# Influence of Non-Behavioral Feautures on Fatal Car Accidents Due to Cell Phone Usage

Madeline M. Warndorf
Graduate Student, EMSE Dept.
George Washington University
Washington, D.C.
mwarndorf65@gwu.edu

***Abstract*— Fatal car accidents related to drivers being distracted by their cell phones can be prevented. This paper presents the analysis of how different non-behavioral variables influence if a fatal accident is related to the use of cell phones by the driver. This paper does not cover any preventative solutions to decrease the number of fatalities due to cell phone usage. This paper focuses on identifying the best supervised classification model for predicting if an accident is related to cell phone usage and identifying the strength of the relationships between the independent variables and dependent variable using the chosen model.**

***Keywords—distracted driving, cell phone usage, supervised classification***

## I. Introduction

Distracted driving has become a rising topic over the past couple of years with the increased presence of cell phones and the impact of social media. Though the use of cell phones has seemed to decrease from 2016 to 2017 [1], it is still a factor in fatal car accidents. Vehicle manufactures continue to create "safer" cars and states are now placing laws that ban use of cell phones in vehicles. However, the problem still relies on the individual deciding be unsafe and use their cell phone while operating their vehicles. The goal of this project is to see how the selected variables collected from the National Highway Traffic Safety Administration (NHTSA) Fatality Analysis Reporting System (FARS) influence if a fatal accident is related to cell phone use. The question is answered by fitting the selected fatality data to a supervised classification model and the coefficients analyzed.

The data is appropriate because it is gathered nationwide, it is consistent, and it involves accidents that were the fatal result of individuals using their cell phones as well those that were not. Using fatality data also helps provide a dataset that is less bias. The two raw datasets (one for 2015 and one for 2016) that were obtained through [2] do not involve the psychological or behavioral information of the driver. It only looks at the environment that the accident happened in, the time of the accident, the age of the driver, if the driver was distracted by their cell phone, and information about the car. Using this information, the datasets can help determine if variables that are not behavioral based influence the decision to operate a cell phone while driving.

This paper was originally going to include analysis on the social aspect of using a cell phone while driving by analyzing a dataset containing tweets that mentioned operating a cell phone while driving. The tweets were collected using George Washington University's Social Feed Manager (SFM). The tweets were then going to be classified as Disapproves, Approves, or Neither/Unknown in regard to supporting or not supporting the action of using a cell phone while driving. The classification was done via Amazon MTurk. The data was not used or represented in this paper because the tweets were collected during the month of April which is Distracted Driving Awareness Month. It caused the dataset to be bias due to the higher number of tweets that Disapprove the act of using a cell phone while driving. Though the Twitter dataset was not analyzed in this paper, Fig. 1 is shown to support the claim that people are openly admitting to using a cell phone while driving.

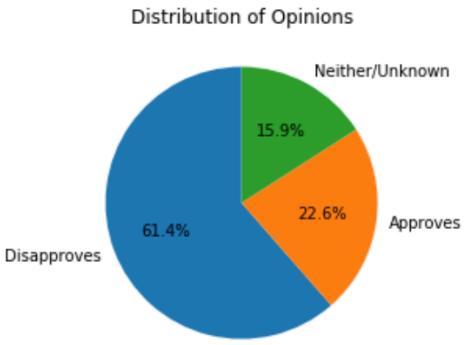


Fig. 1. Distribution of Opinions found from Twitter Social Feed Manager collection using keywords: cell phone use, distracted driving, texting and driving. The full Python analysis and write up is found at [3].

## II. Methodology

### A. Querying FARS Database

The dataset was obtained from the FARS query tool found at [2]. This dataset contains information from fatal car accidents nationwide from the years 2015 and 2016. The two years were

selected because they are the most recent years published besides 2017. The year 2017 was not used because there was an error within the query tool that prevented the data from being accessed.

For both 2015 and 2016, Option 3 (Crash/Vehicle/Driver/Precrash/Occupant) was used and the table shown in Fig. 2 below shows the features that were selected and how they were queried. The features were queried to include only passenger vehicles and light-trucks and only be based on the drivers of the vehicles as defined in [4]. Once the features were selected, the data was downloaded as an Excel file. Fig. 2 shows which features were selected and how they were queried in [2]. The Additional Information column in Fig. 2 explains why the feature was filtered and if it was also cleaned using R.

