbs of Gilroy and Morgan Hill. North of Morgan Hill, US-101 gains an HOV (High Occupancy Vehicle) lane in each direction (expanding to an 8-lane freeway) through the rural area known as Coyote. The freeway wanders in and out of San Jose city limits and unincorporated land for approximately 8 miles. At the junction of State Route 85, US-101 enters the area conventionally accepted as the boundary of the city of San Jose. The route continues as an 8-lane freeway through the junctions of SR-82, I-280/I-680, I-880, and SR-87, then entering the City of Santa Clara. The route continues through the South Bay cities of Sunnyvale, Mountain View, and Palo Alto, finally

running up the Peninsula through San Mateo County to San Francisco.

The study segment of interest for US-101 northbound is 17.1 miles long, starting at milepost 375.31 and ending at milepost 392.37.The study segment for US-101 southbound starts at milepost 392.45 and ends at milepost 375.81, for a total length of 16.6 miles. See Figure 1 and Figure 2 below for schematic diagrams for location of the VDS (vehicle detector stations) along these routes. In the diagrams, VDS ID numbers are truncated to the last four digits, and superimposed on the route.

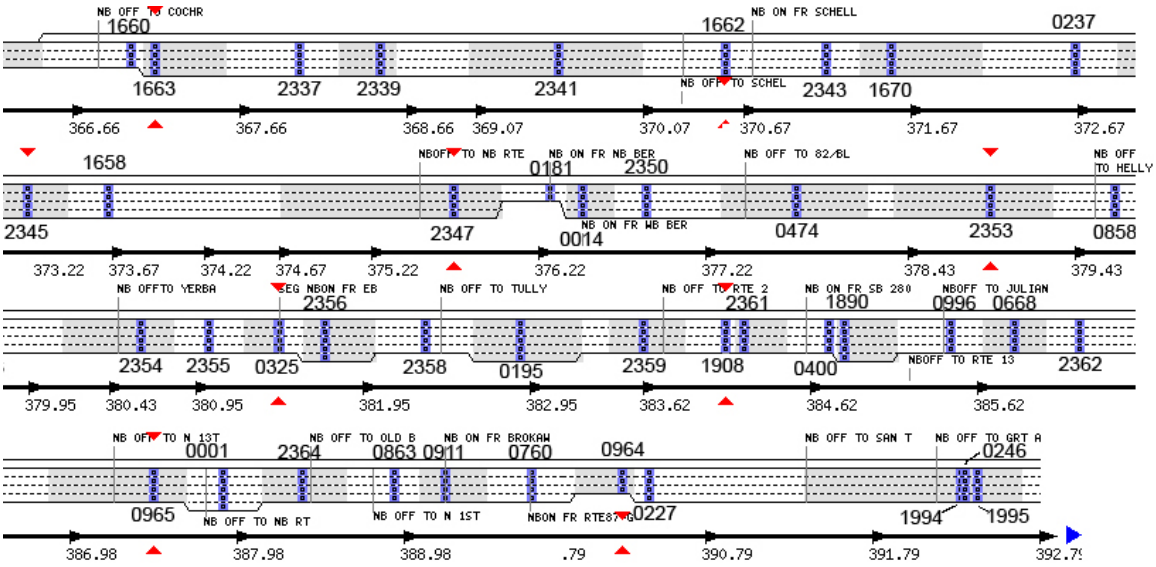

Figure 1. US-101 NB Corridor and VDS Locations

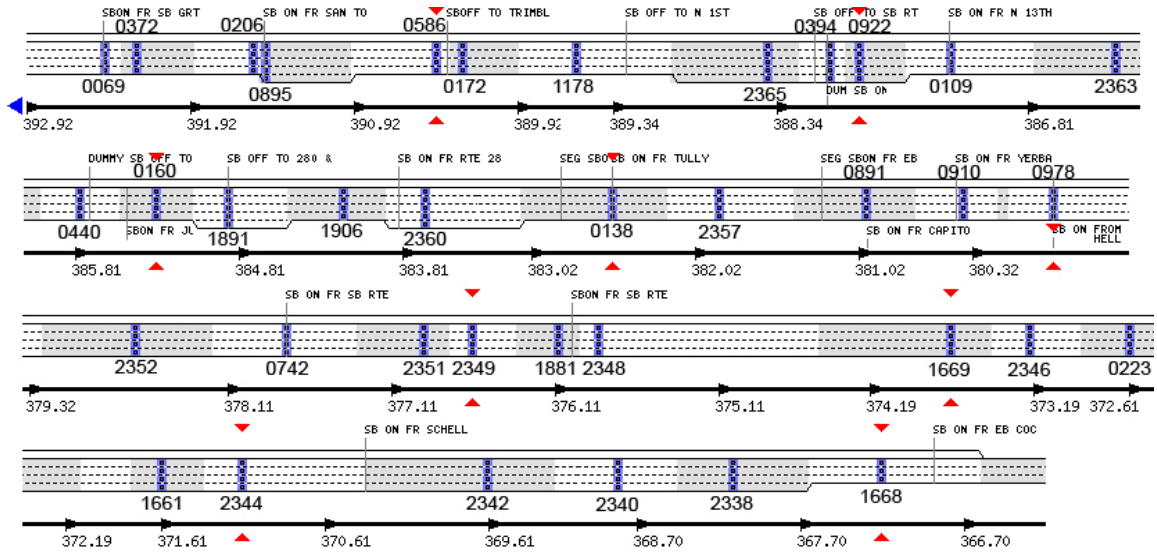

Figure 2. US-101 SB Corridor and VDS Locations

I-880 Freeway

Interstate 880 (also known as the “Nimitz Freeway”) is a 6-lane freeway with no dedicated HOV lanes. Its officially designated beginning is located north of the I-

280 junction. The freeway extends south of this interchange as State Route 17, a freeway running between Santa Cruz, CA and San Jose, CA. Interstate 880 runs north through the city of San Jose for approximately 7 miles, connecting to SR-82, crossing over the SR-87 freeway (with no interchange) and connecting to the US-101 freeway. I-880 next enters the City of Milpitas, and finally crosses the Alameda County line, running up the East Bay to Oakland. An improvement project has been underway since 2010 to reconfigure the I-280/I-880 interchange. The goal is to provide a dedicated NB I-280 to NB I-880 ramp; the connection is currently shared with the busy Stevens Creek Boulevard interchange, causing merging and weaving issues.

The study segment of interest for I-880 NB is 8.1 miles long, starting at milepost 0.13 and ending at milepost 8.27.The study segment for I-880 SB starts at milepost 9.01 and ends at milepost 0.9, for a total length of 8.1 miles. See Figure 3 and Figure 4 below for schematic diagrams of the routes along with VDS locations.

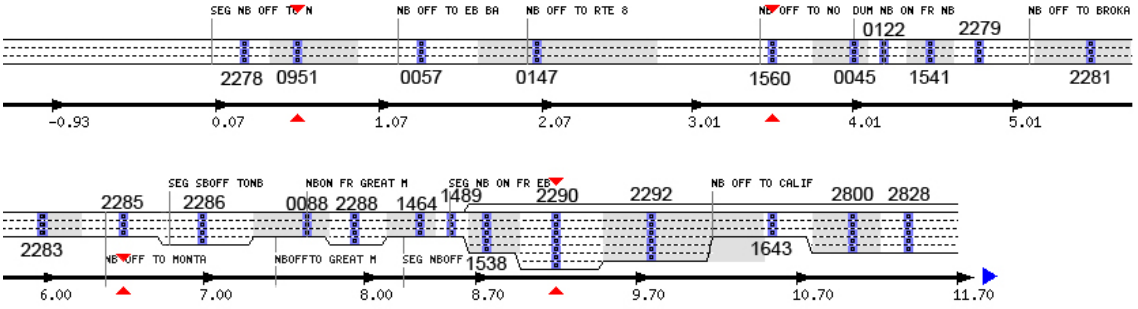

Figure 3. I-880 NB Corridor and VDS Locations

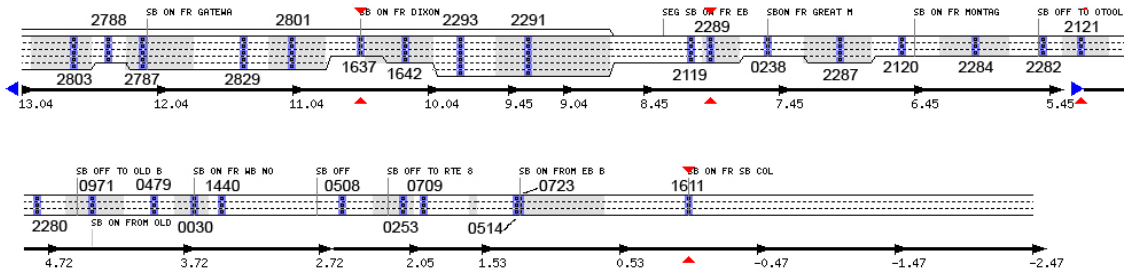

Figure 4. I-880 SB Corridor and VDS Locations

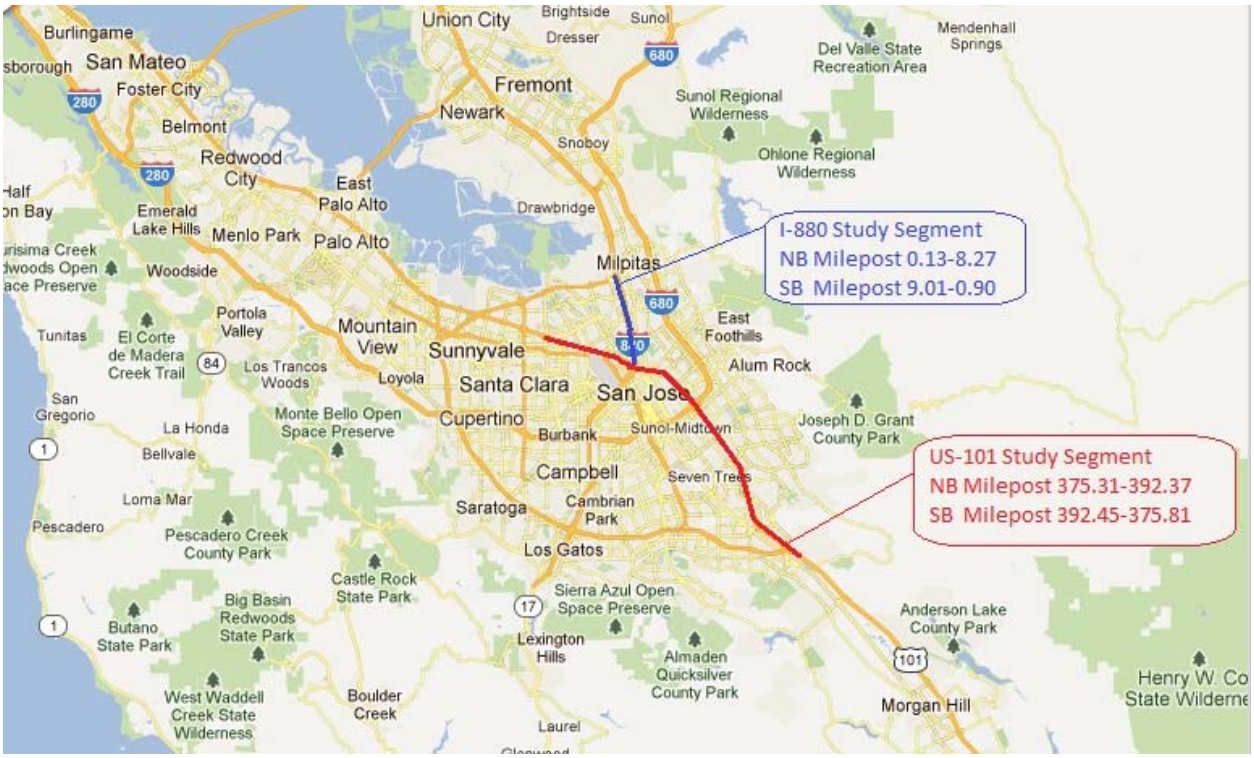

Figure 5. Study Location

DATA COLLECTION AND PREPARATION

Crash Data

This study considers crashes that occurred during a 16-month period from January 2010 through and including April 2011. These days were chosen due to a fairly recent installation of new loop detectors on US-101 in 2009. Crash data was downloaded from the "CHP Incidents" section of Caltrans' Freeway Performance Measurement System (PeMS) database. See Figure 6 for a sample of the downloaded data. Important variables for our analysis contained therein included the incident's unique ID number, time of occurrence, and milepost.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	#	District	Area	Fwy	Start	Duration	Abs Postmile	CA Postmi	Location	Description				
2	472	4	SAN JOSE	SR101-N	1/1/2010 2:45	126	379.56	29.931	NB HELLYER AV ONR TO NB US101	1183 - Traffic Collision - No Details				
3	971	4	SAN JOSE	SR101-N	1/1/2010 9:27	150	383.55	34.113	NB US101 JSO NB I680	1183 - Traffic Collision - No Details				
4	980	4	SAN JOSE	SR101-N	1/1/2010 9:35	0	382.42	32.784	NB US101 AT TULLY RD	1183 - Traffic Collision - No Details				
5	304	4	SAN JOSE	SR101-N	1/5/2010 7:13	40	387.04	37.603	NB US101 JSO N 13TH ST	1183 - Traffic Collision - No Details				
6	419	4	SAN JOSE	SR101-N	1/5/2010 8:05	4	385.22	R35.782	NB US101 JSO E JULIAN ST	1182 - Traffic Collision - Property Damage				
7	1218	4	SAN JOSE	SR101-N	1/5/2010 14:49	53	387.04	37.603	NB US101 JSO N 13TH ST	20002 - Hit and Run - No Injuries				
8	1951	4	SAN JOSE	SR101-N	1/6/2010 19:08	12	391.5	41.862	NB US101 AT MONTAGUE EXWY	1182 - Traffic Collision - Property Damage				
9	1900	4	REDWOOI	SR101-N	1/7/2010 18:33	5	375.31	R26.237	NB US101 JSO SR85	1183 - Traffic Collision - No Details				
10	295	4	HOLLISTEF	SR101-N	1/8/2010 13:38	0	376.32	R26.85	NB US101 JNO BERNAL RD	1183 - Traffic Collision - No Details				
11	1106	4	SAN JOSE	SR101-N	1/8/2010 13:42	11	376.32	R26.85	NB US101 JNO BERNAL RD	1183 - Traffic Collision - No Details				
12	355	4	GOLDEN G	SR101-N	1/9/2010 5:00	0	377.26	R28.185	NB US101 JSO BLOSSOM HILL RD	1183 - Traffic Collision - No Details				
13	362	4	SAN JOSE	SR101-N	1/9/2010 5:13	2	379.36	29.931	NB US101 JSO HELLYER AV	1183 - Traffic Collision - No Details				
14	1722	4	SAN JOSE	SR101-N	1/9/2010 18:16	24	387.6	38.164	NB US101 JSO NB I880	20002 - Hit and Run - No Injuries				
15	1950	4	REDWOOI	SR101-N	1/11/2010 18:38	0	375.31	R26.237	NB US101 JSO SR85	1183 - Traffic Collision - No Details				
16	139	4	SAN JOSE	SR101-N	1/12/2010 3:57	29	387.6	38.164	NB US101 JSO NB I880	1182 - Traffic Collision - Property Damage				
17	2026	4	SAN JOSE	SR101-N	1/12/2010 20:04	1	383.75	34.113	NB US101 TO NB I680 CON	1183 - Traffic Collision - No Details				
18	2045	4	SAN JOSE	SR101-N	1/12/2010 20:16	0	383.75	34.113	NB US101 TO NB I680 CON	1183 - Traffic Collision - No Details				
19	2059	4	SAN JOSE	SR101-N	1/12/2010 20:27	27	382.22	32.784	NB US101 JSO TULLY RD	1182 - Traffic Collision - Property Damage				
20	2062	4	SAN JOSE	SR101-N	1/12/2010 20:29	1	380.77	31.133	NB US101 AT CAPITOL EXWY	1183 - Traffic Collision - No Details				
21	2241	4	SAN JOSE	SR101-N	1/12/2010 22:38	57	382.42	32.784	NB US101 ON EB TULLY RD OFR	1179 - Traffic Collision - Ambulance Responding				
22	2256	4	SAN JOSE	SR101-N	1/12/2010 22:52	66	380.77	31.133	NB US101 ON NB CAPITOL EXWY C	1183 - Traffic Collision - No Details				
23	2260	4	SAN JOSE	SR101-N	1/12/2010 22:55	45	382.42	32.784	NB US101 AT TULLY RD	1183 - Traffic Collision - No Details				
24	2264	4	SAN JOSE	SR101-N	1/12/2010 22:58	0	382.42	32.784	NB US101 AT TULLY RD	1179 - Traffic Collision - Ambulance Responding				
25	2310	4	SAN JOSE	SR101-N	1/12/2010 23:28	0	380.77	31.133	NB US101 AT CAPITOL EXWY	1183 - Traffic Collision - No Details				
26	2319	4	SAN JOSE	SR101-N	1/12/2010 23:37	0	382.42	32.784	NB US101 ON EB TULLY RD OFR	1182 - Traffic Collision - Property Damage				
27	464	4	SAN JOSE	SR101-N	1/13/2010 8:19	13	392.17	42.533	NB US101 AT GREAT AMERICA PKV	1182 - Traffic Collision - Property Damage				
28	651	4	SAN JOSE	SR101-N	1/13/2010 9:33	67	390.43	40.591	NB US101 JNO DE LA CRUZ BLVD	1183 - Traffic Collision - No Details				
29	1622	4	GOLDEN G	SR101-N	1/13/2010 16:50	0	392.17	42.533	GREAT AMERICA PKWY AT NB US1	1183 - Traffic Collision - No Details				
30	2239	4	SAN JOSE	SR101-N	1/14/2010 19:30	0	380.77	31.133	NB US101 AT CAPITOL EXWY	20002 - Hit and Run - No Injuries				
31	648	4	SAN JOSE	SR101-N	1/15/2010 9:35	12	380.77	31.133	NB CAPITOL EXWY ONR TO NB US:	1183 - Traffic Collision - No Details				
32	796	4	GOLDEN G	SR101-N	1/16/2010 11:12	1	377.26	R28.185	NB US101 JSO BLOSSOM HILL RD	1179 - Traffic Collision - Ambulance Responding				
33	1237	4	SAN JOSE	SR101-N	1/16/2010 14:36	0	390.43	40.591	NB US101 JNO DE LA CRUZ BLVD	1183 - Traffic Collision - No Details				
34	385	4	SAN JOSE	SR101-N	1/17/2010 4:52	28	377.26	R28.185	NB US101 JSO BLOSSOM HILL RD	1183 - Traffic Collision - No Details				
35	819	4	SAN JOSE	SR101-N	1/17/2010 12:19	34	383.86	34.224	NB STORY RD ONR TO NB US101	1179 - Traffic Collision - Ambulance Responding				
36	835	4	SAN JOSE	SR101-N	1/17/2010 12:36	59	383.55	34.113	NB US101 JSO NB I680	1183 - Traffic Collision - No Details				
37	846	4	SAN JOSE	SR101-N	1/17/2010 12:41	0	382.42	32.784	NB US101 AT TULLY RD	1183 - Traffic Collision - No Details				
38	876	4	SAN JOSE	SR101-N	1/17/2010 12:55	9	385.62	R35.782	NB US101 JNO MCKEE RD	1179 - Traffic Collision - Ambulance Responding				
39	889	4	SAN JOSE	SR101-N	1/17/2010 13:03	0	383.66	34.224	NB US101 JSO STORY RD	1183 - Traffic Collision - No Details				

Figure 6. CHP Accident Data from PeMS

The predictive models were developed from the crash data from US-101 NB. There were 2176 crashes during the study period, the breakdown of which is shown below in Table 2.

Table 2. Crash Breakdown for US-101 NB

Crash Type	Frequency	Percentage
1181 - Traffic Collision - Minor Injuries	38	1.7%
1182 - Traffic Collision - Property Damage	754	34.7%
1179 - Traffic Collision - Ambulance Responding	257	11.8%
1144 - Possible Fatality	2	0.1%
20002 - Hit and Run - No Injuries	182	8.4%
20001 - Hit and Run - Injuries or Fatalities	5	0.2%
1183 - Traffic Collision - No Details	938	43.1%
Total	2176	100.0%

Traffic Information

Once the crash data were obtained based on the study area milepost boundaries of four freeway corridors, a list of all VDS locations on the study segments was compiled along with their respective mileposts. A sample list is shown below in Figure 7. The variables of interest for this study include the VDS number and milepost.

	A	B	C	D	E	F	G	H	I	J	K	L
1	Fwy	Dist	County	City	CA PM	Abs PM	VDS	ID	Lane	Type	HOV	LDS
3	SR101-S	4	Santa Clara		16.4	365.71	402336	412033	1	Mainline	No	403810
4					16.4	365.71	402336	412034	2	Mainline	No	403810
5					16.4	365.71	402336	412035	3	Mainline	No	403810
6					16.4	365.71	402336	412036	4	Mainline	No	403810
7				Morgan Hill	16.7	366.01	401579	409582	1	Mainline	No	403156
8					16.7	366.01	401579	409583	2	Mainline	No	403156
9					16.7	366.01	401579	409584	3	Mainline	No	403156
10					*SB ON FR EB COCHRAN RD (R17.576)*							
11					*SB ON FR WB COCHRAN RD (R17.857)*							
12	SR101-S	4	Santa Clara	Morgan Hill	17.89	367.2	401668	409824	1	Mainline	No	403286
13					17.89	367.2	401668	409825	2	Mainline	No	403286
14					17.89	367.2	401668	409826	3	Mainline	No	403286
15					*SB OFF TO COCHRAN RD (R18.145)*							
16	SR101-S	4	Santa Clara		18.8	368.11	402338	412041	1	Mainline	No	403812
17					18.8	368.11	402338	412042	2	Mainline	No	403812
18					18.8	368.11	402338	412043	3	Mainline	No	403812
19					18.8	368.11	402338	412044	4	Mainline	No	403812
20				San Jose	19.5	368.81	402340	412049	1	Mainline	No	403814
21					19.5	368.81	402340	412050	2	Mainline	No	403814
22					19.5	368.81	402340	412051	3	Mainline	No	403814
23					19.5	368.81	402340	412052	4	Mainline	No	403814
24					20.3	369.61	402342	412057	1	Mainline	No	403816
25					20.3	369.61	402342	412058	2	Mainline	No	403816
26					20.3	369.61	402342	412059	3	Mainline	No	403816
27					20.3	369.61	402342	412060	4	Mainline	No	403816
28					*SB ON FR SCHELLER AVE (R21.05)*							
29					*SB OFF TO SCHELLER AVE (R21.51)*							
30	SR101-S	4	Santa Clara	San Jose	21.8	371.11	402344	412065	1	Mainline	No	403818
31					21.8	371.11	402344	412066	2	Mainline	No	403818
32					21.8	371.11	402344	412067	3	Mainline	No	403818
33					21.8	371.11	402344	412068	4	Mainline	No	403818
34					22.29	371.6	401661	409802	1	Mainline	No	403279
35					22.29	371.6	401661	409803	2	Mainline	No	403279
36					22.29	371.6	401661	409804	3	Mainline	No	403279
37					22.29	371.6	401661	409805	4	Mainline	No	403279
38					23.29	372.6	400223	402377	1	Mainline	No	402861
39					23.29	372.6	400223	402378	2	Mainline	No	402861
40					23.29	372.6	400223	402379	3	Mainline	No	402861
41					23.29	372.6	400223	402380	4	Mainline	No	402861

Figure 7. VDS List by Milepost

Traffic data from these VDS locations were downloaded from the "Data Clearinghouse" section of PeMS for the entirety of Caltrans District 4 (Bay Area) for the 16-month study period. The downloaded data included the following variables for each VDS: time and date, milepost and average speed, volume, and lane-occupancy information measured every 30 seconds by corresponding VDS. It is worth mentioning that among these variables only volume and lane-occupancy are measured variables and the 30-second average speed is calculated (in the database) using these two measurements. Refer to Figure8 for a sample of the

downloaded raw loop detector data.

The next step in the data collection process was to match the traffic data to the corresponding crash events. The crash time and locations were known from the crash database (see sample in Figure 6) as described above. Each crash event was merged with corresponding traffic data from six VDS locations. These six locations included three nearest VDS to the location of crash in the upstream direction and three in the downstream direction. The spatial arrangement of locations is shown later in this chapter (See Figure 11). VDS stations were typically spaced between 0.5 and 0.8 miles apart. The time horizon for each event was the period up to 20 minutes before the crash and five minutes after the crash time. The period of 0-5 minutes after the crash was only used to verify the incident's occurrence (and is typically only relevant for incident detection); it will therefore not be discussed further in this thesis.

	VAR1	VAR2	VAR3	VAR4	VAR5	VAR6	VAR7	VAR8	VAR9	VAR10	VAR11	VAR12	VAR13
1	01JAN10:00:00:18	400223	1	0.0078	78	1	0.0089	86	1	0.0078	78	0	0
2	01JAN10:00:00:18	400237	2	0.01	86	2	0.0117	78	3	0.0206	71	3	0.045
3	01JAN10:00:00:18	401178	0	0	149	2	0.01	149	4	0.01	149	2	0.01
4	01JAN10:00:00:18	401464	1	0.0044	86	2	0.0139	65	3	0.0106	49	.	.
5	01JAN10:00:00:18	401489	1	0.0067	81	1	0.005	96	0	0	0	.	.
6	01JAN10:00:00:18	401538	0	0	0	0	0	0	0	0	0	2	0.0133
7	01JAN10:00:00:18	401642	2	0.0122	78	2	0.0139	71	3	0.0189	71	2	0.0111
8	01JAN10:00:00:18	401658	1	0.0078	71	1	0.005	86	3	0.0139	65	0	0
9	01JAN10:00:00:18	401660	1	0.0094	78	1	0.0078	78	0	0	0	.	.
10	01JAN10:00:00:18	401661	2	0.0172	78	1	0.0078	78	0	0	0	1	0.0117
11	01JAN10:00:00:18	401662	0	0	0	0	0	0	1	0.0094	78	1	0.0106
12	01JAN10:00:00:18	401663	0	0	0	0	0	0	2	0.0167	78	0	0
13	01JAN10:00:00:18	401668	1	0.0061	86	4	0.0283	78	1	0.0078	86	.	.
14	01JAN10:00:00:18	401669	0	0	0	1	0.0078	71	2	0.0122	78	0	0
15	01JAN10:00:00:18	401670	0	0	0	0	0	0	1	0.0056	97	0	0
16	01JAN10:00:00:18	401906	0	0	0	0	0	0	1	0.0072	78	1	0.0078
17	01JAN10:00:00:18	401907	0	0	0	4	0.0311	65	0	0	0	.	.
18	01JAN10:00:00:18	401908	3	0.0189	71	1	0.0056	65	3	0.0194	84	2	0.0133
19	01JAN10:00:00:18	401909	0	0	0	0	0	0	0	0	0	.	.
20	01JAN10:00:00:18	401995	0	0	0	1	0.0044	86	1	0.0044	86	0	0
21	01JAN10:00:00:18	402119	0	0	0	0	0	0	0	0	0	0	0
22	01JAN10:00:00:18	402120	0	0	0	2	0.015	96	1	0.0128	109	.	.
23	01JAN10:00:00:18	402121	2	0.0128	78	4	0.0194	71	0	0	0	.	.
24	01JAN10:00:00:21	400001	3	0.0222	78	0	0	0	1	0.0067	71	1	0.0078
25	01JAN10:00:00:21	400014	0	0	0	1	0.0061	71	1	0.0067	9	1	0.0072
26	01JAN10:00:00:21	400030	0	0	0	0	0	0	1	0.0089	65	.	.
27	01JAN10:00:00:21	400045	0	0	0	5	0.0367	60	0	0	0	.	.
28	01JAN10:00:00:21	400057	0	0	0	0	0	0	0	0	0	.	.
29	01JAN10:00:00:21	400069	0	0	0	2	0.0161	67	3	0.0128	71	1	0.0056
30	01JAN10:00:00:21	400088	0	0	0	0	0	0	0	0	0	.	.
31	01JAN10:00:00:21	400109	0	0	0	1	0.0083	65	3	0.0233	65	3	0.0122
32	01JAN10:00:00:21	400138	4	0.1539	14	0	0	0	3	0.0211	65	6	0.0233

Figure 8. Raw Data from VDS, obtained through PeMS

Non-crash Events

Since the modeling approach adopted here was binary classification we also collected traffic data for non-crash cases. The traffic data corresponding to the 'non-crash' cases would be representative of the 'normal' conditions on the freeways as opposed to the traffic data corresponding to the crash cases (described in the previous section) which represent crash prone conditions. To represent 'normal' traffic conditions for the freeway we generated a sample of random traffic conditions. As the crashes occurred both on and off-peak, both on and off-peak non-crashes were generated to sample overall traffic conditions. To generate random non-crash sample, the total study period was divided into one minute periods from which a random sample of times could be selected as the "time of non-crash" event. Similarly milepost location for non-crash cases could also be drawn from any milepost from the beginning to the end of the corresponding corridor. From all possible combinations of date-time and mileposts a sample of non-crash cases were derived. To adequately represent 'normal' conditions for every crash event used in the analysis there were 10 "non-crash" events. A previous study tested different ratios of crash to non-crash events and found; it was found that the number of non-crashes included had no effect on the classification accuracy of the model (Pande, Mohamed Abdel-Aty, and Hsia 2005). A snapshot of the process generating the random non-crash sample can be seen in Figure 9. One may observe that the function "randbetween" from excel is used in the process.

A2 =RANDBETWEEN(37531,39237)

	A	B	C	D	E	F	G	H	I	J	K	L	M
1											Date	Time	Milepost
2	38470.00	384.7	323	11/20/2010	1/1/2010	994	0.690278	0:00	16:34		11/20/2010	16:34	384.7
3	39016.00	390.16	271	9/29/2010	1/1/2010	313	0.217361	0:00	5:13		9/29/2010	5:13	390.16
4	38664.00	386.64	230	8/19/2010	1/1/2010	1020	0.708333	0:00	17:00		8/19/2010	17:00	386.64
5	38818.00	388.18	114	4/25/2010	1/1/2010	625	0.434028	0:00	10:25		4/25/2010	10:25	388.18
6	38858.00	388.58	225	8/14/2010	1/1/2010	25	0.017361	0:00	0:25		8/14/2010	0:25	388.58
7	39229.00	392.29	212	8/1/2010	1/1/2010	1271	0.882639	0:00	21:11		8/1/2010	21:11	392.29
8	38755.00	387.55	23	1/24/2010	1/1/2010	671	0.465972	0:00	11:11		1/24/2010	11:11	387.55
9	38421.00	384.21	47	2/17/2010	1/1/2010	1087	0.754861	0:00	18:07		2/17/2010	18:07	384.21
10	38584.00	385.84	481	4/27/2011	1/1/2010	299	0.207639	0:00	4:59		4/27/2011	4:59	385.84
11	38024.00	380.24	296	10/24/2010	1/1/2010	253	0.175694	0:00	4:13		10/24/2010	4:13	380.24
12	38015.00	380.15	81	3/23/2010	1/1/2010	1249	0.867361	0:00	20:49		3/23/2010	20:49	380.15
13	39009.00	390.09	109	4/20/2010	1/1/2010	747	0.51875	0:00	12:27		4/20/2010	12:27	390.09
14	38375.00	383.75	198	7/18/2010	1/1/2010	937	0.650694	0:00	15:37		7/18/2010	15:37	383.75
15	38730.00	387.3	419	2/24/2011	1/1/2010	1006	0.698611	0:00	16:46		2/24/2011	16:46	387.3
16	37541.00	375.41	472	4/18/2011	1/1/2010	1383	0.960417	0:00	23:03		4/18/2011	23:03	375.41
17	38552.00	385.52	456	4/2/2011	1/1/2010	12	0.008333	0:00	0:12		4/2/2011	0:12	385.52
18	38143.00	381.43	196	7/16/2010	1/1/2010	318	0.220833	0:00	5:18		7/16/2010	5:18	381.43
19	37941.00	379.41	229	8/18/2010	1/1/2010	1335	0.927083	0:00	22:15		8/18/2010	22:15	379.41
20	38311.00	383.11	60	3/2/2010	1/1/2010	1096	0.761111	0:00	18:16		3/2/2010	18:16	383.11
21	37756.00	377.56	306	11/3/2010	1/1/2010	442	0.306944	0:00	7:22		11/3/2010	7:22	377.56
22	38390.00	383.9	450	3/27/2011	1/1/2010	615	0.427083	0:00	10:15		3/27/2011	10:15	383.9
23	37715.00	377.15	182	7/2/2010	1/1/2010	887	0.615972	0:00	14:47		7/2/2010	14:47	377.15
24	38555.00	385.55	94	4/5/2010	1/1/2010	42	0.029167	0:00	0:42		4/5/2010	0:42	385.55
25	39043.00	390.43	300	10/28/2010	1/1/2010	369	0.25625	0:00	6:09		10/28/2010	6:09	390.43
26	37972.00	379.72	197	7/17/2010	1/1/2010	320	0.222222	0:00	5:20		7/17/2010	5:20	379.72
27	38286.00	382.86	397	2/2/2011	1/1/2010	484	0.336111	0:00	8:04		2/2/2011	8:04	382.86
28	38468.00	384.68	23	1/24/2010	1/1/2010	1116	0.775	0:00	18:36		1/24/2010	18:36	384.68
29	38856.00	388.56	453	3/30/2011	1/1/2010	383	0.265972	0:00	6:23		3/30/2011	6:23	388.56
30	38859.00	388.59	128	5/9/2010	1/1/2010	13	0.009028	0:00	0:13		5/9/2010	0:13	388.59
31	39064.00	390.64	410	2/15/2011	1/1/2010	535	0.371528	0:00	8:55		2/15/2011	8:55	390.64
32	38104.00	381.04	206	7/26/2010	1/1/2010	852	0.591667	0:00	14:12		7/26/2010	14:12	381.04

Figure 9. Random Generation of "Non-Crash" Events

The nearest three VDS in both upstream and downstream directions of the event location milepost were also identified for all of these non-crash events. Time horizon (from 20 minutes before the crash up to 5 minutes after the crash) was also the same as the crash events. See Figure 10 below for a sample spread-sheet of this identification process. For any crash, the station arrangement con-vention is depicted in Figure 11.

	A	B	D	H	I	J	K	BK	BL	BM	BN	BO	BP	BQ	BR	BS	BT	BU	BV
1				VDS ID:	401578	401586	401660	US3 PM	US2 PM	US1 PM	DS1 PM	DS2 PM	DS3 PM	US3 ID	US2 ID	US1 ID	DS1 ID	DS2 ID	DS3 ID
2	Date	Time	Abs Postmi	VDS PM:	365.57	365.97	367.01												
3	2/8/2011	8:26	378.27		-12.7	-12.3	-11.26	376.49	376.87	377.77	378.93	379.68	380.63	400014	402350	400474	402353	400858	402354
4	1/27/2010	13:12	383.23		-17.66	-17.26	-16.22	381.73	382.33	382.9	383.63	384.12	384.23	402356	402358	400195	402359	401908	402361
5	2/1/2011	10:15	388.61		-23.04	-22.64	-21.6	387.48	387.9	388.37	388.92	389.23	389.74	400965	400001	402364	400863	400911	400760
6	3/31/2010	5:57	387.66		-22.09	-21.69	-20.65	385.85	386.24	387.48	387.9	388.37	388.92	400668	402362	400965	400001	402364	400863
7	4/7/2011	14:40	378.16		-12.59	-12.19	-11.15	376.49	376.87	377.77	378.93	379.68	380.63	400014	402350	400474	402353	400858	402354
8	3/23/2010	8:37	378.75		-13.18	-12.78	-11.74	376.49	376.87	377.77	378.93	379.68	380.63	400014	402350	400474	402353	400858	402354
9	1/22/2010	18:43	386.73		-21.16	-20.76	-19.72	385.47	385.85	386.24	387.48	387.9	388.37	400996	400668	402362	400965	400001	402364
10	4/4/2011	7:19	387.92		-22.35	-21.95	-20.91	386.24	387.48	387.9	388.37	388.92	389.23	402362	400965	400001	402364	400863	400911
11	2/13/2010	3:30	378.76		-13.19	-12.79	-11.75	376.49	376.87	377.77	378.93	379.68	380.63	400014	402350	400474	402353	400858	402354
12	12/26/2010	8:33	388.31		-22.74	-22.34	-21.3	386.24	387.48	387.9	388.37	388.92	389.23	402362	400965	400001	402364	400863	400911
13	2/26/2011	22:57	387.83		-22.26	-21.86	-20.82	385.85	386.24	387.48	387.9	388.37	388.92	400668	402362	400965	400001	402364	400863
14	1/24/2010	3:03	379.47		-13.9	-13.5	-12.46	376.87	377.77	378.93	379.68	380.63	381.03	402350	400474	402353	400858	402354	402355
15	1/12/2010	16:59	379.52		-13.95	-13.55	-12.51	376.87	377.77	378.93	379.68	380.63	381.03	402350	400474	402353	400858	402354	402355
16	7/11/2010	13:25	388.96		-23.39	-22.99	-21.95	387.9	388.37	388.92	389.23	389.74	390.29	400001	402364	400863	400911	400760	400964
17	4/3/2010	6:00	391.47		-25.9	-25.5	-24.46	389.74	390.29	390.45	392.32	392.34	392.42	400760	400964	400227	401994	400246	401995
18	3/26/2010	6:01	391.7		-26.13	-25.73	-24.69	389.74	390.29	390.45	392.32	392.34	392.42	400760	400964	400227	401994	400246	401995
19	3/6/2011	20:24	381.92		-16.35	-15.95	-14.91	381.45	381.46	381.73	382.33	382.9	383.63	400325	400591	402356	402358	400195	402359
20	10/22/2010	14:29	390.55		-24.98	-24.58	-23.54	389.74	390.29	390.45	392.32	392.34	392.42	400760	400964	400227	401994	400246	401995
21	1/16/2010	9:48	381.95		-16.38	-15.98	-14.94	381.45	381.46	381.73	382.33	382.9	383.63	400325	400591	402356	402358	400195	402359
22	9/16/2010	10:33	387.17		-21.6	-21.2	-20.16	385.47	385.85	386.24	387.48	387.9	388.37	400996	400668	402362	400965	400001	402364
23	1/8/2010	13:11	384.48		-18.91	-18.51	-17.47	383.63	384.12	384.23	384.74	384.83	385.47	402359	401908	402361	400400	401890	400996
24	11/28/2010	14:32	389.34		-23.77	-23.37	-22.33	388.37	388.92	389.23	389.74	390.29	390.45	402364	400863	400911	400760	400964	400227
25	2/1/2011	11:00	385.29		-19.72	-19.32	-18.28	384.23	384.74	384.83	385.47	385.85	386.24	402361	400400	401890	400996	400668	402362
26	2/25/2010	22:35	380.49		-14.92	-14.52	-13.48	377.77	378.93	379.68	380.63	381.03	381.45	400474	402353	400858	402354	402355	400325
27	8/20/2010	15:02	389.41		-23.84	-23.44	-22.4	388.37	388.92	389.23	389.74	390.29	390.45	402364	400863	400911	400760	400964	400227
28	3/13/2010	19:03	386.95		-21.38	-20.98	-19.94	385.47	385.85	386.24	387.48	387.9	388.37	400996	400668	402362	400965	400001	402364
29	10/19/2010	17:38	376.01		-10.44	-10.04	-9	373.17	373.65	375.72	376.29	376.49	376.87	402345	401658	402347	400181	400014	402350
30	4/17/2011	13:27	388.81		-23.24	-22.84	-21.8	387.48	387.9	388.37	388.92	389.23	389.74	400965	400001	402364	400863	400911	400760
31	3/29/2011	10:53	386.68		-21.11	-20.71	-19.67	385.47	385.85	386.24	387.48	387.9	388.37	400996	400668	402362	400965	400001	402364
32	7/16/2010	7:29	388.31		-22.74	-22.34	-21.3	386.24	387.48	387.9	388.37	388.92	389.23	402362	400965	400001	402364	400863	400911
33	8/19/2010	0:10	391.19		-25.62	-25.22	-24.18	389.74	390.29	390.45	392.32	392.34	392.42	400760	400964	400227	401994	400246	401995
34	11/29/2010	1:38	385.57		-20	-19.6	-18.56	384.74	384.83	385.47	385.85	386.24	387.48	400400	401890	400996	400668	402362	400965
35	2/26/2010	1:14	391.7		-26.13	-25.73	-24.69	389.74	390.29	390.45	392.32	392.34	392.42	400760	400964	400227	401994	400246	401995
36	12/2/2010	22:23	379.25		-13.68	-13.28	-12.24	376.87	377.77	378.93	379.68	380.63	381.03	402350	400474	402353	400858	402354	402355
37	2/25/2011	1:27	385.37		-19.8	-19.4	-18.36	384.23	384.74	384.83	385.47	385.85	386.24	402361	400400	401890	400996	400668	402362
38	3/23/2011	18:43	386.98		-21.41	-21.01	-19.97	385.47	385.85	386.24	387.48	387.9	388.37	400996	400668	402362	400965	400001	402364
39	9/7/2010	4:46	385.15		-19.58	-19.18	-18.14	384.23	384.74	384.83	385.47	385.85	386.24	402361	400400	401890	400996	400668	402362