| Features | Fields Selected | Additional Information |
|---|---|---|
| Case Number | All | |
| State Number | All | |
| Vehicle Number | All | |
| Crash Day | All | |
| Crash Month | All | |
| Crash Year | All | 2015 and 2016 |
| Land Use | 1, 2 | 1 = Rural, 2 = Urban[a] |
| Age | All | This is addressed in the R cleaning process (Age became restricted to all ages under 997 to remove error codes). |
| Person Type | 1 | Include only driver[a] |
| Body Type | 1-11, 14-16, 19-22, 29-33, 39-41, 48, 49 | Include only passenger and light-truck vehicles[a] |
| Number of Fatalities in Vehicle | All | |
| Number of Occupants in Vehicle | All | This is addressed in the R cleaning process (Number of Occupants became restricted to all number of occupants under 99 to remove error codes). |
| Vehicle Model Year | All | This is addressed in the R cleaning process (Model Year became restricted to all years under 9999 to remove error codes). |
| Driver Distracted By | All | This is addressed in the R cleaning process (Accidents that had Distraction code 5, 6, and/or 15[a] were recoded as 1 else 0). |

[a] [4]

Fig. 2. Features selected from FARS Encyclopedia based on the desired information.

## B. Cleaning in R Studio

The two data Excel files obtained from the FARS Encyclopedia database [2] were cleaned using R Studio. The R code used to clean the two data Excel files [5]. There were several features that were cleaned in order to remove the codes associated with unknown. Accidents involving an Age that was greater than 997 were removed. Accidents involving a vehicle with a Model Year equal to or greater than 9999 were removed. Finally, accidents involving the Number of Occupants equal to or greater than 99 were also removed. The total number of accidents that were removed from the raw data based on the criteria above was 773 out of 79,190 [5].

For both of the case years, the Driver Distracted By column was split into three columns such that each column would have one variable. Using the three new columns, the column that has the binary representation of cell phone use was created. Fig. 2 shows the codes that were recoded to be represented as 1 and those that did not were represented as 0 in the new column Cellphone Use. The original Driver Distracted By columns were then dropped. Then the two data frames were combined into the final dataset that is found at [6]. The final dataset has nine categorical features (State Number, Day, Month, Year, Land Use, Age, Body Type, Model Year, and Cellphone Use) and two continuous features (Vehicle Fatal Count and Number of Occupants). Please note that Body Type refers the body type of the vehicle.

# III. Analysis

## A. Dataset Analysis

The initial analysis of the final dataset concluded that there were 830 accidents that involved cell phone usage out of the 78,417. That is means that only 1.058% of the data represent accidents involving cell phone usage [7]. Using this information, the data was undersampled so that each label was represented evenly. The modes and levels for the categorical features and the means for the continuous features are listed below in Table I.

TABLE I. Categorical Features and Continuous Features

| Categorical | | | Continuous | |
|---|---|---|---|---|
| ***Feature*** | ***Number of Levels*** | ***Mode*** | ***Feature*** | ***Mean*** |
| State Number | 49 | 48 | Number of Fatalities in Vehicle | 0.61 |
| Day | 31 | 5 | Number of Occupants in Vehicle | 1.51 |
| Month | 12 | 12 | | |
| Year | 2 | 2016 | | |
| Land Use | 2 | 1 | | |
| Age | 78 | 19 | | |
| Body Type | 18 | 4 | | |
| Model Year | 39 | 2004 | | |

Prior to analyzing how the features influence if the accident is related to cell phone usage based on the supervised classification model, the relationship between each categorical feature and cell phone usage were visualized using stacked bar graphs. Fig. 3 shows the most interesting stacked bar graphs that resulted from this analysis. The graph Age Verses Cellphone Usage shows that the younger drivers have higher proportions of accidents involving cell phone usage as compared to drivers that are older Fig. 3a. The same is shown in Fig. 3b with the vehicles with newer model years as compared to vehicles with older model years.

To get a better look of the distribution of cell phone usage among these two features. The following histograms were produced Fig 4. The orange histograms for each feature represent the accidents involving cell phone usage while the blue represent the accidents that do not. It is interesting to note that the distribution frequencies for each of these features follow the same pattern seen in the proportion graphs in Fig. 3.