Figure 10. Identification of Nearest Three Upstream and Downstream VDS

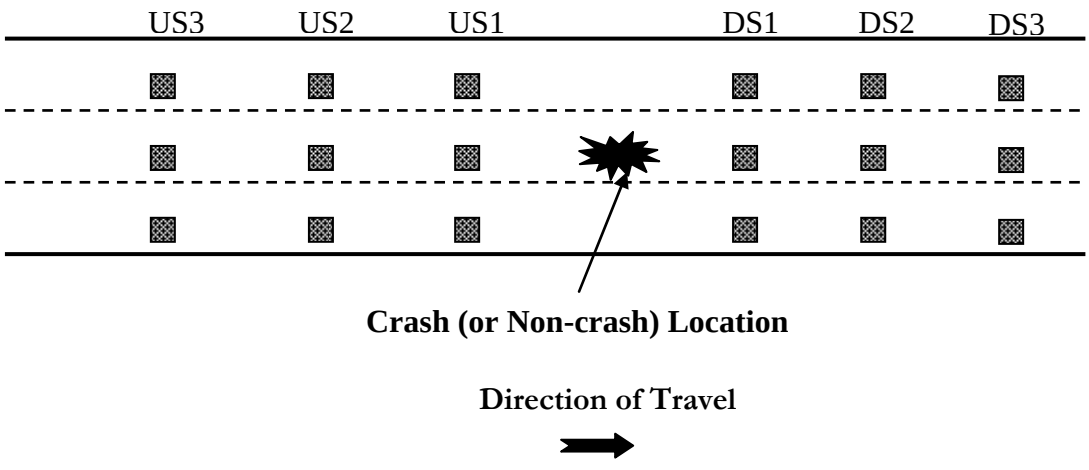

Figure 11. Arrangement of the Loop Detector Stations

The upstream station ids in the order of increasing distance from the crash site are US1, US2, US3 while downstream stations in the order of increasing distance from the crash site are named DS1, DS2, DS3 (yellow highlighted cells in Figure 10). In addition to the ids the spreadsheet snapshot also shows the mileposts

that were identified for these VDS locations (brown highlighted in Figure 10).

Data Aggregation

One of the previous studies (Pande and Mohamed Abdel-Aty 2006a) noted that there is significant noise in the raw 30-second loop detector data and therefore they are not suited for modeling purposes. Hence, for each of the 6 VDS locations (3 upstream and 3 downstream) identified for both crash and non-crash events, individual variables were averaged across all lanes, and aggregated into five minute intervals. These intervals are: 0-5 minutes *after* the crash (time slice 0), 0-5 minutes *before* the crash (time slice 1), 5-10 minutes *before* the crash (time slice 2), 10-15 minutes *before* the crash (time slice 3), and 15-20 minutes *before* the crash (time slice 4). For these time slices, standard deviations of the variables were also calculated since past studies documented in the literature review noted variation in traffic parameters was critically associated with the freeway crash potential.

As time slice 0 occurs after the crash, it is only relevant to incident detection and will not be further analyzed or discussed in this thesis. These four 5-minute intervals preceding a crash were selected based on previous research by Pande. Generally, the model will predict more accurately the closer the analysis interval to the crash time. However, there must also be sufficient time for a traffic management center to identify crash-prone conditions and deploy countermeasures; it is therefore likely that only the time slice 2, 3, and 4 models will be relevant, considering the overarching aims of this research (proactive crash management).

The nomenclature for these average and standard deviations is of the form '**XYZα_β**'. 'X' takes the value A or S for average and standard deviation, respectively; while 'Y' takes the value S or V or O for speed or volume or lane-occupancy, respectively. 'Zα' takes the value of U1, U2, U3 or D1, D2, D3 depending on the station to which a traffic parameter belongs (nearest upstream/downstream station relative to the crash location being U1/D1 and subsequent detectors being U2/D2 and U3/D3, respectively). 'β' takes up the values 1, 2,3, or 4 referring to aforementioned four time slices. Hence, 'ASD1_2' and 'AVU1_2' represent average speed on station DS1 over time slice 2 and average volume on station US1 over time slice 2, respectively. It should be noted that all these averages and standard deviations were calculated for crash as well as non-crash cases.

CONCLUDING REMARKS

This chapter described the process of gathering traffic data corresponding to crash and non-crash events from four different freeway corridors, I-880 NB/SB and US-101NB/SB in the city of San Jose. The chapter also included information about the VDS locations along the freeway corridors as well. In the next chapter these data are used to estimate and test statistical (binary logistic regression) and data mining (classification trees) models for classifying crash prone vs. normal conditions on the freeways.

IV. MODELING TOOLS, ANALYSIS, AND RESULTS

This study applies two different modeling tools, namely, logistic regression and classification trees, to identify crash prone conditions. These tools are applied to data from US-101 NB section in order to estimate the models. The models estimated from US-101 NB data are then applied to US-101 SB and I-880 NB/SB segments. This chapter first provides the details of the statistical and data mining methods and then the analysis and results.

LOGISTIC REGRESSION

In a logistic regression setting the function of dependent variables yielding a linear function of the independent variables would be the logit transformation.

$$g(x) = \ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0 + \beta_1 x \qquad (1)$$

Where $\pi(x) = E(Y|x)$ is the conditional mean of Y (dependent variable representing crash occurrence; Y=1 in this case) given independent variable x when the logistic distribution is used. Under the assumption that the logit is linear in continuous covariate x, the equation for the logit would be $g(x)$.Once the model (i.e., the coefficient βs) is estimated for the binary target variable it can be used to score any dataset that contains the required input variable to the model (i.e., x). The output of the model is in the form of a posterior probability of crash occurrence, lying between 0 and 1. Note that the same formulation may be extended to multiple independent variables as would be the case in this research. In case of multiple independent variables, a standard stepwise variable selection method

will be used to finalize the set of variables that are significantly associated with the crash occurrence. The details of logistic regression and stepwise variable selection procedure may be found in standard text on logistic regression and binary data modeling (e.g. Collett (1991) and Hosner and Lemeshow (1989)).

DECISION TREES

A classification tree represents segmentation of data created by applying a series of simple rules. Each rule assigns an observation to a group based on the value of an input. One rule is applied after another, resulting in a hierarchy of groups within groups. The hierarchy is called a tree, and each group is called a node. The final or terminal nodes are called leaves. For each leaf, a decision is made and applied to all observations in that leaf. Decision trees are one of the most widely utilized tools in data mining applications besides neural networks and may be used for classification of categorical variables as well as for continuous targets. The latter application, of course, is not relevant here. The advantage of classification tree over other modeling tools, such as neural networks, is that it produces a model that may represent interpretable English rules or logic statements. The other advantage associated with trees, compared to logistic regression models, is that no assumptions are necessary about the data and the model form. In the next subsection theoretical details of the classification trees are described. Since we would invariably deal with binary target variable (Y=1 for crash and Y=0 for non-crash) in this study the details of the methodology are provided in the context of a binary target. Neural networks and decision tree algorithms have been successfully used to develop classification models for crash severity

as a function of potentially correlated categorical factors Sohn and Shin (2001), and more recently, to demonstrate significant correlation between speed differentials upstream/downstream and crash risk (Pande & Abdel-Aty 2006).

Decision Tree Methodology for Binary Classification

The basic element in classification tree construction is to split each (non-terminal) node such that the descendant nodes are 'purer' than the parent node. A completely 'pure' node would be one that has all its observations belonging to the same class. To achieve this, a set of candidate split rules is created, which consists of all possible splits for all variables included in the analysis. For example, for a dataset with 200 observations and 5 input variables there would be up to 200*5=1000 splits available at the root node. These splits are then evaluated based on a criterion to choose amongst various available splits at every non-terminal node (including the root node). Gini Index is used as the measure (i.e., 'purity' functions) to rank candidate splits for a binary target variable. This measure was proposed by Breiman et al. (1984).

One of these criteria is applied recursively to the descendants, which become the parents to successive splits, and so on. The splitting process is continued until the criteria of minimum reduction in impurity (i.e., reduction in Gini Index) and/or minimum size of a node are satisfied. To stop the splitting process one may also choose the classification accuracy over the validation dataset (i.e., the dataset not used for estimating the splits) as the criterion. The classification accuracy may be assessed after every split and the process may be terminated if the classification accuracy declines after a particular split. The output from the classifica-

tion tree model is also the posterior probability of an observation being a crash (a number lying between 0 and 1).

Note that these tools are selected since they can provide not only a measure for crash vs. non-crash classification but also the variables included in the model can be explained. Neural networks were also considered as a tool but were not used due to their 'black box' nature. In other words, neural networks were not used here since the results are not transparent in terms of the effect of individual independent variables on the output.

METHOD FOR ANALYSIS OF CLASSIFICATION PERFORMANCE

There were some critical issues that needed to be addressed before proceeding with the modeling exercise. The crashes, however frequent on the corridors under consideration, are still rare events. Sampling their actual proportion in the dataset would mean that the sample would almost exclusively consist of non-crash cases (crash cases would be even less than 0.001 %). It is reasonable to assume that the crash prone conditions, which would be worth issuing warnings, are more frequent than the crashes themselves. For any model intended to be applied in real-time the ideal sample composition for modeling would have proportion of the two competing events same as that in reality. However, there is no way, at this stage anyway, to estimate the proportion of crash prone conditions on the freeway. Also, since the number of warnings beyond a certain point would mean "unreasonable" number of false alarms; the decision from the models cannot be positive (i.e., a crash) in the vicinity of 50% of the time. Hence, a sample with equal number of crash and non-crash cases would not make an ideal sam-

ple. At this point, 10% was deemed to be an appropriate ratio for crash vs. non-crash cases. Therefore, in the sample there were 10 non-crash cases for each crash.

Due to imbalance in the proportions of crashes vs. non-crash cases model performance evaluation issue becomes complicated. The output of these models (for any observation) is the posterior probability of the crash. As mentioned Posterior probability is a number between 0 and 1. The closer it is to unity the more likely, according to the model, it is for that observation to be a crash. Usually the overall classification accuracy based on a pre-selected threshold is an appropriate measure to judge the performance of the model. However, with only 9.1% of the crashes in the sample (1 crash for 10 non-crash cases) used for modeling, 90.9% overall classification accuracy could be achieved by a model that merely classifies every data point as non-crash. Such a model would of course be useless for crash prone conditions identification. Also, since the classification performance of the models would vary based on the cut-off set on the output from the models (i.e., the posterior probability) even the classification accuracy over each individual class (at a certain cut-off) would not be appropriate to compare performance of competing models. It will only reflect the performance of the model at a predetermined threshold on output posterior probability. It is especially true here since we have two different classes of models and their outputs are calibrated differently. The same threshold can potentially produce varying results for these two different classes of modeling techniques. Therefore, a well-calibrated measure of performance evaluation was needed instead and it is pre-

sented below.

Performance Evaluation Measure

To evaluate, the performance of the estimated models were applied to a dataset consisting of the input variables. The output of these models (for any observation) is the posterior probability of the crash. The closer posterior probability is to unity, the more likely, according to the model, it is for that observation to be a crash. To assess the performance of any model observations, the output dataset were sorted by the estimated posterior probability. In the sorted group, the top 10% of observations would be those that are most likely to be a crash, according to the model. The performance of a model may be measured by determining the proportion of crashes captured within various deciles[1] of posterior probability. Since these models are intended to identify an event as rare as crashes, to choose among competing models the proportion of crashes captured within the first few deciles must be critically examined. It was decided that the best model among a set of competing models would be the one capturing the highest percentage of crashes within the first three deciles (i.e., 30^{th} percentile). As mentioned earlier due to imbalance in the proportions of crash and non-crash cases in the sample, overall classification accuracy over validation dataset would not be a good measure for model performance evaluation. In the next section the mod-

[1]Decile is defined as any of nine points that divide a distribution of ranked scores into equal intervals with each interval containing one-tenth of the scores

eling and results are discussed for logistic regression followed by classification tree based analysis.

LOGISTIC REGRESSION ANALYSIS

Overview

The analysis was conducted by estimating multivariate logistic regression modeling using US-101 NB data. The statistical analysis software package SAS (SAS Institute, 2001) was used to fit the regression models. The target variable for these logistic regression models was Y taking value 0 for non-crash cases and 1 for crash cases. The independent variables of interest were: average speed, standard deviation of speed, average volume, standard deviation of volume, average lane-occupancy, and standard deviation of lane-occupancy calculated over each VDS location and time slice. It should be noted that all three of these traffic parameters (speed, volume, and lane-occupancy) were not included simultaneously in any model and speed-based models were created separately from the volume and occupancy-based models, as the study VDS were all based on single loop detectors. This implies that speed was calculated from the volume and occupancy data and not independently measured. Including them in the same regression model would have led to unacceptable level of correlation in independent variables. A stepwise selection process was used to identify the most significant variables, and the model coefficients were estimated for these significant variables.

In all, a total of 30 different logistic regression models were estimated. These models were estimated with traffic information from 4time slices (ranging from 0-

20 minutes before the crash in 5 minute intervals) and three different sets of VDS locations. For each VDS and time slice combination there were two models: one model based on Caltrans' derived speed information, and the other model based on independently measured volume and lane-occupancy information. The crash risk estimation models are identified as predX_Y_Z; where

- X represents the number of VDS stations upstream and downstream of the crash (or non-crash) location (1, 2, or 3) contributing traffic information to the model
- Y represents the time slice number (1,2,3, or 4) as described in the previous chapter
- Z represents whether the model uses speed information (s) or volume and lane-occupancy information (v)

For example:

pred1_4_s represents that the model is developed from dataset of *speed* observations from the *one* nearest VDS both upstream and downstream of the crash, over the period of *15-20 minutes before* the crash occurred. Note that this model utilized traffic data from a total of two VDS locations

pred3_4_s represents the dataset of *speed* observations from the *nearest three loop detectors both upstream and downstream of the crash*, over the period of *15-20 minutes before the crash* occurred. Note that this model utilized traffic data from a total of six VDS locations, with two of the VDS locations being the same as pred1_4_s.

These 30 models were then applied to the dataset used to estimate the models containing observations (both crash and non-crash events for US-101 NB), and the posterior probability of the observation being a crash was estimated for each observation. These models were then compared with each other in terms of the cumulative proportion of crashes correctly identified within 30% observations which according to the model were most likely to be a crash. It is the criterion selected based on the discussion provided in the previous section. It is worth mentioning that the percentage of crashes identified by each model can also be examined in the context of the "*performance*" of a random baseline 'model' which represents the percentage of crashes identified in the sample if one randomly assigns observations as crash and non-crash. Of course in any set of 30 percent observations such a 'model' would be able to correctly identify 30% of crashes in the dataset. Any model can be assessed for its classification based on the difference between crashes it identifies within the first three deciles vs. 30%. 30% is the percentage of crashes that can be identified by the random baseline 'model'.

Following this criterion the best model was selected from subsets of one, two, and three upstream/downstream VDS models. Traffic parameters from time-slice 1,being too close to time of the crash used in a model, would leave absolutely no leeway in terms of time available to process, analyze and disseminate the information that may in turn be used to avoid crashes. Hence, in the following section the models from variables measured only during time slice 2, 3, or 4 are given further consideration.

The single loops analyzed in this study collect raw volume and occupancy data

and use a predetermined effective vehicle length (g-factor) to calculate average speed; this stands in contrast with dual loops which can measure speeds directly. Acknowledging that this g-factor will vary by lane, time of day, and loop sensitivity, PeMS calculates a g-factor for each loop for every 5-minute period during an average week to improve the accuracy of the speed estimates. The smoothened g-factor factor is then applied to the real-time VDS data to obtain speeds. These real-time reported speeds are then smoothened with an exponential filter, which is weighted based on traffic flow to produce reasonable estimates of speed (that is, lower flow conditions require more smoothening).

In general, it was found that the volume and occupancy (v) models had a much higher classification accuracy at the 30^{th} percentile than the speed (s) models. This is understandable, as the speeds derived by the PeMS algorithm are inherently less reflective of field conditions than looking at the actual VDS data. Additionally, only the volume and occupancy data are reported live by Caltrans districts (in a variety of methods, including XML feed over TPC, SQLnet, and raw controller packets over RPC); speeds must be post-processed from this transmitted data. Only the volume and occupancy models will be further considered in this thesis, for reasons of model reliability and applicability in a real-time framework. The following section described the US101 NB models for all crashes and non-crash data available from the freeway.

All Crash Model Comparison

The best models for the former case, using the 30 percentile selection criteria, are summarized below in Table 3.

Table 3. Selection of Best Models for All Crashes

Model Name	Time Slice	Cumulative % Of Crashes Captured Within first three deciles (30th percentile)
1-VDS Upstream and Downstream Models		
pred1_4_v	**4**	**53.463**
pred1_3_v	3	52.276
pred1_2_v	2	50.069
2-VDS Upstream and Downstream Models		
pred2_4_v	4	56.546
pred2_3_v	3	56.711
pred2_2_v	**2**	**57.524**
3-VDS Upstream and Downstream Models		
pred3_4_v	**4**	**61.749**
pred3_3_v	3	61.264
pred3_2_v	2	60.000

It was found that the **best 1-VDS model** used volume and occupancy data from the fourth time slice, **pred1_4_v**. The **best 2-VDS model** used volume and occupancy data from the second time slice, **pred2_2_v**. The **best 3-VDS model** used volume and occupancy data from the fourth time slice, **pred3_4_v**. These models are in 'bold' in Table 3 above. It is noteworthy that when one uses data from more VDS locations the classification accuracy increases. The model from 3-VDS upstream and downstream is able to identify more than 61% of the crashes on US-101 NB segment, which is a noticeable (31%) improvement over the random baseline 'model'.

Model Details

This section provides the coefficients of the best 1-VDS, 2-VDS, and 3-VDS logistic regression models. Tables 4, 5, and 6 show the best 1-VDS, 2-VDS, and 3-VDS model, respectively. The tables only show the variables included in the

models based on the stepwise selection procedure. In addition to the model parameters, the tables also provide the corresponding p-value for the model coefficients. A p-value less than 0.05 indicates that the variable is significant at 95% confidence level. Positive (negative) coefficient means that as the value of the corresponding variable increases the crash risk measure increases (decreases).

Table 4. Model Coefficients for the Best 1-VDS Model

Parameter	Estimate	Pr > ChiSq
AVDS1_4	0.1	<.0001
AVUS1_4	0.08	<.0001
AODS1_4	1.72	<.0001
AOUS1_4	0.87	0.0058
SVDS1_4	0.05	0.1355
SVUS1_4	-0.1	0.0035
SODS1_4	-0.57	0.2157

[a]*Syntax:*

Column 1: *A = average;* *S = standard deviation*
Column 2: *O = occupancy;* *V = volume* *S = speed*
Columns 3&4: *DS = downstream;* *US = upstream*
E.g.: *AODS = average occupancy downstream*
Red text denotes statistical significance at the 95% confidence level.

Table 5. Model Coefficients for the Best 2-VDS Model

Parameter	Estimate	Pr > ChiSq
AVDS1_2	0.05	0.0138
AVUS1_2	0.04	0.027
AODS1_2	0.91	0.0171
AOUS1_2	1.5	0.0443
SVDS1_2	0.07	0.0997
SOUS1_2	-0.93	0.2134
AVDS2_2	0.05	0.0442
AVUS2_2	0.08	0.0013
AODS2_2	1.45	0.0158
AOUS2_2	-1.49	0.139
SVDS2_2	-0.22	<.0001
SVUS2_2	-0.08	0.081
SODS2_2	-1.85	0.0061
SOUS2_2	2.87	0.0024

Table 6. Model Coefficients for the Best 3-VDS Model

Parameter[a]	Estimate	Pr > ChiSq (p-value)
AVUS1_4	0.13	<.0001
SVDS1_4	0.08	0.0407
SVUS1_4	-0.18	0.0017
AVUS2_4	0.06	0.0318
AODS2_4	-1.24	0.0389
SVDS2_4	-0.09	0.0454
SVUS2_4	-0.1	0.068
SOUS2_4	2.7	<.0001
AVDS3_4	-0.11	<.0001
AVUS3_4	0.11	<.0001
AODS3_4	1.87	0.0112
AOUS3_4	3.04	0.0003
SVDS3_4	0.16	0.0023
SODS3_4	-1.82	0.0169
SOUS3_4	-1.32	0.1416

It can be observed from the model coefficients that the standard deviation of occupancy downstream of a freeway location is negatively associated with crash risk; i.e., if standard deviation of lane-occupancy decreases the crash risk increases. Also, variables representing average occupancy downstream (AODS*_*) have a positive coefficient in all models indicating if there is increased lane-occupancy (i.e., congestion) downstream of a site then the crash likelihood

increases. Since the specific crash type is not known it is not possible to relate these coefficients with the relevant crash mechanism. However, it can be said that these coefficients might be more readily associated with conditions prone to rear-end crashes, which are the most common crash type on urban freeways.

Model Application for Assessing Transferability

Transferability evaluation is one of the biggest contributions of this research project, that is the potential to apply the predictive model developed on one freeway segment to other similar facilities in the nearby area. As was discussed in the literature review, previous studies have either failed to address the issue (which is critical to real-time application in a network) or tried to apply the model on dissimilar facilities (such as in a different study area) and subsequently failed to attain good classification accuracy.

To assess transferability, coefficients of regression models shown in Tables 4, 5, and 6 were used to score the combined crash and non-crash data for the other three corridors on US-101 SB, I-880 NB, and I-880 SB. Again, for each observation in these datasets a posterior probability output was obtained. We then examined the proportion of crashes in the dataset correctly identified within the 30% observations having the highest posterior probability. The cumulative percentages of identified crashes for each model on each of the three corridors are depicted in Figure 12. Note that the model which identifies higher proportion of crashes within 30^{th} percentile is considered a better model.

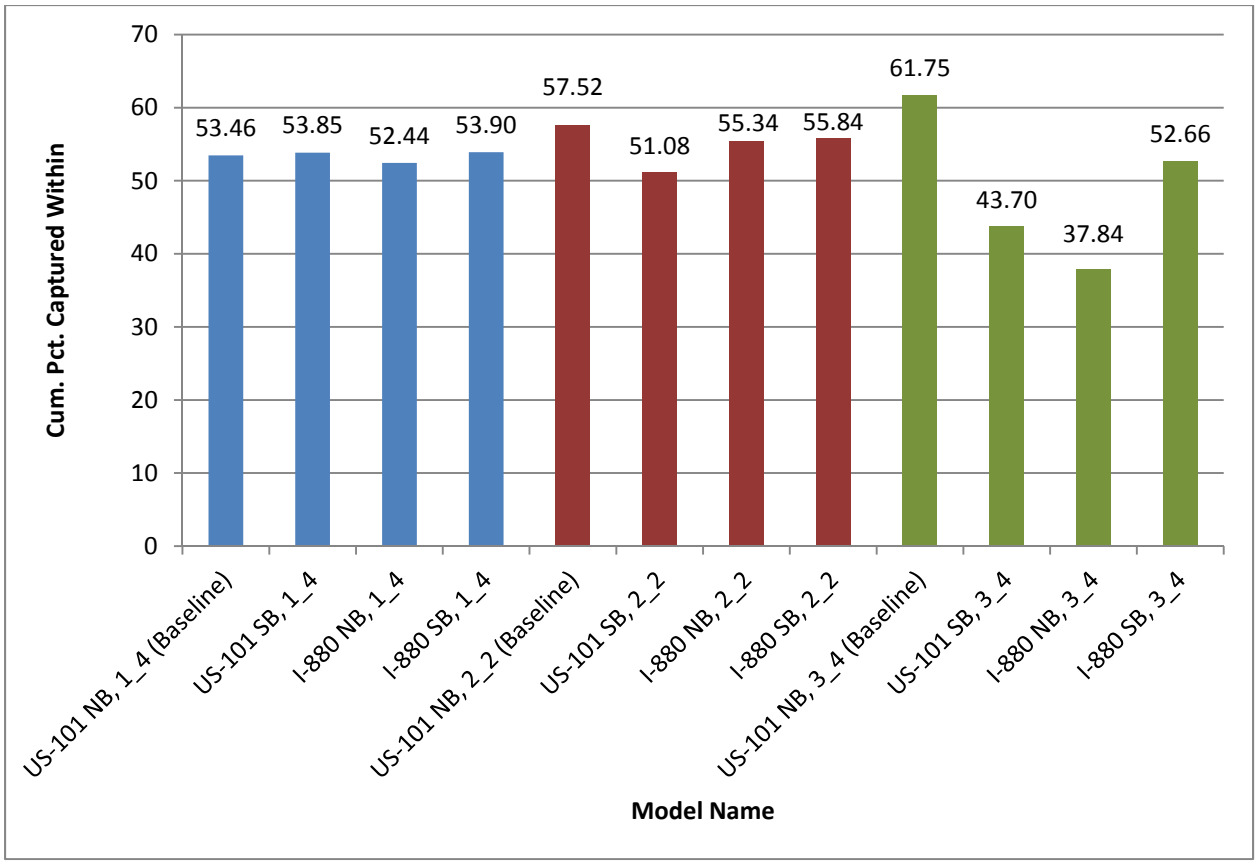

Figure 12. Transferability Analysis for the Three Models

Tables showing this same data for each model are presented below in Tables 7 to 9. These tables also show the information used by the model in terms of up-stream/downstream stations as well as time slice.

Table 7. Classification Accuracy of Best 1 VDS Model applied to Other Freeways, All Crashes

Best 1 VDS Model Name:	**pred1_4**
VDS US/DS:	**1**
Time Slice:	**4**
Mins Before Crash:	15-20
Selection Criteria:	30%
Segment	**Percent Captured Within**
US-101 NB (Estimation baseline)	**53.463**
US-101 SB	53.846
I-880 NB	52.439
I-880 SB	53.898

Table 8. Classification Accuracy of Best 2 VDS Model applied to Other Freeways, All Crashes

Best 2 VDS Model Name:	**pred2_2**
VDS US/DS:	**2**
Time Slice:	**2**
Mins Before Crash:	5-10
Selection Criteria:	30%
Segment	**Percent Captured Within**
US-101 NB (Estimation baseline)	**57.524**
US-101 SB	51.083
I-880 NB	55.340
I-880 SB	55.844

Table 9. Classification Accuracy of Best 3 VDS Model applied to Other Freeways, All Crashes

Best 3 VDS Model Name:	**pred3_4**
VDS US/DS:	**3**
Time Slice:	**4**
Mins Before Crash:	15-20
Selection Criteria:	30%
Segment	**Percent Captured Within**
US-101 NB (Estimation baseline)	**61.749**
US-101 SB	43.700
I-880 NB	37.838
I-880 SB	52.660

It can be clearly seen that both the 1-VDS and 2-VDS models work comparably well on nearby freeways, as on the same roadway for which they were developed. The 3-VDS model developed for US-101 NB is a much less accurate predictor of crashes on the nearby roadway segments. In other words, the 1-VDS and 2-VDS models are easily transferable while 3-VDS model does not seem to be transferable.

It appears that the 3-VDS model is overfitting; we believe that traffic conditions that far away from the crash location (approx. 1.5 miles in each direction) do not have a real relationship with crash risk 15-20 minutes later. This is why the overfitting is happening on the training data; when tested with an unseen dataset, the model is not performing very well.

In the next section, the analysis is repeated for crashes that occurred on the weekdays between the hours of 5:00 AM through 10:00 PM.

Daytime-Only Models

Daytime only models were estimated since late night crashes were postulated to be more likely to occur due to driver error or driving conditions (e.g., under influence), rather than measurable traffic conditions. The modeling process and model comparison was identical to above except for the fact that the regression models were estimated using data only for crashes and non-crash cases between the weekday hours of 5:00am and 10:00pm. A summary of the model results is shown below in Table 10.

Table 10. Selection of Best Models for Daytime-Only Crashes

Model Name	Time Slice	Cumulative % of Crashes Captured within 30th Percentile (US-101 NB)
1-VDS Models		
pred1_4_v	**4**	**52.362**
pred1_3_v	3	51.866
pred1_2_v	2	50.000
2-VDS Models		
pred2_4_v	**4**	**57.724**
pred2_3_v	3	56.873
pred2_2_v	2	55.285
3-VDS Models		
pred3_4_v	**4**	**60.324**
pred3_2_v	3	59.438
pred3_3_v	2	59.438

Again, it was found that the volume-occupancy models' performance was better than those based on calculated speed information in almost every case. So the speed models were dropped from the analysis. The best 1, 2, and 3 VDS models all used volume and occupancy data from the fourth time slice.

A comparison of the daytime-only results to the all-crash results is shown below in Table 12. Note that Tables 10 and 11 show the performance of the models on the US-101 NB dataset, which was also used to estimate the model.

Table 11. Best Three Models for All Crashes and Daytime-Only Crashes

All Crashes			Daytime-Only Crashes		
Model Name	**Time Slice**	**Cum. Pct. Captured**	**Model Name**	**Time Slice**	**Cum. Pct. Captured**
1-VDS					
pred1_4_v	4	53.463	pred1_4_v	4	52.362
2-VDS					
pred2_2_v	2	57.524	Pred2_4_v	4	57.724
3-VDS					
pred3_4_v	4	61.749	pred3_4_v	4	60.324

It can be observed that there is not an appreciable difference in the performance of the all-crash models compared to the daytime-only crash models. Hence, it is not advantageous to estimate the model only for daytime crashes. This is the reason why the transferability analysis for daytime only crashes is also not discussed here. In the next section, models are estimated using classification trees, which is a data mining tool.

CLASSIFICATION TREE ANALYSIS

Overview

Classification tree models are one of the more often utilized data mining tools. One concern with these models is that they tend to over-fit the data which affects their future performance on the unseen datasets. Therefore, standard practice in data mining analysis is to estimate a model with a “training dataset” using 70% of the available observations, and then validate the model using the remaining 30%. Validating them with the unseen dataset helps identify a more robust model in terms of its performance on new datasets.

Similar to the logistic regression approach there were 30 different classification tree models that were estimated. From the 30 models, those using data from time slice 1 were excluded based on the reasons discussed in the last section. In the case of classification tree models it was observed that the speed models were generally better than the volume-occupancy models. These classification tree models were compared using the same metric used for the logistic regression models, which is the proportion of validation dataset crashes identified within

the top 30 percentile.

Selection of Best Model

In addition to different 1-VDS, 2-VDS, and 3-VDS models from different time slices we estimated a classification tree model with just time of crash (and non-crash) as input. This model was estimated in order to ensure that the models are providing real differentiation between crash prone and normal traffic conditions. If the models using traffic data are providing valuable information about crash risk then these models should perform much better than the model with just time of crash/non-crash information. It turns out that these models do in fact perform much better than the time of crash information only model. Table 12 shows the classification performance of best 1-VDS, 2-VDS, and 3-VDS models along with time of crash model. It is clear that while time of the crash model performs better than the random baseline 'model' (i.e., identifies more than 30% crashes); the model is significantly worse than the models using traffic information. It is worth mentioning that the model performance in Table 12 is on the 30% validation set aside from the whole US-101 NB dataset. The results shown below in Table 12 and Figure 13 identify the **2-VDS, time slice 3 model** as the most accurate classifier on the validation dataset.

Table 12. Classification Accuracy of Classification Tree Models

Model Name	US/DS VDS Locations	Time Slice	Cumulative % of Crashes Captured Within 30th Percentile (Validation dataset)
Pred1_4_s	1	4	56.662
Pred2_3_s	**2**	**3**	**58.647**
Pred3_3_s	3	3	56.309
Time of Crash Model	-	-	43.771

Model Details

Classification tree models are a series of "if-then" rules to classify the observations. The exact set of rules for the best model is provided in the Appendix. The variables analyzed through classification trees for crash vs. no-crash classification can be ranked by combination of the number of times they appear in various rules and number of observations they contribute in classifying. For the best classification tree model (Pred2_3_s) variables included in the model were ranked as follows:

1. SSDS2_3
2. ASDS2_3
3. ASUS1_3
4. SSUS1_3
5. ASUS2_3
6. SSDS1_3
7. ASDS1_3

According to the list above, standard deviation and averages of speed at the second downstream VDS are the two most significant variables, respectively. The results are consistent with the past studies which have found the turbulence in speed downstream of a location is significantly associated with crash risk on urban freeways. It is worth mentioning that the standard deviation of speed at the second upstream VDS (SSUS2_3) was the only variable that was found to be not associated with the crash likelihood.

Transferability Analysis

The best classification tree model (Pred2_3_s) was applied to complete sets of data from the US-101 SB and I-880 NB/SB. In addition to these three nearby freeway corridors, the model was also applied on the complete set of US-101 NB data itself. It was done since the results shown in Table 12 are based on applying the tree model on the validation dataset (i.e. 30% of observations from US-101 NB). Note that the classification accuracy is higher in Table 13 (61.897%) for US-101 NB than over the validation dataset (58.657%; Table 12) since the complete set also includes the 70% training data as well. Applying the model on US-101 NB dataset allows us to compare the tree model performance with the logistic regression model from the previous section.

Table 13. Classification Accuracy of Best US-101 NB Classification Tree Model on Other Freeways

Facility	Proportion of Crashes Identified within 30 percentile (Classification tree model)	Proportion of Crashes Identified within 30 percentile (Logistic regression model)
US-101 NB (Estimation Baseline)	**61.897**	**61.749**
US-101 SB	46.505	43.700
I-880 NB	40.674	37.838
I-880 SB	50.368	52.660

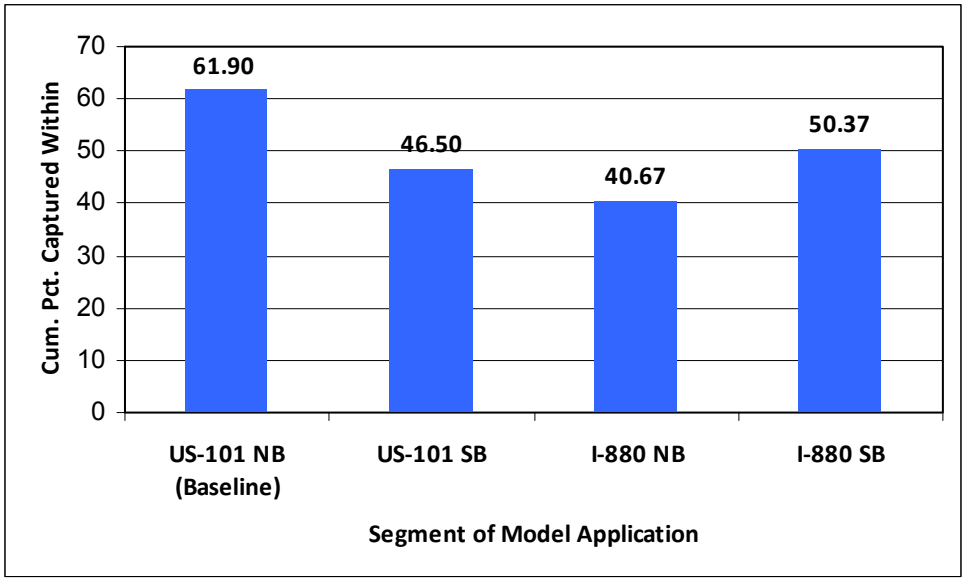

Figure 13. Transferability of the Best US-101 NB Model

Note that the best classification model performs slightly worse on the other freeways as was the case with best logistic regression model. I-880 SB is the corridor where the model estimated from US-101 NB data seems to be most readily transferable based on the classification performance. In the next chapter, the conclusions from this analysis are discussed in terms of the real-time crash risk implementation framework.