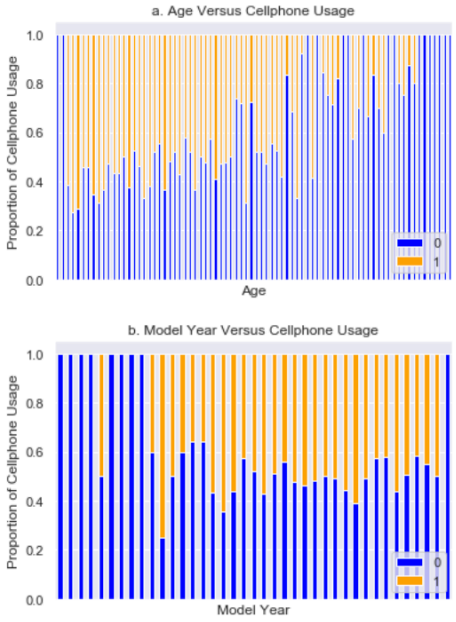


Fig. 3. a) Proporation of accidents involving cell phone usage, labeled as 1 and in orange, for the feature Age. The feature Age ranged from 14 to 93. b) Proporation of accidents involving cell phone usage, labeled as 1 and in orange, for the feature Model Year. The model years ranged from 1968 to 2017.

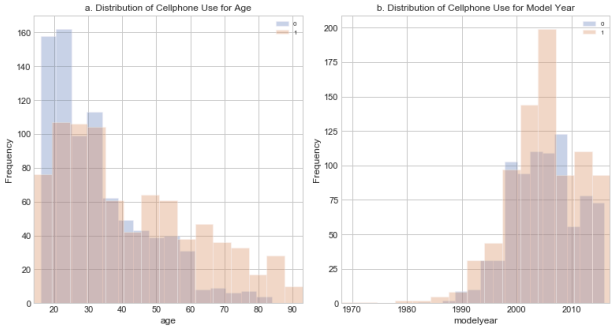


Fig. 4. a) Combined graph of the histogram of accidents involving cell phone usage, represented by 1 and in orange, and the historgram of accidents not involving cell phone usage, represented by 0 and in blue, for the feature Age. b) Combined graph of the histogram of accidents involving cell phone usage, represented by 1 and in orange, and the historgram of accidents not involving cell phone usage, represented by 0 and in blue, for the feature Model Year.

## B. *Principal Component Analysis (PCA)*

*1) Covariance Matrix, Eigenvectors and Eigenvalues:* To determine if dimension reduction would benefit the model a Principal Component Analysis was done on the data. After standardizing the data, the Covariance Matrix found in Fig. 5 was formed. The highlighted cells indicate a high covariance level. The features that had the highest covariance levels were Day, Year, Land Use, Body Type, and Number of Occupants in Vehicle.

Once the covariance matrix was formed, the following eigenvectors and eigenvalues were found. Fig. 6 shows the eigenvectors and eigenvalues for the data.

| **Covariance Matrix** | State Number | Day | Month | Year | Land Use | Age | Body Type | Number of Fatalities in Vehicle | Number of Occupants in Vehicle | Model Year |
|---|---|---|---|---|---|---|---|---|---|---|
| State Number | 1.00000 | 0.00718 | 0.00095 | 0.00299 | 0.00474 | -0.00182 | -0.00923 | -0.00446 | -0.00123 | 0.00228 |
| Day | 0.00718 | 1.00000 | -0.00078 | -0.01532 | -0.00819 | 0.00130 | 0.00023 | 0.00281 | -0.01105 | -0.00215 |
| Month | 0.00095 | -0.00079 | 1.00000 | 0.00066 | 0.00014 | -0.00483 | -0.00251 | -0.00214 | -0.00580 | 0.00543 |
| Year | 0.00299 | -0.01531 | 0.00066 | 1.00000 | 0.01165 | 0.00201 | -0.00281 | -0.00646 | -0.02130 | 0.00376 |
| Land Use | 0.00474 | -0.00819 | 0.00014 | 0.01165 | 1.00000 | 0.00417 | 0.00017 | -0.00239 | -0.01631 | 0.00461 |
| Age | -0.00182 | 0.00130 | -0.00483 | 0.00201 | 0.00416 | 1.00000 | 0.00128 | -0.00172 | -0.00388 | -0.00219 |
| Body Type | -0.00923 | 0.00023 | -0.00251 | -0.00281 | 0.00017 | 0.00128 | 1.00000 | 0.00703 | 0.01909 | -0.00467 |
| Number of Fatalities in Vehicle | -0.00463 | 0.00281 | -0.00214 | -0.00646 | -0.00239 | -0.00172 | 0.00703 | 1.00000 | 0.00555 | 0.00081 |
| Number of Occupants in Vehicle | -0.00123 | -0.01105 | -0.00581 | -0.02130 | -0.01963 | -0.00388 | 0.01909 | 0.00555 | 1.00000 | 0.00005 |
| Model Year | 0.00228 | -0.00215 | 0.00543 | 0.00375 | 0.00461 | -0.00219 | -0.00467 | 0.00081 | 0.00005 | 1.00000 |