V. REAL TIME APPLICATION FRAMEWORK

PROCEDURE

The models developed here may be applied in real-time, as they are capable of classifying the traffic patterns measured at VDS into posterior probability. A step-by-step procedure is shown graphically in Figure 14, and described below.

We first try to obtain data from three VDS upstream and downstream of the location of interest, as the 3-VDS models are the best estimators of crash risk (on the corridor for which the original model was developed). If all the VDS are in good health after a data check, the 5-minute averages and standard deviations of traffic variables are calculated for each location. Estimated model coefficients (for logistic regression models) or if—then rules (for the classification tree) models can be applied to obtain the measure of crash risk at the middle of the section.

If data from all 6 VDS stations are not available due to intermittent loop failures, a check for data availability is applied for the 2-VDS model (total of 4 VDS needed). Using the same procedure as described above for the 3-VDS application, traffic parameters are calculated an input into a model. As was noted in the transferability discussion, models developed using 2 VDS on nearby freeways are transferable to other roadways. If models have not yet been developed specifically for the segment of interest (which indeed perform the best), a 2 VDS model from a nearby roadway can be applied by the system and used to estimate crash risk.

If there is only enough good data to run a 1-VDS model, the same procedure is

applied as was for the 2-VDS model. Traffic parameters are calculated from the VDS data and input into the calibrated 1-VDS model for the segment (if available). If a model has not yet been specifically developed for the location, a 1-VDS model from another freeway can be applied to produce a reasonably accurate assessment of crash risk.

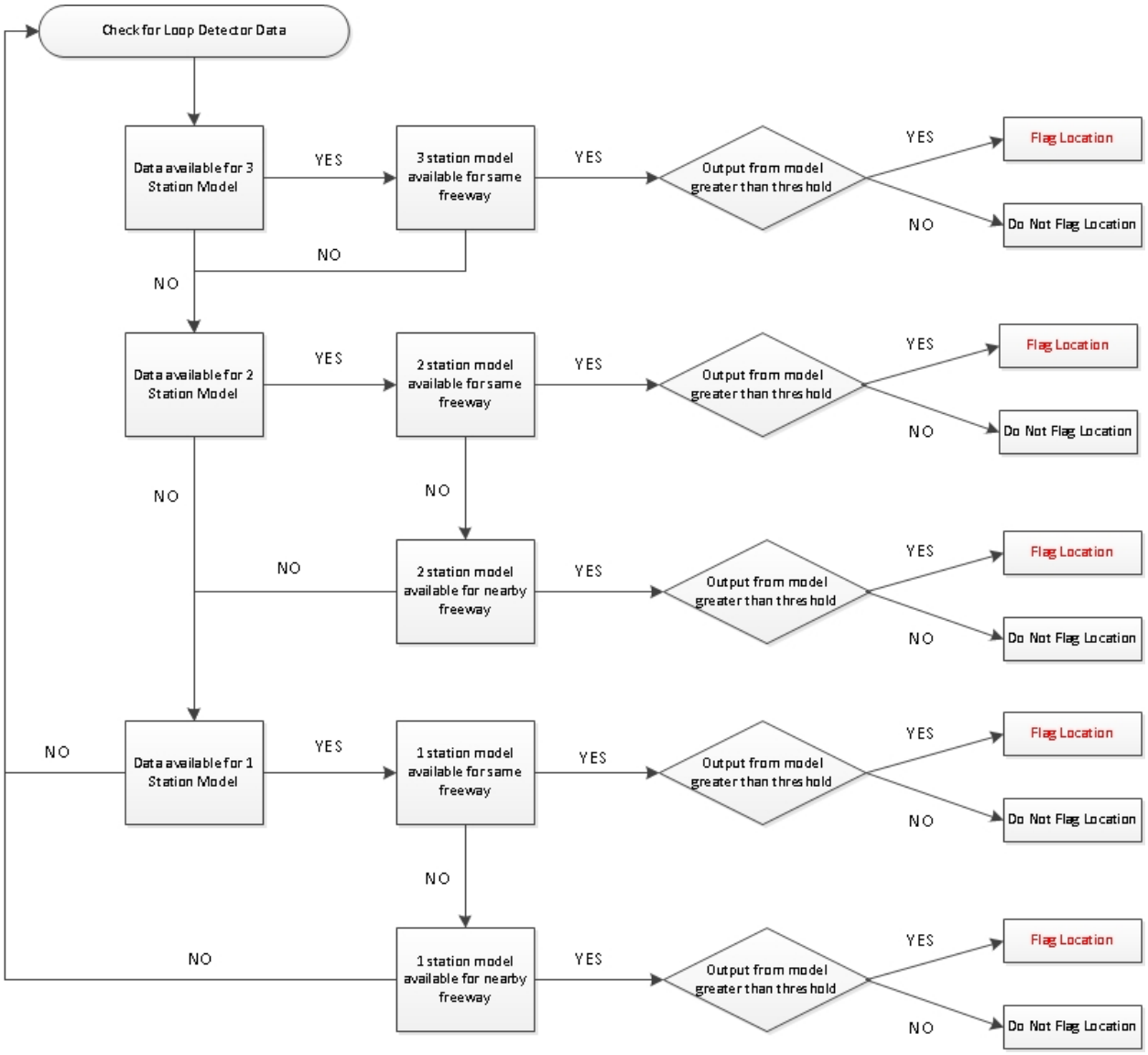

Figure 14. Real-time Application Procedure

If the output posterior probability for a segment of freeway is consistently high then the traffic management authorities can keep their crash mitigation squad on alert so that the impacts of crash occurrence may be minimized. Also, if there are some freeway segments where the models trigger the warning more often than the other locations, these segments may be closely watched through the freeway cameras. This will help recognize any problems associated with these locations. Another application for the findings of this research could be formulation of VSL (variable speed limit) and/or Ramp metering strategies that can reduce the estimated likelihood of crashes. These strategies can be tested using microscopic traffic simulation models.

REAL-TIME APPLICATION ISSUES

1-VDS vs. 2-VDS vs. 3-VDS Models

It should be noted that even though the 1 VDS models may not always achieve classification accuracy as well as the 3 VDS model for the same corridor, the advantage of using those models is that they have more tolerant data requirements. Since the 3-VDS models require that the data be available from 6 simultaneous VDS locations. If even one of the VDS is malfunctioning then the 3 VDS model cannot be applied. 1-VDS model on the other hand, requires data from only 2 VDS locations.

False-Alarms

The formulation of the problem along with the solution approach adopted here is similar to incident detection. In fact, the authors did estimate some models that

used the data 0-5 minutes after the crash. However, the objective of the analysis is to identify crash prone conditions i.e., the conditions in which drivers are more likely to make errors resulting crashes, rather than pinpoint the occurrence of a crash. Conditions prior to crashes (present research problem) are not as readily identifiable (possibly due to significant human factor involvement) as the conditions following the crashes (approach for incident detection). Crashes being such rare events, it is not possible to avoid the false alarms.

Adopting the approach used here for assessing classification models (cases with highest 30% crash risk measure output) even the modest 30% positive decisions would result in a significant number of 'false alarms' throughout the day. One may bring it down to an extent by using a higher threshold (e.g., 20 percentile value for the posterior probability), it would still remain significant. Traffic parameters from time-slice 1, if used as inputs instead of the parameters from time-slice 2, can also be expected to provide slight improvement. However, time-slice 1 being too close to time of the crash would leave absolutely no leeway in terms of time available to process, analyze and disseminate the information that may in turn be used to avoid crashes. Hence, it is our opinion that the warning of crash prone is provided, if at all, not as an event prediction but as a heightened measure of crash risk.

It is also worth mentioning that 'false alarms' are not as detrimental in the present application as they are in case of incident detection algorithms. In fact, the ultimate goal of this research would, or at least should be, to 'achieve' a 'false alarm' every time a crash warning is issued. The goal would be based on the expecta-

tion that with some form of proactive countermeasure or warnings to the motorists, potential crashes following the crash prone conditions may be avoided. The justification or inevitability of false alarm does not mean that an unlimited number of warnings could be issued; especially if the information based on the model output is being transferred to the drivers on the freeway. The reason for being judicious about the number of warnings would be to ensure that the drivers do not perceive the number of warnings to be “too many” and become immune to them. The whole notion of warnings and drivers’ reaction to them are beyond the scope of the present work and require further investigation.

VI. MODEL ROBUSTNESS

No model will classify an event as a crash or non-crash with perfect accuracy. However, it is important to identify situations in which a model performs better than in other situations. Hence, these models' outputs for US-101 NB, as well as to data from the three other freeway segments, were then assessed for their classification performance in a variety of situations. This analysis of robustness has not been carried out in the similar studies and may help in identifying location and times of day/days of week for which additional training of the neural network may be warranted. To study the robustness of the models, for each model (1-VDS and 2-VDS models discussed above), all cases (crash and non-crash) were sorted in descending posterior probability output so the ones most likely to be a crash were at the top and the least likely ones at the bottom. All non-crashes in the top 10% observations (most likely to be crashes according to the model) were labeled as "false positives" and all crashes in the bottom 10% of observations (least likely to be crashes according to the model) were labeled as "false negatives." This process was repeated for events on all four freeway segments.

To examine the robustness of the model, we examined patterns in these "false positives" and "false negatives": day of week/time of day (off-peak, morning peak, or afternoon peak), and location of the crash/non-crash case. While potentially significant, incident duration could not be analyzed in this framework since the California Highway Patrol database from PeMS was missing this information for most of the cases. The findings for the false positives and false negatives for each model were compared to the model performance on all crash and non-

crash cases. It should be noted that the false negative (crash cases deemed safe by the model) is less conclusive due to the smaller sample size, although there are clearly observable trends. The trends shown below are from 2-VDS model for US-101 NB.

TIME OF DAY AND DAY OF WEEK

Figure 15 shows that while more than 80% of overall data was from the off-peak locations, among the “false positives” and “false negatives” off-peak periods represented a smaller proportion. The morning peak is overrepresented in “false positives”. It indicates that while the model deems the morning peak conditions to be crash prone, there are fewer crashes. It may be caused by the fact that the drivers are more attentive during morning peak periods and are able to successively navigate through crash prone conditions.

While the trends is not as pronounced in the afternoon, it appears that there are more false negatives indicating that in the afternoon drivers end up in crashes even when the model is not detecting these conditions. While drivers’ fatigue may play a role here, it could also be caused by the fact that congestion in the afternoon can back up much faster and those conditions are not captured by the model, since it uses data from up to 10 minutes before the crash.

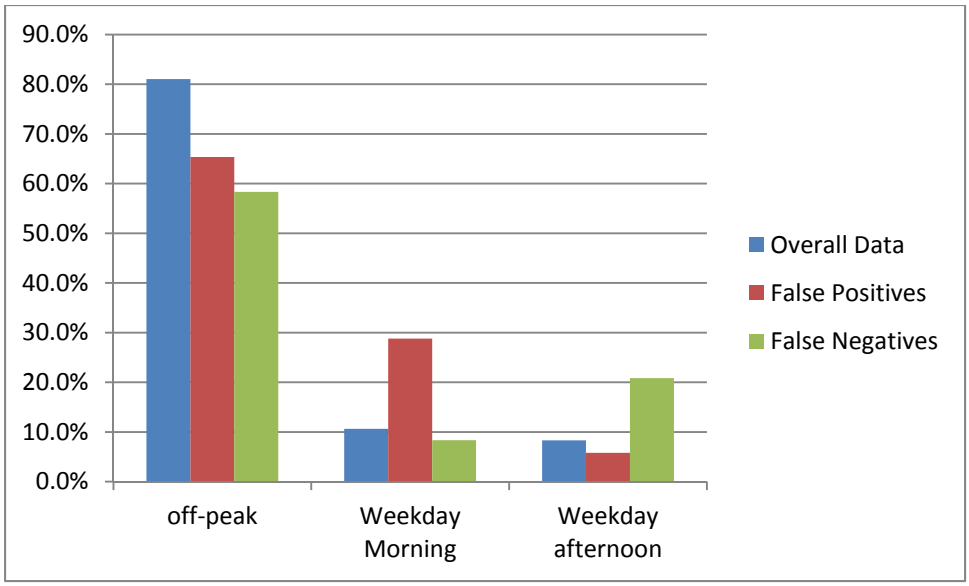

Figure 15. Robustness of the Best Model

LOCATION

We next evaluated whether there are any locations that were overrepresented in the misclassifications. The first upstream VDS location for all “false positives” and “false negatives” was determined as a subset of the original spatial distribution of all incidents. While most locations had the false positives and false negatives consistent with their proportion in the overall data, there were three locations that were noteworthy on US-101 NB:

VDS 401890: Higher percentage of “false positives”: Figure 16 shows that this VDS is located at the US-101/I-280/I-680 interchange, where a large amount of weaving, merging activity may lead to higher speed variations. Higher level of turbulence prevailing in this location means that the drivers need to carefully navigate through this section, since the model deems this location to be crash prone more often than others.

VDS 400858 and 400195: Higher percentage of false negatives: Figures 17 and 18 show that these locations are on long, straight US-101 NB segments, where other factors (driver errors at high speed) are likely to be responsible for more crashes.

It is worth mentioning that while results from all freeways demonstrated these trends; the trends from the other freeways mirror US-101 NB results to the degree of how well the original predictive US-101 NB model fit the other data. For example, I-880 NB was closest to the US-101 NB in terms of crash identification and hence the trends on I-880 matched most closely to the US-101 NB trends.

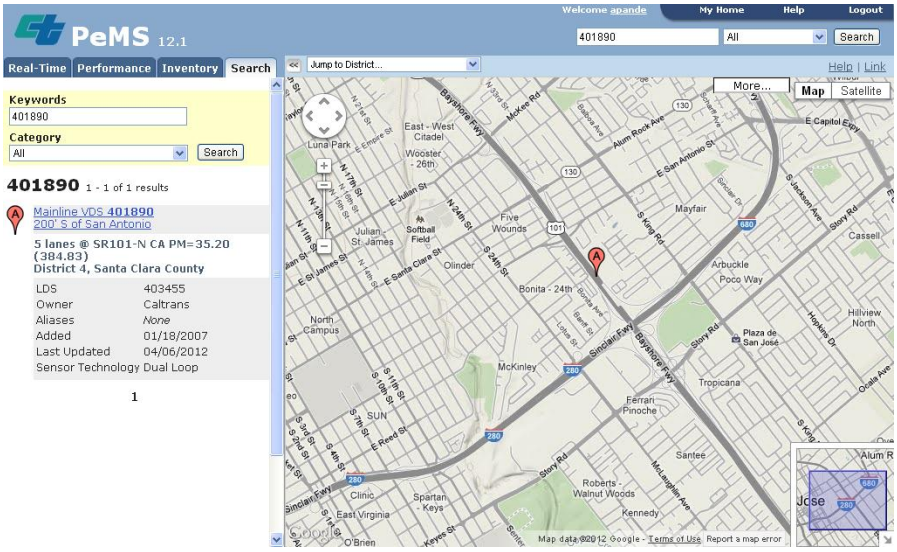

Figure 16. Location Map of VDS 401890 (High False Positives)

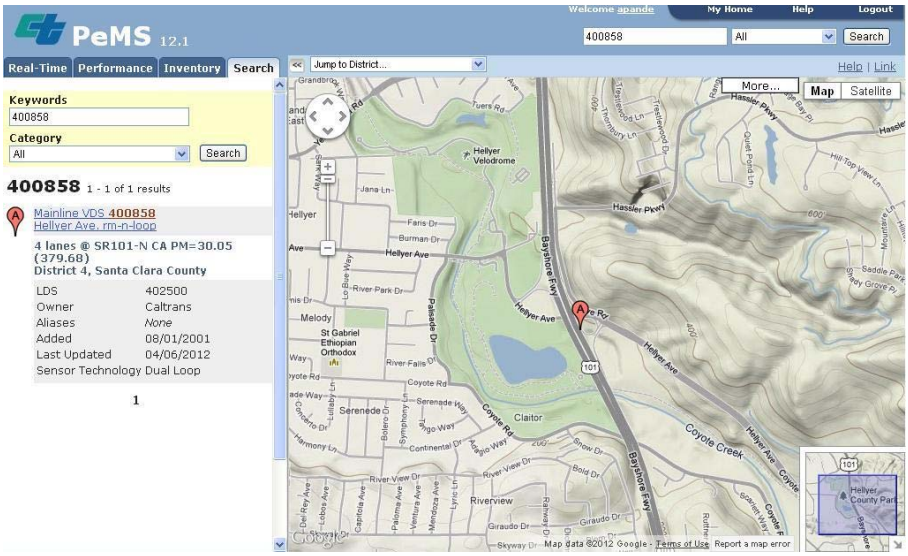

Figure 17. Location Map of VDS 400858 (High False Negatives)

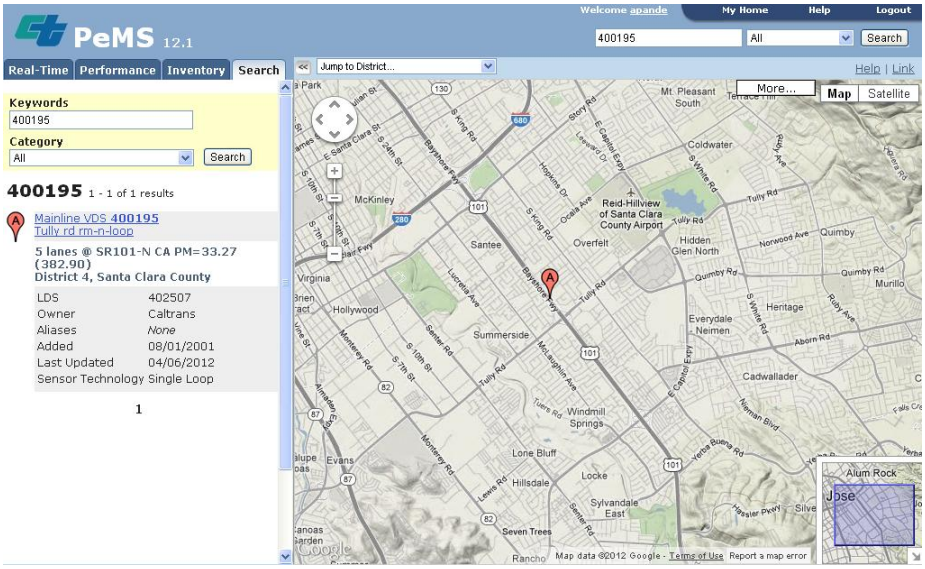

Figure 18. Location Map of VDS 400195 (High False Negatives)

VII. CONCLUSIONS

The objective of this research was to develop and assess transferability of a methodology to link ITS-archived data with historical crashes on instrumented corridors in the San Jose metropolitan area. A detailed database was assembled for all crashes that occurred on four major corridors in the area for a 16-month duration and linked them to the archived loop detector data from the surrounding VDS locations. The analysis of the models' classification results showed that the continuous output of the models (i.e., posterior probability) can in fact be related to real-time crash risk.

TRANSFERABILITY ANALYSIS

While crash risk assessment models have been developed for freeways in US (I-4 in Orlando, FL), Canada, and Netherlands, this research advances the body of knowledge with regards to transferability of the models. Specifically, this project critically examined the performance of models estimated with data from the US-101 NB corridor on nearby corridors (US-101 SB, I-880 NB, and I-880 SB). It was found that the model from one corridor can be applied to other corridors, although the classification performance of the models is not as good as it is on the same corridor.

Answering the question of transferability is important since uninterrupted flow facilities from the same region tend to have similar types of data collection infrastructure. The conclusion from this study on the transferability of the same model can be beneficial to freeways where such infrastructure is either currently under

development or has been recently put in place. The crash risk on such sections can be estimated from a transferable model from the freeways nearby.

Another interesting finding was that the models that use data from smaller section surrounding the crash location (1-VDS) transfer better to the nearby corridors and provide performance comparable to US-101 NB. 3-VDS logistic regression models did not transfer as well to the other corridors. One possible reason is that including traffic data from a larger segment leads to crash risk being influenced by variability in geometric factors. Over a smaller segment the geometrical factors do not vary as widely and therefore the model transfers better to corridors that may have different geometric design.

As logistic regression models include more and more VDS locations, the classification accuracy increases for the freeway segment from which they were estimated. However, it seems to come at a trade-off since these models perform worse when applied to other nearby freeways compared to models that use data from fewer VDS locations. The modeling with data only from weekdays did not change the classification results in any significant way and hence the proposed models used data for all crashes. The classification tree models have comparable classification accuracy to the logistic regression models. The US-101 NB classification tree model was again a more accurate predictor of I-880 SB crashes than for the other two roadways, though not nearly as accurate as for the US-101 NB crashes (as was the case for regression models).

MODEL ROBUSTNESS

As was the case with transferability, the other previously unanswered question pertained to whether the misclassifications from crash risk estimation models are concentrated on certain situations of time of day/day of week or locations.

It is worth noting that, while this research establishes that models for most locations may be transferable from one freeway to the other, some locations on the same freeway may require additional training for crash risk estimation (e.g., the US-101 NB section near the I-680/I-280 interchange). This study provided a framework to flag these locations for additional model training, through analysis of “false positives” and “false negatives” by locations. On a system of freeways, these locations with higher “false positives” or “false negatives” may be combined together from different facilities by not restricting the freeway crash risk estimation model by the corridor.

FUTURE WORK

Improvements to this Research

This research used random generation of both times and locations in order to generate non-crash events. In order to reduce variability in the modeling effort, fixing the location to the actual crash location and then randomizing times should be considered. In addition, using the lasso (instead of stepwise) selection procedure for logistic regression has been suggested to reduce bias in the coefficient estimates.

Additional Topics of Study

For the purposes of this study, all detectable incidents were treated the same in the modeling procedure. It would be an interesting topic of future work to analyze incidents in terms of intensity. Potential measures of severity might include the number of lanes closed, incident duration, and resulting effects on traffic flow.

ABBREVIATIONS AND ACRONYMS

AADT: Annual Average Daily Traffic

Cal Poly: California Polytechnic State University, San Luis Obispo

CHP: California Highway Patrol

ITS: Intelligent Transportation System

HOV: High Occupancy Vehicle

MTI: Mineta Transportation Institute

PeMS: Performance Measurement System, an online database containing both real-time and archived traffic data collected on state facilities.

SJSU: San José State University

VDS: Vehicle Detection Station, the controller cabinets at a fixed milepost which process incoming loop detector data from across all lanes

VSL: Variable Speed Limit

BIBLIOGRAPHY

Abdel-Aty, M., and F. Abdalla. 2004. Linking roadway geometrics and real-time traffic characteristics to model daytime freeway crashes using generalized estimating equations for correlated data. In Washington, D.C.

Abdel-Aty, Mohamed, Nizam Uddin, Anurag Pande, Fathy Abdalla, and Liang Hsia. 2004. “Predicting Freeway Crashes from Loop Detector Data by Matched Case-Control Logistic Regression.” *Transportation Research Record: Journal of the Transportation Research Board* 1897 (-1) (January 1): 88-95. doi:10.3141/1897-12.

Abdel-Aty, Mohamed, Nizam Uddin, and Anurag Pande. 2005. “Split Models for Predicting Multivehicle Crashes During High-Speed and Low-Speed Operating Conditions on Freeways.” *Transportation Research Record: Journal of the Transportation Research Board* 1908 (-1) (January 1): 51-58. doi:10.3141/1908-07.

Abdel-Aty, Mohamed, and Anurag Pande. 2005. “Identifying crash propensity using specific traffic speed conditions.” *Journal of Safety Research* 36 (1): 97-108. doi:10.1016/j.jsr.2004.11.002.

Abdelwahab, Hassan, and Mohamed Abdel-Aty. 2001. “Development of Artificial Neural Network Models to Predict Driver Injury Severity in Traffic Accidents at Signalized Intersections.” *Transportation Research Record: Journal of the Transportation Research Board* 1746 (-1) (January 1): 6-13. doi:10.3141/1746-02.

———. 2002. “Artificial Neural Networks and Logit Models for Traffic Safety Analysis of Toll Plazas.” *Transportation Research Record: Journal of the Transportation Research Board* 1784 (-1) (January 1): 115-125. doi:10.3141/1784-15.

Abdulhai, Baher, and Stephen G. Ritchie. 1999. “Enhancing the universality and transferability of freeway incident detection using a Bayesian-based neural network.” *Transportation Research Part C: Emerging Technologies* 7 (5) (October): 261-280. doi:10.1016/S0968-090X(99)00022-4.

Al-Deek, H., A. Garib, and A. Radwan. 1998. A new method for estimating freeway incident congestion. Text. http://cat.inist.fr/?aModele=afficheN&cpsidt=2958605.

Al-Deek, H., S. Ishak, and A. Khan. 1996. Impact of freeway geometric and incident characteristics on incident detection. Text. http://cat.inist.fr/?aModele=afficheN&cpsidt=2477103.

Awad, Wael, and Bruce Janson. 1998. “Prediction Models for Truck Accidents at Freeway Ramps in Washington State Using Regression and Artificial Intelligence Techniques.” *Transportation Research Record: Journal of the Transportation Research Board* 1635 (-1) (January 1): 30-36. doi:10.3141/1635-04.

Breiman, L., J. H. Freidman, R. A. Olshen, and C. J. Stone. 1984. *Classification and Regression Trees*. Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software.

Cheu, Ruey L., and Stephen G. Ritchie. 1995. “Automated detection of lane-

blocking freeway incidents using artificial neural networks." *Transportation Research Part C: Emerging Technologies* 3 (6) (December): 371-388. doi:10.1016/0968-090X(95)00016-C.

Collett, D. 1991. *Modelling Binary Data*. Chapman and Hall.

Gayah, V.V., C. Dos Santos, M. Abdel-Aty, A. Dhindsa, and J. Dilmore. 2006. Evaluating ITS strategies for real-time freeway safety improvement. In *Intelligent Transportation Systems Conference, 2006. ITSC '06. IEEE*, 1114-1119. doi:10.1109/ITSC.2006.1707371.

Golob, Thomas F., Wilfred W. Recker, and Veronica M. Alvarez. 2004a. "Tool to Evaluate Safety Effects of Changes in Freeway Traffic Flow." *Journal of Transportation Engineering* 130 (2) (March): 222-230. doi:10.1061/(ASCE)0733-947X(2004)130:2(222).

———. 2004b. "Freeway safety as a function of traffic flow." *Accident Analysis & Prevention* 36 (6) (November): 933-946. doi:10.1016/j.aap.2003.09.006.

Golob, Thomas F., and Wilfred W. Recker. 2003. "Relationships Among Urban Freeway Accidents, Traffic Flow, Weather, and Lighting Conditions." *Journal of Transportation Engineering* 129 (4) (July): 342-353. doi:10.1061/(ASCE)0733-947X(2003)129:4(342).

———. 2004. "A method for relating type of crash to traffic flow characteristics on urban freeways." *Transportation Research Part A: Policy and Practice* 38 (1) (January): 53-80. doi:10.1016/j.tra.2003.08.002.

Hand, D. J., Heikki Mannila, and Padhraic Smyth. 2001. *Principles of data mining*. MIT Press, August 1.

Hosner, D. W., and S. Lemeshow. 1989. *Applied Logistic Regression*. Wiley & Sons.

Hughes, R., and F. Council. 1999. On establishing relationship(s) between freeway safety and peak period operations: Performance measurement and methodological considerations. In Washington, D.C.

Ishak, S., and C. Alecsandru. 2005. Analysis of freeway pre-incident, post-incident, and non-incident conditions using second-order spatio-temporal traffic performance measures. In Washington, D.C.

Ishak, Sherif, and Haitham Al-Deek. 1999. "Performance of Automatic ANN-Based Incident Detection on Freeways." *Journal of Transportation Engineering* 125 (4) (July): 281-290. doi:10.1061/(ASCE)0733-947X(1999)125:4(281).

Kockelman, K., and J. Ma. 2004. Freeway speeds and speed variations preceding crashes, within and across lanes. In Washington, D.C.

Lee, Chris, Bruce Hellinga, and Frank Saccomanno. 2003. "Real-Time Crash Prediction Model for Application to Crash Prevention in Freeway Traffic." *Transportation Research Record: Journal of the Transportation Research Board* 1840 (-1) (January 1): 67-77. doi:10.3141/1840-08.

———. 2004. "Assessing Safety Benefits of Variable Speed Limits." *Transportation Research Record: Journal of the Transportation Research Board* 1897 (-1) (January 1): 183-190. doi:10.3141/1897-24.

Lee, Chris, Frank Saccomanno, and Bruce Hellinga. 2002. "Analysis of Crash Precursors on Instrumented Freeways." *Transportation Research Record:*

Journal of the Transportation Research Board 1784 (-1) (January 1): 1-8. doi:10.3141/1784-01.

Madanat, S., and P. Liu. 1995. *A prototype system for real-time incident likelihood prediction*. IDEA Project Final Report (ITS-2). Washington, D.C.: Transportation Research Board. http://pubsindex.trb.org/view.aspx?id=465172.

Mussone, Lorenzo, Andrea Ferrari, and Marcello Oneta. 1999. "An analysis of urban collisions using an artificial intelligence model." *Accident Analysis & Prevention* 31 (6) (November): 705-718. doi:10.1016/S0001-4575(99)00031-7.

Nezamuddin, N., Nan Jiang, Jianming Ma, Ti Zhang, and S. Travis Waller. 2011. Active Traffic Management Strategies: Implications for freeway operations and traffic safety. In Washington, D.C.

Oh, C., J. Oh, S. Ritchie, and M. Chang. 2001. Real time estimation of freeway accident likelihood. In Washington, D.C.

Pande, Anurag. 2003. Classification of real-time traffic speed patterns to predict crashes on freeways. University of Central Florida.

———. 2005. Applying Hybrid Models for Real-Time Crash Risk Assessment On Freeways. University of Central Florida.

Pande, Anurag, Mohamed Abdel-Aty, and Liang Hsia. 2005. "Spatiotemporal Variation of Risk Preceding Crashes on Freeways." *Transportation Research Record: Journal of the Transportation Research Board* 1908 (-1) (January 1): 26-36. doi:10.3141/1908-04.

Pande, Anurag, and Mohamed Abdel-Aty. 2006a. "Comprehensive Analysis of the Relationship Between Real-Time Traffic Surveillance Data and Rear-End Crashes on Freeways." *Transportation Research Record: Journal of the Transportation Research Board* 1953 (-1) (January 1): 31-40. doi:10.3141/1953-04.

———. 2006b. "Assessment of freeway traffic parameters leading to lane-change related collisions." *Accident Analysis & Prevention* 38 (5) (September): 936-948. doi:10.1016/j.aap.2006.03.004.

———. 2007. "Multiple-Model Framework for Assessment of Real-Time Crash Risk." *Transportation Research Record: Journal of the Transportation Research Board* 2019 (-1) (December 1): 99-107. doi:10.3141/2019-13.

———. 2008. "A Computing Approach Using Probabilistic Neural Networks for Instantaneous Appraisal of Rear-End Crash Risk." *Computer-Aided Civil and Infrastructure Engineering* 23 (7) (October): 549-559. doi:10.1111/j.1467-8667.2008.00559.x.

Park, S., and S. Ritchie. 2004. Exploring the relationship between freeway speed variance, lane changing, and vehicle hetrogenity. In Washington, D.C.

Pham, Minh-Hai, Nour-Eddin El Faouzi, and André-Gilles Dumont. 2011. Real-time identification of risk-prone traffic patterns taking into account weather conditions. In Washington, D.C.

SAS Institute. 2001. *SAS*. SAS Institute.

Sayed, Tarek, and Walid Abdelwahab. 1998. "Comparison of Fuzzy and Neural Classifiers for Road Accidents Analysis." *Journal of Computing in Civil*

Engineering 12 (1) (January): 42-47. doi:10.1061/(ASCE)0887-3801(1998)12:1(42).

Sohn, S., and H. Shin. 2001. “Pattern recognition for road traffic accident severity in Korea.” *Ergonomics* 44 (1): 107-117.

Songchitruksa, Praprut, and Kevin Balke. 2006. “Assessing Weather, Environment, and Loop Data for Real-Time Freeway Incident Prediction.” *Transportation Research Record: Journal of the Transportation Research Board* 1959 (-1) (January 1): 105-113. doi:10.3141/1959-12.

Vorko, Ariana, and Franjo Jovic. 2000. “Multiple attribute entropy classification of school-age injuries.” *Accident Analysis & Prevention* 32 (3) (May): 445-454. doi:10.1016/S0001-4575(99)00069-X.

Xu, Chengcheng, Liu Pan, Wei Wang, and Yu Chunjun. 2011. Exploration and identification of hazardous traffic flow states before crash occurrences on freeways. In Washington, D.C.

Zhang, C., J. Ivan, W. El-Dessouki, and E. Anagnostou. 2005. Relative risk analysis for studying the impact of adverse weather conditions and congestion on traffic accidents. In Washington, D.C.

Zhou, Min, and Virginia Sisiopiku. 1997. “Relationship Between Volume-to-Capacity Ratios and Accident Rates.” *Transportation Research Record: Journal of the Transportation Research Board* 1581 (-1) (January 1): 47-52. doi:10.3141/1581-06.