Fig. 5. Covariance matrix found in [7] and formed using Excel. The yellow colored cells indicate high covariance between the two features.

| **Eigenvector 1** | **Eigenvector 2** | **Eigenvector 3** | **Eigenvector 4** | **Eigenvector 5** | **Eigenvector 6** | **Eigenvector 7** | **Eigenvector 8** | **Eigenvector 9** | **Eigenvector 10** | **Eigenvalues** |
|---|---|---|---|---|---|---|---|---|---|---|
| 0.17344699 | -0.09189816 | 0.12224104 | 0.48004616 | 0.4267538 | 0.07737812 | -0.3508597 | -0.23789747 | -0.57465671 | 0.13434818 | 1.46743406 |
| 0.19139282 | -0.42063614 | -0.08879243 | -0.40346233 | -0.31090218 | -0.00317099 | 0.32540159 | -0.00824728 | -0.63644519 | 0.08989531 | 0.5950561 |
| -0.38531641 | -0.66974382 | -0.12092452 | 0.35673564 | 0.23991697 | 0.09604472 | 0.38848119 | 0.05118809 | 0.2007861 | 0.02130183 | 1.17987944 |
| -0.11334113 | 0.29093063 | -0.49446706 | -0.00279181 | 0.14483058 | 0.57014301 | 0.14449835 | 0.04482721 | -0.22488256 | -0.48837212 | 1.09937488 |
| -0.13179695 | -0.25311315 | 0.0259024 | -0.65971037 | 0.58684451 | -0.04433495 | -0.32512949 | -0.05050098 | 0.06863563 | -0.15281407 | 0.85728843 |
| 0.34068772 | -0.33467459 | 0.35639032 | 0.08219073 | -0.18002943 | 0.37738369 | -0.28812944 | 0.5031823 | 0.10596464 | -0.34042161 | 0.88748182 |
| -0.27063052 | 0.2927923 | 0.2702341 | -0.06269013 | 0.28135198 | 0.01772551 | 0.22631141 | 0.68265044 | -0.25876234 | 0.31688439 | 1.02352232 |
| -0.61993037 | 0.02888695 | 0.46820885 | 0.03122581 | -0.26433862 | -0.11724055 | -0.07192426 | -0.19459577 | -0.25852366 | -0.44851082 | 0.9919628 |
| 0.38678003 | 0.13575982 | 0.47857155 | -0.03336863 | 0.31652632 | 0.03054791 | 0.59376265 | -0.26928762 | 0.09691438 | -0.25737501 | 0.94338763 |
| -0.17713862 | 0.01888608 | 0.26285148 | -0.17752202 | -0.12909358 | 0.70734587 | -0.03926929 | -0.3264496 | 0.12122613 | 0.47733607 | 0.96064025 |

Fig. 6. Eigenvectors and Eigenvalues found in [7] and the table is formed in Excel.

*2) Explained Variance and Scree Plot:* Fig. 7a shows the cumulative explained variance and the indiviual explained variance for each principal component. Fig. 7b shows the Scree Plot that was generated from the eigenvalues. Both plots show that it takes a great amount of components to reach a decent amount of explained variance.

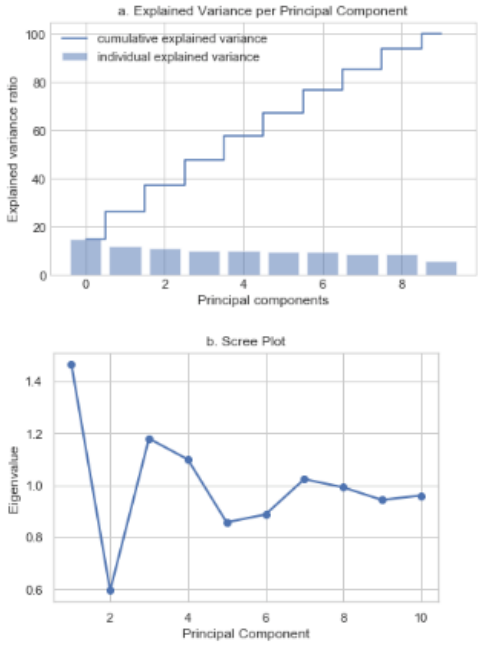


Fig. 7. a) Explained Variance per Principal Component. b) Scree Plot of the eigenvalues for each principal component.

*3) Determining Number of Components:* The first technique that was used to determine the number of principal

components to use if dimension reduction is chosen is to use the number of components that have eigenvalues greater than one. There were four eigenvalues that were greater than one. Therefore, the number of components that were retained were four. The cumulative explained variance for the first four components is 47.673% [7].