APPENDIX A: SAMPLE CODE

BUILD MODELS FROM 101 NB CRASH AND NON-CRASH DATA

```
proc logistic data=sas_sjsu.us101nb_crash_nocrash;

where (1<weekday(crashday)<7 and 18000<=crashtime<=79200);

model y(event='1')=AVUS1_0 AVDS1_0SVUS1_0 SVDS1_0 AOUS1_0 AODS1_0
SOUS1_0 SODS1_0

/ selection=stepwise

slentry=0.3

slstay=0.35

details

lackfit;

output out=dayonly.pred1_0_vo p=phat lower=lcl upper=ucl

predprob=(individual crossvalidate);

run;


proc sort data=dayonly.pred1_0_vo;

by descending IP_1;

run;


proc logistic data=sas_sjsu.us101nb_crash_nocrash;

where (1<weekday(crashday)<7 and 18000<=crashtime<=79200);

model y(event='1')=AVUS1_1 AVDS1_1SVUS1_1 SVDS1_1 AOUS1_1 AODS1_1
SOUS1_1 SODS1_1

/ selection=stepwise

slentry=0.3

slstay=0.35

details

lackfit;

output out=dayonly.pred1_1_vo p=phat lower=lcl upper=ucl

predprob=(individual crossvalidate);
```

```
run;


proc sort data=dayonly.pred1_1_vo;
by descending IP_1;
run;


proc logistic data=sas_sjsu.us101nb_crash_nocrash;
where (1<weekday(crashday)<7 and 18000<=crashtime<=79200);
model y(event='1')=AVUS1_2 AVDS1_2SVUS1_2 SVDS1_2 AOUS1_2 AODS1_2
SOUS1_2 SODS1_2
/ selection=stepwise
slentry=0.3
slstay=0.35
details
lackfit;
output out=dayonly.pred1_2_vo p=phat lower=lcl upper=ucl
predprob=(individual crossvalidate);
run;


proc sort data=dayonly.pred1_2_vo;
by descending IP_1;
run;


proc logistic data=sas_sjsu.us101nb_crash_nocrash;
where (1<weekday(crashday)<7 and 18000<=crashtime<=79200);
model y(event='1')=AVUS1_3 AVDS1_3SVUS1_3 SVDS1_3 AOUS1_3 AODS1_3
SOUS1_3 SODS1_3
/ selection=stepwise
slentry=0.3
slstay=0.35
```

```
details
lackfit;
output out=dayonly.pred1_3_vo p=phat lower=lcl upper=ucl
predprob=(individual crossvalidate);
run;

proc sort data=dayonly.pred1_3_vo;
by descending IP_1;
run;



proc logistic data=sas_sjsu.us101nb_crash_nocrash;
where (1<weekday(crashday)<7 and 18000<=crashtime<=79200);
model y(event='1')=AVUS1_4    AVDS1_4SVUS1_4      SVDS1_4 AOUS1_4     AODS1_4
SOUS1_4         SODS1_4
/ selection=stepwise
slentry=0.3
slstay=0.35
details
lackfit;
output out=dayonly.pred1_4_vo p=phat lower=lcl upper=ucl
predprob=(individual crossvalidate);
run;

proc sort data=dayonly.pred1_4_vo;
by descending IP_1;
run;

proc logistic data=sas_sjsu.us101nb_crash_nocrash;
where (1<weekday(crashday)<7 and 18000<=crashtime<=79200);
```

```
model y(event='1')=AVUS1_0     AVDS1_0SVUS1_0       SVDS1_0 AOUS1_0      AODS1_0
SOUS1_0         SODS1_0 AVUS2_0      AVDS2_0 SVUS2_0      SVDS2_0                AOUS2_0
          AODS2_0 SOUS2_0      SODS2_0

/ selection=stepwise

slentry=0.3

slstay=0.35

details

lackfit;

output out=dayonly.pred2_0_vo p=phat lower=lcl upper=ucl

predprob=(individual crossvalidate);

run;


proc sort data=dayonly.pred2_0_vo;

by descending IP_1;

run;


proc logistic data=sas_sjsu.us101nb_crash_nocrash;

where (1<weekday(crashday)<7 and 18000<=crashtime<=79200);

model y(event='1')=AVUS1_1     AVDS1_1SVUS1_1       SVDS1_1 AOUS1_1      AODS1_1
SOUS1_1         SODS1_1 AVUS2_1      AVDS2_1 SVUS2_1      SVDS2_1                AOUS2_1
          AODS2_1 SOUS2_1      SODS2_1

/ selection=stepwise

slentry=0.3

slstay=0.35

details

lackfit;

output out=dayonly.pred2_1_vo p=phat lower=lcl upper=ucl

predprob=(individual crossvalidate);

run;


proc sort data=dayonly.pred2_1_vo;
```

```
by descending IP_1;
run;


proc logistic data=sas_sjsu.us101nb_crash_nocrash;
where (1<weekday(crashday)<7 and 18000<=crashtime<=79200);
model y(event='1')=AVUS1_2 AVDS1_2SVUS1_2 SVDS1_2 AOUS1_2 AODS1_2
SOUS1_2 SODS1_2 AVUS2_2 AVDS2_2 SVUS2_2 SVDS2_2 AOUS2_2
AODS2_2 SOUS2_2 SODS2_2
/ selection=stepwise
slentry=0.3
slstay=0.35
details
lackfit;
output out=dayonly.pred2_2_vo p=phat lower=lcl upper=ucl
predprob=(individual crossvalidate);
run;


proc sort data=dayonly.pred2_2_vo;
by descending IP_1;
run;


proc logistic data=sas_sjsu.us101nb_crash_nocrash;
where (1<weekday(crashday)<7 and 18000<=crashtime<=79200);
model y(event='1')=AVUS1_3 AVDS1_3SVUS1_3 SVDS1_3 AOUS1_3 AODS1_3
SOUS1_3 SODS1_3 AVUS2_3 AVDS2_3 SVUS2_3 SVDS2_3 AOUS2_3
AODS2_3 SOUS2_3 SODS2_3
/ selection=stepwise
slentry=0.3
slstay=0.35
details
lackfit;
```

```
output out=dayonly.pred2_3_vo p=phat lower=lcl upper=ucl
predprob=(individual crossvalidate);
run;

proc sort data=dayonly.pred2_3_vo;
by descending IP_1;
run;

proc logistic data=sas_sjsu.us101nb_crash_nocrash;
where (1<weekday(crashday)<7 and 18000<=crashtime<=79200);
model y(event='1')=AVUS1_4 AVDS1_4SVUS1_4 SVDS1_4 AOUS1_4 AODS1_4
SOUS1_4 SODS1_4 AVUS2_4 AVDS2_4 SVUS2_4 SVDS2_4 AOUS2_4
AODS2_4 SOUS2_4 SODS2_4
/ selection=stepwise
slentry=0.3
slstay=0.35
details
lackfit;
output out=dayonly.pred2_4_vo p=phat lower=lcl upper=ucl
predprob=(individual crossvalidate);
run;

proc sort data=dayonly.pred2_4_vo;
by descending IP_1;
run;

proc logistic data=sas_sjsu.us101nb_crash_nocrash;
where (1<weekday(crashday)<7 and 18000<=crashtime<=79200);
  model y(event='1')=AVUS1_0 AVDS1_0 SVUS1_0 SVDS1_0 AOUS1_0 AODS1_0
SOUS1_0 SODS1_0 AVUS2_0 AVDS2_0 SVUS2_0 SVDS2_0 AOUS2_0
AODS2_0 SOUS2_0 SODS2_0 AVUS3_0 AVDS3_0 SVUS3_0 SVDS3_0
AOUS3_0 AODS3_0 SOUS3_0 SODS3_0
```

```
              / selection=stepwise
slentry=0.3
slstay=0.35
details
lackfit;
output out=dayonly.pred3_0_vo p=phat lower=lcl upper=ucl
predprob=(individual crossvalidate);
run;


proc sort data=dayonly.pred3_0_vo;
by descending IP_1;
run;


proc logistic data=sas_sjsu.us101nb_crash_nocrash;
where (1<weekday(crashday)<7 and 18000<=crashtime<=79200);
  model y(event='1')=AVUS1_1  AVDS1_1 SVUS1_1      SVDS1_1 AOUS1_1      AODS1_1
SOUS1_1         SODS1_1 AVUS2_1      AVDS2_1 SVUS2_1      SVDS2_1                AOUS2_1
          AODS2_1 SOUS2_1      SODS2_1 AVUS3_1      AVDS3_1 SVUS3_1      SVDS3_1
AOUS3_1         AODS3_1 SOUS3_1      SODS3_1
              / selection=stepwise
slentry=0.3
slstay=0.35
details
lackfit;
output out=dayonly.pred3_1_vo p=phat lower=lcl upper=ucl
predprob=(individual crossvalidate);
run;


proc sort data=dayonly.pred3_1_vo;
by descending IP_1;
run;
```

```
proc logistic data=sas_sjsu.us101nb_crash_nocrash;

where (1<weekday(crashday)<7 and 18000<=crashtime<=79200);

  model y(event='1')=AVUS1_2  AVDS1_2 SVUS1_2      SVDS1_2 AOUS1_2      AODS1_2
SOUS1_2         SODS1_2 AVUS2_2      AVDS2_2 SVUS2_2      SVDS2_2                AOUS2_2
          AODS2_2 SOUS2_2      SODS2_2 AVUS3_2      AVDS3_2 SVUS3_2      SVDS3_2
AOUS3_2         AODS3_2 SOUS3_2      SODS3_2

            / selection=stepwise

slentry=0.3

slstay=0.35

details

lackfit;

output out=dayonly.pred3_2_vo p=phat lower=lcl upper=ucl

predprob=(individual crossvalidate);

run;


proc sort data=dayonly.pred3_2_vo;

by descending IP_1;

run;


proc logistic data=sas_sjsu.us101nb_crash_nocrash;

where (1<weekday(crashday)<7 and 18000<=crashtime<=79200);

  model y(event='1')=AVUS1_3  AVDS1_3 SVUS1_3      SVDS1_3 AOUS1_3      AODS1_3
SOUS1_3         SODS1_3 AVUS2_3      AVDS2_3 SVUS2_3      SVDS2_3                AOUS2_3
          AODS2_3 SOUS2_3      SODS2_3 AVUS3_3      AVDS3_3 SVUS3_3      SVDS3_3
AOUS3_3         AODS3_3 SOUS3_3      SODS3_3

            / selection=stepwise

slentry=0.3

slstay=0.35

details

lackfit;

output out=dayonly.pred3_3_vo p=phat lower=lcl upper=ucl
```

```
predprob=(individual crossvalidate);
run;

proc sort data=dayonly.pred3_3_vo;
by descending IP_1;
run;

proc logistic data=sas_sjsu.us101nb_crash_nocrash;
where (1<weekday(crashday)<7 and 18000<=crashtime<=79200);
  model y(event='1')=AVUS1_4  AVDS1_4 SVUS1_4      SVDS1_4 AOUS1_4      AODS1_4
SOUS1_4         SODS1_4 AVUS2_4      AVDS2_4 SVUS2_4      SVDS2_4                AOUS2_4
          AODS2_4 SOUS2_4      SODS2_4 AVUS3_4      AVDS3_4 SVUS3_4      SVDS3_4
AOUS3_4         AODS3_4 SOUS3_4      SODS3_4
             / selection=stepwise
slentry=0.3
slstay=0.35
details
lackfit;
output out=dayonly.pred3_4_vo p=phat lower=lcl upper=ucl
predprob=(individual crossvalidate);
run;

proc sort data=dayonly.pred3_4_vo;
by descending IP_1;
run;

proc logistic data=sas_sjsu.us101nb_crash_nocrash;
where (1<weekday(crashday)<7 and 18000<=crashtime<=79200);
model y(event='1')=ASUS1_0     ASDS1_0SSUS1_0       SSDS1_0
/ selection=stepwise
slentry=0.3
```

```
slstay=0.35
details
lackfit;
output out=dayonly.pred1_0_s p=phat lower=lcl upper=ucl
predprob=(individual crossvalidate);
run;

proc sort data=dayonly.pred1_0_s;
by descending IP_1;
run;

proc logistic data=sas_sjsu.us101nb_crash_nocrash;
where (1<weekday(crashday)<7 and 18000<=crashtime<=79200);
model y(event='1')=ASUS1_1     ASDS1_1SSUS1_1       SSDS1_1
/ selection=stepwise
slentry=0.3
slstay=0.35
details
lackfit;
output out=dayonly.pred1_1_s p=phat lower=lcl upper=ucl
predprob=(individual crossvalidate);
run;

proc sort data=dayonly.pred1_1_s;
by descending IP_1;
run;

proc logistic data=sas_sjsu.us101nb_crash_nocrash;
where (1<weekday(crashday)<7 and 18000<=crashtime<=79200);
model y(event='1')=ASUS1_2     ASDS1_2SSUS1_2       SSDS1_2
```

```
/ selection=stepwise
slentry=0.3
slstay=0.35
details
lackfit;
output out=dayonly.pred1_2_s p=phat lower=lcl upper=ucl
predprob=(individual crossvalidate);
run;

proc sort data=dayonly.pred1_2_s;
by descending IP_1;
run;

proc logistic data=sas_sjsu.us101nb_crash_nocrash;
where (1<weekday(crashday)<7 and 18000<=crashtime<=79200);
model y(event='1')=ASUS1_3     ASDS1_3SSUS1_3       SSDS1_3
/ selection=stepwise
slentry=0.3
slstay=0.35
details
lackfit;
output out=dayonly.pred1_3_s p=phat lower=lcl upper=ucl
predprob=(individual crossvalidate);
run;

proc sort data=dayonly.pred1_3_s;
by descending IP_1;
run;

proc logistic data=sas_sjsu.us101nb_crash_nocrash;
```

```
where (1<weekday(crashday)<7 and 18000<=crashtime<=79200);
model y(event='1')=ASUS1_4     ASDS1_4SSUS1_4       SSDS1_4
/ selection=stepwise
slentry=0.3
slstay=0.35
details
lackfit;
output out=dayonly.pred1_4_s p=phat lower=lcl upper=ucl
predprob=(individual crossvalidate);
run;

proc sort data=dayonly.pred1_4_s;
by descending IP_1;
run;

proc logistic data=sas_sjsu.us101nb_crash_nocrash;
where (1<weekday(crashday)<7 and 18000<=crashtime<=79200);
model y(event='1')=ASUS1_0     ASDS1_0SSUS1_0       SSDS1_0 ASUS2_0      ASDS2_0
SSUS2_0          SSDS2_0
/ selection=stepwise
slentry=0.3
slstay=0.35
details
lackfit;
output out=dayonly.pred2_0_s p=phat lower=lcl upper=ucl
predprob=(individual crossvalidate);
run;

proc sort data=dayonly.pred2_0_s;
by descending IP_1;
```

```
run;


proc logistic data=sas_sjsu.us101nb_crash_nocrash;
where (1<weekday(crashday)<7 and 18000<=crashtime<=79200);
model y(event='1')=ASUS1_1 ASDS1_1SSUS1_1 SSDS1_1 ASUS2_1 ASDS2_1
SSUS2_1 SSDS2_1
/ selection=stepwise
slentry=0.3
slstay=0.35
details
lackfit;
output out=dayonly.pred2_1_s p=phat lower=lcl upper=ucl
predprob=(individual crossvalidate);
run;



proc sort data=dayonly.pred2_1_s;
by descending IP_1;
run;


proc logistic data=sas_sjsu.us101nb_crash_nocrash;
where (1<weekday(crashday)<7 and 18000<=crashtime<=79200);
model y(event='1')=ASUS1_2 ASDS1_2SSUS1_2 SSDS1_2 ASUS2_2 ASDS2_2
SSUS2_2 SSDS2_2
/ selection=stepwise
slentry=0.3
slstay=0.35
details
lackfit;
output out=dayonly.pred2_2_s p=phat lower=lcl upper=ucl
```

```
predprob=(individual crossvalidate);

run;


proc sort data=dayonly.pred2_2_s;

by descending IP_1;

run;


proc logistic data=sas_sjsu.us101nb_crash_nocrash;

where (1<weekday(crashday)<7 and 18000<=crashtime<=79200);

model y(event='1')=ASUS1_3    ASDS1_3SSUS1_3      SSDS1_3 ASUS2_3     ASDS2_3
SSUS2_3         SSDS2_3

/ selection=stepwise

slentry=0.3

slstay=0.35

details

lackfit;

output out=dayonly.pred2_3_s p=phat lower=lcl upper=ucl

predprob=(individual crossvalidate);

run;


proc sort data=dayonly.pred2_3_s;

by descending IP_1;

run;


proc logistic data=sas_sjsu.us101nb_crash_nocrash;

where (1<weekday(crashday)<7 and 18000<=crashtime<=79200);

model y(event='1')=ASUS1_4    ASDS1_4SSUS1_4      SSDS1_4 ASUS2_4     ASDS2_4
SSUS2_4         SSDS2_4

/ selection=stepwise

slentry=0.3
```

```
slstay=0.35
details
lackfit;
output out=dayonly.pred2_4_s p=phat lower=lcl upper=ucl
predprob=(individual crossvalidate);
run;

proc sort data=dayonly.pred2_4_s;
by descending IP_1;
run;

proc logistic data=sas_sjsu.us101nb_crash_nocrash;
where (1<weekday(crashday)<7 and 18000<=crashtime<=79200);
model y(event='1')=ASUS1_0 ASDS1_0SSUS1_0 SSDS1_0 ASUS2_0 ASDS2_0
SSUS2_0 SSDS2_0 ASUS3_0 ASDS3_0 SSUS3_0 SSDS3_0
/ selection=stepwise
slentry=0.3
slstay=0.35
details
lackfit;
output out=dayonly.pred3_0_s p=phat lower=lcl upper=ucl
predprob=(individual crossvalidate);
run;

proc sort data=dayonly.pred3_0_s;
by descending IP_1;
run;

proc logistic data=sas_sjsu.us101nb_crash_nocrash;
where (1<weekday(crashday)<7 and 18000<=crashtime<=79200);
```

```
model y(event='1')=ASUS1_1    ASDS1_1SSUS1_1      SSDS1_1 ASUS2_1     ASDS2_1
SSUS2_1         SSDS2_1 ASUS3_1     ASDS3_1 SSUS3_1     SSDS3_1

/ selection=stepwise

slentry=0.3

slstay=0.35

details

lackfit;

output out=dayonly.pred3_1_s p=phat lower=lcl upper=ucl

predprob=(individual crossvalidate);

run;


proc sort data=dayonly.pred3_1_s;

by descending IP_1;

run;


proc logistic data=sas_sjsu.us101nb_crash_nocrash;

where (1<weekday(crashday)<7 and 18000<=crashtime<=79200);

model y(event='1')=ASUS1_2    ASDS1_2SSUS1_2      SSDS1_2 ASUS2_2     ASDS2_2
SSUS2_2         SSDS2_2 ASUS3_2     ASDS3_2 SSUS3_2     SSDS3_2

/ selection=stepwise

slentry=0.3

slstay=0.35

details

lackfit;

output out=dayonly.pred3_2_s p=phat lower=lcl upper=ucl

predprob=(individual crossvalidate);

run;


proc sort data=dayonly.pred3_2_s;

by descending IP_1;
```

```
run;

proc logistic data=sas_sjsu.us101nb_crash_nocrash;
where (1<weekday(crashday)<7 and 18000<=crashtime<=79200);
model y(event='1')=ASUS1_3 ASDS1_3SSUS1_3 SSDS1_3 ASUS2_3 ASDS2_3
SSUS2_3 SSDS2_3 ASUS3_3 ASDS3_3 SSUS3_3 SSDS3_3
/ selection=stepwise
slentry=0.3
slstay=0.35
details
lackfit;
output out=dayonly.pred3_3_s p=phat lower=lcl upper=ucl
predprob=(individual crossvalidate);
run;

proc sort data=dayonly.pred3_3_s;
by descending IP_1;
run;

proc logistic data=sas_sjsu.us101nb_crash_nocrash;
where (1<weekday(crashday)<7 and 18000<=crashtime<=79200);
model y(event='1')=ASUS1_4 ASDS1_4SSUS1_4 SSDS1_4 ASUS2_4 ASDS2_4
SSUS2_4 SSDS2_4 ASUS3_4 ASDS3_4 SSUS3_4 SSDS3_4
/ selection=stepwise
slentry=0.3
slstay=0.35
details
lackfit;
output out=dayonly.pred3_4_s p=phat lower=lcl upper=ucl
predprob=(individual crossvalidate);
```

```
run;


proc sort data=dayonly.pred3_4_s;
by descending IP_1;
run;
```

COMPARE MODELS TO FIND BEST THREE

```
%inc "E:\code\gainlift_mac.sas";

ods graphics on;


%GainLift(data=dayonly.pred1_0_vo, groups=10, oneplot=CCAPT , response=y, p=IP_1, event=1,out=dayonly.pctile_pred1_0_vo);

datadayonly.pctile_pred1_0_vo; set dayonly.pctile_pred1_0_vo; modelname='pred1_0_vo'; run;


%GainLift(data=dayonly.pred1_1_vo, groups=10, oneplot=CCAPT , response=y, p=IP_1, event=1,out=dayonly.pctile_pred1_1_vo);

datadayonly.pctile_pred1_1_vo; set dayonly.pctile_pred1_1_vo; modelname='pred1_1_vo'; run;


%GainLift(data=dayonly.pred1_2_vo, groups=10, oneplot=CCAPT , response=y, p=IP_1, event=1,out=dayonly.pctile_pred1_2_vo);

datadayonly.pctile_pred1_2_vo; set dayonly.pctile_pred1_2_vo; modelname='pred1_2_vo'; run;


%GainLift(data=dayonly.pred1_3_vo, groups=10, oneplot=CCAPT , response=y, p=IP_1, event=1,out=dayonly.pctile_pred1_3_vo);

datadayonly.pctile_pred1_3_vo; set dayonly.pctile_pred1_3_vo; modelname='pred1_3_vo'; run;


%GainLift(data=dayonly.pred1_4_vo, groups=10, oneplot=CCAPT , response=y, p=IP_1, event=1,out=dayonly.pctile_pred1_4_vo);

datadayonly.pctile_pred1_4_vo; set dayonly.pctile_pred1_4_vo; modelname='pred1_4_vo'; run;


%GainLift(data=dayonly.pred2_0_vo, groups=10, oneplot=CCAPT , response=y, p=IP_1, event=1,out=dayonly.pctile_pred2_0_vo);

datadayonly.pctile_pred2_0_vo; set dayonly.pctile_pred2_0_vo; modelname='pred2_0_vo'; run;


%GainLift(data=dayonly.pred2_1_vo, groups=10, oneplot=CCAPT , response=y, p=IP_1, event=1,out=dayonly.pctile_pred2_1_vo);

datadayonly.pctile_pred2_1_vo; set dayonly.pctile_pred2_1_vo; modelname='pred2_1_vo'; run;
```

%GainLift(data=dayonly.pred2_2_vo, groups=10, oneplot=CCAPT , response=y, p=IP_1, event=1,out=dayonly.pctile_pred2_2_vo);

datadayonly.pctile_pred2_2_vo; set dayonly.pctile_pred2_2_vo; modelname='pred2_2_vo'; run;

%GainLift(data=dayonly.pred2_3_vo, groups=10, oneplot=CCAPT , response=y, p=IP_1, event=1,out=dayonly.pctile_pred2_3_vo);

datadayonly.pctile_pred2_3_vo; set dayonly.pctile_pred2_3_vo; modelname='pred2_3_vo'; run;

%GainLift(data=dayonly.pred2_4_vo, groups=10, oneplot=CCAPT , response=y, p=IP_1, event=1,out=dayonly.pctile_pred2_4_vo);

datadayonly.pctile_pred2_4_vo; set dayonly.pctile_pred2_4_vo; modelname='pred2_4_vo'; run;

%GainLift(data=dayonly.pred3_0_vo, groups=10, oneplot=CCAPT , response=y, p=IP_1, event=1,out=dayonly.pctile_pred3_0_vo);

datadayonly.pctile_pred3_0_vo; set dayonly.pctile_pred3_0_vo; modelname='pred3_0_vo'; run;

%GainLift(data=dayonly.pred3_1_vo, groups=10, oneplot=CCAPT , response=y, p=IP_1, event=1,out=dayonly.pctile_pred3_1_vo);

datadayonly.pctile_pred3_1_vo; set dayonly.pctile_pred3_1_vo; modelname='pred3_1_vo'; run;

%GainLift(data=dayonly.pred3_2_vo, groups=10, oneplot=CCAPT , response=y, p=IP_1, event=1,out=dayonly.pctile_pred3_2_vo);

datadayonly.pctile_pred3_2_vo; set dayonly.pctile_pred3_2_vo; modelname='pred3_2_vo'; run;

%GainLift(data=dayonly.pred3_3_vo, groups=10, oneplot=CCAPT , response=y, p=IP_1, event=1,out=dayonly.pctile_pred3_3_vo);

datadayonly.pctile_pred3_3_vo; set dayonly.pctile_pred3_3_vo; modelname='pred3_3_vo'; run;

%GainLift(data=dayonly.pred3_4_vo, groups=10, oneplot=CCAPT , response=y, p=IP_1, event=1,out=dayonly.pctile_pred3_4_vo);

datadayonly.pctile_pred3_4_vo; set dayonly.pctile_pred3_4_vo; modelname='pred3_4_vo'; run;

%GainLift(data=dayonly.pred1_0_s, groups=10, oneplot=CCAPT , response=y, p=IP_1,

event=1,out=dayonly.pctile_pred1_0_s);

datadayonly.pctile_pred1_0_s; set dayonly.pctile_pred1_0_s; modelname='pred1_0_s'; run;

%GainLift(data=dayonly.pred1_1_s, groups=10, oneplot=CCAPT , response=y, p=IP_1, event=1,out=dayonly.pctile_pred1_1_s);

datadayonly.pctile_pred1_1_s; set dayonly.pctile_pred1_1_s; modelname='pred1_1_s'; run;

%GainLift(data=dayonly.pred1_2_s, groups=10, oneplot=CCAPT , response=y, p=IP_1, event=1,out=dayonly.pctile_pred1_2_s);

datadayonly.pctile_pred1_2_s; set dayonly.pctile_pred1_2_s; modelname='pred1_2_s'; run;

%GainLift(data=dayonly.pred1_3_s, groups=10, oneplot=CCAPT , response=y, p=IP_1, event=1,out=dayonly.pctile_pred1_3_s);

datadayonly.pctile_pred1_3_s; set dayonly.pctile_pred1_3_s; modelname='pred1_3_s'; run;

%GainLift(data=dayonly.pred1_4_s, groups=10, oneplot=CCAPT , response=y, p=IP_1, event=1,out=dayonly.pctile_pred1_4_s);

datadayonly.pctile_pred1_4_s; set dayonly.pctile_pred1_4_s; modelname='pred1_4_s'; run;

%GainLift(data=dayonly.pred2_0_s, groups=10, oneplot=CCAPT , response=y, p=IP_1, event=1,out=dayonly.pctile_pred2_0_s);

datadayonly.pctile_pred2_0_s; set dayonly.pctile_pred2_0_s; modelname='pred2_0_s'; run;

%GainLift(data=dayonly.pred2_1_s, groups=10, oneplot=CCAPT , response=y, p=IP_1, event=1,out=dayonly.pctile_pred2_1_s);

datadayonly.pctile_pred2_1_s; set dayonly.pctile_pred2_1_s; modelname='pred2_1_s'; run;

%GainLift(data=dayonly.pred2_2_s, groups=10, oneplot=CCAPT , response=y, p=IP_1, event=1,out=dayonly.pctile_pred2_2_s);

datadayonly.pctile_pred2_2_s; set dayonly.pctile_pred2_2_s; modelname='pred2_2_s'; run;

%GainLift(data=dayonly.pred2_3_s, groups=10, oneplot=CCAPT , response=y, p=IP_1, event=1,out=dayonly.pctile_pred2_3_s);

datadayonly.pctile_pred2_3_s; set dayonly.pctile_pred2_3_s; modelname='pred2_3_s'; run;

%GainLift(data=dayonly.pred2_4_s, groups=10, oneplot=CCAPT , response=y, p=IP_1, event=1,out=dayonly.pctile_pred2_4_s);

datadayonly.pctile_pred2_4_s; set dayonly.pctile_pred2_4_s; modelname='pred2_4_s'; run;

%GainLift(data=dayonly.pred3_0_s, groups=10, oneplot=CCAPT , response=y, p=IP_1, event=1,out=dayonly.pctile_pred3_0_s);

datadayonly.pctile_pred3_0_s; set dayonly.pctile_pred3_0_s; modelname='pred3_0_s'; run;

%GainLift(data=dayonly.pred3_1_s, groups=10, oneplot=CCAPT , response=y, p=IP_1, event=1,out=dayonly.pctile_pred3_1_s);

datadayonly.pctile_pred3_1_s; set dayonly.pctile_pred3_1_s; modelname='pred3_1_s'; run;

%GainLift(data=dayonly.pred3_2_s, groups=10, oneplot=CCAPT , response=y, p=IP_1, event=1,out=dayonly.pctile_pred3_2_s);

datadayonly.pctile_pred3_2_s; set dayonly.pctile_pred3_2_s; modelname='pred3_2_s'; run;

%GainLift(data=dayonly.pred3_3_s, groups=10, oneplot=CCAPT , response=y, p=IP_1, event=1,out=dayonly.pctile_pred3_3_s);

datadayonly.pctile_pred3_3_s; set dayonly.pctile_pred3_3_s; modelname='pred3_3_s'; run;

%GainLift(data=dayonly.pred3_4_s, groups=10, oneplot=CCAPT , response=y, p=IP_1, event=1,out=dayonly.pctile_pred3_4_s);

datadayonly.pctile_pred3_4_s; set dayonly.pctile_pred3_4_s; modelname='pred3_4_s'; run;

datadayonly.final_compare;

set dayonly.pctile_pred1_0_s

dayonly.pctile_pred1_1_s

dayonly.pctile_pred1_2_s

dayonly.pctile_pred1_3_s

dayonly.pctile_pred1_4_s

```
dayonly.pctile_pred2_0_s
dayonly.pctile_pred2_1_s
dayonly.pctile_pred2_2_s
dayonly.pctile_pred2_3_s
dayonly.pctile_pred2_4_s
dayonly.pctile_pred3_0_s
dayonly.pctile_pred3_1_s
dayonly.pctile_pred3_2_s
dayonly.pctile_pred3_3_s
dayonly.pctile_pred3_4_s
dayonly.pctile_pred1_0_vo
dayonly.pctile_pred1_1_vo
dayonly.pctile_pred1_2_vo
dayonly.pctile_pred1_3_vo
dayonly.pctile_pred1_4_vo
dayonly.pctile_pred2_0_vo
dayonly.pctile_pred2_1_vo
dayonly.pctile_pred2_2_vo
dayonly.pctile_pred2_3_vo
dayonly.pctile_pred2_4_vo
dayonly.pctile_pred3_0_vo
dayonly.pctile_pred3_1_vo
dayonly.pctile_pred3_2_vo
dayonly.pctile_pred3_3_vo
dayonly.pctile_pred3_4_vo;
run;

procgplot data=dayonly.final_compare;
whereSelectedPct=30;
plotCumPctCaptured*modelname;
```

```
run;
```

SCORING US-101 SB AND I-880 DATA FOR BEST 1 VDS MODEL

```
proc logistic data=SAS_SJSU.us101nb_crash_nocrashoutmodel=results2.pred1_4_vo_model;

where (1<weekday(crashday)<7 and 18000<=crashtime<=79200);

model y(event='1')=AVUS1_4 AVDS1_4SVUS1_4 SVDS1_4 AOUS1_4 AODS1_4
SOUS1_4 SODS1_4

/ selection=stepwise

slentry=0.3

slstay=0.35

details

lackfit;

output out=results2.pred1_4_vo p=phat lower=lcl upper=ucl

predprob=(individual crossvalidate);

run;


/*pred1_2 name convention for the input to the model*/


proc logistic inmodel=results2.pred1_4_vo_model;

score data=sas_sjsu.crash_nocrash_us101sb out=results2.us101sb_pred1_4_vo;

run;


proc logistic inmodel=results2.pred1_4_vo_model;

score data=sas_sjsu.crash_nocrash_880nb out=results2.i880nb_pred1_4_vo;

run;


proc logistic inmodel=results2.pred1_4_vo_model;

score data=sas_sjsu.crash_nocrash_880sb out=results2.i880sb_pred1_4_vo;

run;
```

COMPARING BEST MODELS FOR EACH DATASET

```
%inc "E:\code\gainlift_mac.sas";

ods graphics on;


%GainLift(data=results2.us101sb_pred1_4_vo, groups=10, oneplot=CCAPT , response=y, p=P_1, event=1,out=results2.pctile_us101sb_pred1_4_vo);

dataresults2.pctile_us101sb_pred1_4_vo; set results2.pctile_us101sb_pred1_4_vo; modelname='us101sb_pred1_4_vo'; run;


%GainLift(data=results2.i880nb_pred1_4_vo, groups=10, oneplot=CCAPT , response=y, p=P_1, event=1,out=results2.pctile_i880nb_pred1_4_vo);

dataresults2.pctile_i880nb_pred1_4_vo; set results2.pctile_i880nb_pred1_4_vo; modelname='i880nb_pred1_4_vo'; run;


%GainLift(data=results2.i880sb_pred1_4_vo, groups=10, oneplot=CCAPT , response=y, p=P_1, event=1,out=results2.pctile_i880sb_pred1_4_vo);

dataresults2.pctile_i880sb_pred1_4_vo; set results2.pctile_i880sb_pred1_4_vo; modelname='i880sb_pred1_4_vo'; run;


%GainLift(data=results2.us101sb_pred2_1_vo, groups=10, oneplot=CCAPT , response=y, p=P_1, event=1,out=results2.pctile_us101sb_pred2_1_vo);

dataresults2.pctile_us101sb_pred2_1_vo; set results2.pctile_us101sb_pred2_1_vo; modelname='us101sb_pred2_1_vo'; run;


%GainLift(data=results2.i880nb_pred2_1_vo, groups=10, oneplot=CCAPT , response=y, p=P_1, event=1,out=results2.pctile_i880nb_pred2_1_vo);

dataresults2.pctile_i880nb_pred2_1_vo; set results2.pctile_i880nb_pred2_1_vo; modelname='i880nb_pred2_1_vo'; run;


%GainLift(data=results2.i880sb_pred2_1_vo, groups=10, oneplot=CCAPT , response=y, p=P_1, event=1,out=results2.pctile_i880sb_pred2_1_vo);

dataresults2.pctile_i880sb_pred2_1_vo; set results2.pctile_i880sb_pred2_1_vo; modelname='i880sb_pred2_1_vo'; run;
```

```
%GainLift(data=results2.us101sb_pred3_1_vo, groups=10, oneplot=CCAPT , response=y, p=P_1, event=1,out=results2.pctile_us101sb_pred3_1_vo);

dataresults2.pctile_us101sb_pred3_1_vo; set results2.pctile_us101sb_pred3_1_vo; modelname='us101sb_pred3_1_vo'; run;


%GainLift(data=results2.i880nb_pred3_1_vo, groups=10, oneplot=CCAPT , response=y, p=P_1, event=1,out=results2.pctile_i880nb_pred3_1_vo);

dataresults2.pctile_i880nb_pred3_1_vo; set results2.pctile_i880nb_pred3_1_vo; modelname='i880nb_pred3_1_vo'; run;


%GainLift(data=results2.i880sb_pred3_1_vo, groups=10, oneplot=CCAPT , response=y, p=P_1, event=1,out=results2.pctile_i880sb_pred3_1_vo);

dataresults2.pctile_i880sb_pred3_1_vo; set results2.pctile_i880sb_pred3_1_vo; modelname='i880sb_pred3_1_vo'; run;


data results2.final_compare_3_best;

set results2.pctile_us101sb_pred1_4_vo

results2.pctile_i880nb_pred1_4_vo

results2.pctile_i880sb_pred1_4_vo

results2.pctile_us101sb_pred2_1_vo

results2.pctile_i880nb_pred2_1_vo

results2.pctile_i880sb_pred2_1_vo

results2.pctile_us101sb_pred3_1_vo

results2.pctile_i880nb_pred3_1_vo

results2.pctile_i880sb_pred3_1_vo;

run;


procgplot data=results2.final_compare_3_best;

whereSelectedPct=30;

plotCumPctCaptured*modelname;

run;
```

BEST CLASSIFICATION TREE MODEL RULES

IF ASDS2_3 < 13.64

AND 27.750069233 <= SSDS2_3

THEN

NODE : 6

N : 29

0 : 79.3%

1 : 20.7%

IF 34.787389945<= SSDS1_3

AND 13.64 <= ASDS2_3

AND 27.750069233 <= SSDS2_3

THEN

NODE : 13

N : 590

0 : 98.8%

1 : 1.2%

IF 14.902777778<= ASUS2_3

AND SSDS2_3 < 3.2564497325

AND ASDS2_3 < 62.4125

THEN

NODE : 15

N : 644

0 : 94.6%

1 : 5.4%

IF 62.4125 <= ASDS2_3 < 76.825

AND SSDS2_3 < 5.2618221829

THEN

NODE : 18

N : 123

0 : 85.4%

1 : 14.6%

IF 76.825 <= ASDS2_3

AND SSDS2_3 < 5.2618221829

THEN

NODE : 19

N : 85

0 : 65.9%

1 : 34.1%

IF 46.647222222<= ASUS2_3

AND SSDS1_3 < 34.787389945

AND 13.64 <= ASDS2_3

AND 27.750069233 <= SSDS2_3

THEN

NODE : 23

N : 1017

0 : 95.8%

1 : 4.2%

IF SSUS1_3 < 30.757920412

AND ASUS2_3 < 14.902777778

AND SSDS2_3 < 3.2564497325

AND ASDS2_3 < 62.4125

THEN

NODE : 26

N : 191

0 : 77.0%

1 : 23.0%

IF 30.757920412<= SSUS1_3

AND ASUS2_3 < 14.902777778

AND SSDS2_3 < 3.2564497325

AND ASDS2_3 < 62.4125

THEN

NODE : 27

N : 120

0 : 92.5%

1 : 7.5%

IF 51.675 <= ASDS2_3 < 62.4125

AND 21.013899321 <= SSDS2_3 < 27.750069233

THEN

NODE : 33

N : 327

0 : 90.5%

1 : 9.5%

IF SSDS1_3 < 6.7492863341

AND 16.562533892 <= SSUS1_3

AND 5.2618221829 <= SSDS2_3 < 27.750069233

AND 62.4125 <= ASDS2_3

THEN

NODE : 38

N : 252

0 : 99.6%

1 : 0.4%

IF ASDS1_3 < 34.65

AND ASUS2_3 < 46.647222222

AND SSDS1_3 < 34.787389945

AND 13.64 <= ASDS2_3

AND 27.750069233 <= SSDS2_3

THEN

NODE : 40

N : 548

0 : 94.0%

1 : 6.0%

IF ASDS2_3 < 11.6375

AND ASUS1_3 < 31.6375

AND 3.2564497325 <= SSDS2_3 < 21.013899321

THEN

NODE : 50

N : 41

0 : 78.0%

1 : 22.0%

IF 11.6375 <= ASDS2_3 < 62.4125

AND ASUS1_3 < 31.6375

AND 3.2564497325 <= SSDS2_3 < 21.013899321

THEN

NODE : 51

N : 125

0 : 48.0%

1 : 52.0%

IF ASDS1_3 < 25.185

AND 31.6375 <= ASUS1_3

AND 3.2564497325 <= SSDS2_3 < 21.013899321

AND ASDS2_3 < 62.4125

THEN

NODE : 52

N : 67

0 : 89.6%

1 : 10.4%

IF 25.185 <= ASDS1_3

AND 31.6375 <= ASUS1_3

AND 3.2564497325 <= SSDS2_3 < 21.013899321

AND ASDS2_3 < 62.4125

THEN

NODE : 53

N : 228

0 : 71.9%

1 : 28.1%

IF ASDS2_3 < 20.2625

AND 21.013899321 <= SSDS2_3 < 27.750069233

THEN

NODE : 54

N : 63

0 : 95.2%

1 : 4.8%

IF 20.2625 <= ASDS2_3 < 51.675

AND 21.013899321 <= SSDS2_3 < 27.750069233

THEN

NODE : 55

N : 124

0 : 71.8%

1 : 28.2%

IF SSUS1_3 < 5.4405194116

AND ASUS1_3 < 66.672222222

AND 5.2618221829 <= SSDS2_3 < 27.750069233

AND 62.4125 <= ASDS2_3

THEN

NODE : 58

N : 449

0 : 89.5%

1 : 10.5%

IF 5.4405194116<= SSUS1_3 < 16.562533892

AND ASUS1_3 < 66.672222222

AND 5.2618221829 <= SSDS2_3 < 27.750069233

AND 62.4125 <= ASDS2_3

THEN

NODE : 59

N : 339

0 : 77.9%

1 : 22.1%

IF SSDS1_3 < 20.613077874

AND 66.672222222 <= ASUS1_3

AND SSUS1_3 < 16.562533892

AND 5.2618221829 <= SSDS2_3 < 27.750069233

AND 62.4125 <= ASDS2_3

THEN

NODE : 60

N : 1586

0 : 93.4%

1 : 6.6%

IF 20.613077874<= SSDS1_3

AND 66.672222222 <= ASUS1_3

AND SSUS1_3 < 16.562533892

AND 5.2618221829 <= SSDS2_3 < 27.750069233

AND 62.4125 <= ASDS2_3

THEN

NODE : 61

N : 433

0 : 86.4%

1 : 13.6%

IF 5.2618221829<= SSDS2_3 < 5.7513568455

AND 6.7492863341 <= SSDS1_3

AND 16.562533892 <= SSUS1_3

AND 62.4125 <= ASDS2_3

THEN

NODE : 64

N : 20

0 : 80.0%

1 : 20.0%

IF 5.7513568455<= SSDS2_3 < 27.750069233

AND 6.7492863341 <= SSDS1_3

AND 16.562533892 <= SSUS1_3

AND 62.4125 <= ASDS2_3

THEN

NODE : 65

N : 1074

0 : 95.0%

1 : 5.0%

IF 13.64 <= ASDS2_3 < 39.7875

AND 34.65 <= ASDS1_3

AND ASUS2_3 < 46.647222222

AND SSDS1_3 < 34.787389945

AND 27.750069233 <= SSDS2_3

THEN

NODE : 68

N : 66

0 : 80.3%

1 : 19.7%

IF 39.7875 <= ASDS2_3

AND 34.65 <= ASDS1_3

AND ASUS2_3 < 46.647222222

AND SSDS1_3 < 34.787389945

AND 27.750069233 <= SSDS2_3

THEN

NODE : 69

N : 85

0 : 94.1%

1 : 5.9%