The second technique used to determine the number of principal components to keep was by determining the knee of the Scree Plot in Fig. 7b. There appears to be three different knees that can be chosen. For the sake of this paper, the first five components were retained. The cumulative explained variance for the first five components is 57.587%.

*4) Determining Whether to Use PCA or Not:* It was determined that PCA will not be used because the cumulative explained variance from both techniques were lower than desired. Also since half of the components needed to obtain slightly greater than half of the cumulative explained variance it is not worth reducing the number of dimesions. For the purpose of this project, all of the dimensions were retained.

### *C. Deciding on the Model*

SKLearn's GridSearchCV was used to determine which supervised classification model to use. The evaluation metric that was used to score the best model was the F1 scores. The supervised classification models selected to be compared were logistic regression models, Multinomial Naïve Bayes models, support vector classifier (SVC) models using the radial basis function kernel, and linear support vector classifier (LinearSVC) models. The evaluation metric F1 score was used because the model should be concerned on the balance of sensitivity and specificity since it is predicting if an accident is related to cell phone usage. The results shown in Fig. 8 from the GridSearchCV, indicates that the SVC model has the highest F1 score and the logistic regression model has the second highest [7].

```
__MaxEnt (Logistic Regression)__
Best F1 Score: 0.591
Best Parameters set:
{'C': 100.0, 'class_weight': None, 'dual': False, 'fit_intercept': True, 'intercept_scal
ing': 1, 'max_iter': 100, 'multi_class': 'warn', 'n_jobs': None, 'penalty': 'l2', 'rando
m_state': None, 'solver': 'liblinear', 'tol': 0.0001, 'verbose': 0, 'warm_start': False}


__Naive Bayes (Multinomial)__
Best Score: 0.565
Best Parameters set:
{'alpha': 0.01, 'class_prior': None, 'fit_prior': True}


__Support Vector RBF Classifier__
Best F1 Score: 0.641
Best Parameters set:
{'C': 1.0, 'cache_size': 200, 'class_weight': None, 'coef0': 0.0, 'decision_function_sha
pe': 'ovr', 'degree': 3, 'gamma': 0.1, 'kernel': 'rbf', 'max_iter': -1, 'probability': F
alse, 'random_state': None, 'shrinking': True, 'tol': 0.001, 'verbose': False}


__Support Vector Linear Classifier__
Best F1 Score: 0.470
Best Parameters set:
{'C': 1000.0, 'class_weight': None, 'dual': True, 'fit_intercept': True, 'intercept_scal
ing': 1, 'loss': 'squared_hinge', 'max_iter': 5000, 'multi_class': 'ovr', 'penalty': 'l
2', 'random_state': None, 'tol': 0.0001, 'verbose': 0}
```

Fig. 8. Results shown in [7] from the GridSearchCV of each supervised classification model based on using the scoring metric F1 score.

To decide on which supervised classification model to use, each model's classification reports were analyzed. The results shown in Fig. 9 show that the logistic regression has the highest average accuracy at 0.612. The logistic regression model also has the highest precision, recall, and F1 scores from all of the models at 0.56.

One final evaluation metric was used to solidify the choice of using the logistic regression model as the best model. The final evaluation metric was analyzing the model's ROC Curve and AUC. Fig. 10 shows the ROC Curves for each of the models. Table II shows that the logistic regression model has the highest AUC of 0.609 when compared to the other models.

```
__MaxEnt (Logistic Regression)__
Model: LogisticRegression(C=100, class_weight=None, dual=False, fit_intercept=True,
          intercept_scaling=1, max_iter=100, multi_class='warn',
          n_jobs=None, penalty='l2', random_state=None, solver='liblinear',
          tol=0.0001, verbose=0, warm_start=False)
Average accuracy score for Logistic Regression classifier: 0.612
___Classification Report___
              precision    recall  f1-score   support

           0       0.60      0.63      0.61       182
           1       0.52      0.49      0.51       151

   micro avg       0.56      0.56      0.56       333
   macro avg       0.56      0.56      0.56       333
weighted avg       0.56      0.56      0.56       333

__Naive Bayes (Multinomial)__
Model: MultinomialNB(alpha=0.01, class_prior=None, fit_prior=True)
Average accuracy score for Multinomial Naive Bayes classifier: 0.597
___Classification Report___
              precision    recall  f1-score   support

           0       0.58      0.63      0.61       182
           1       0.50      0.45      0.48       151

   micro avg       0.55      0.55      0.55       333
   macro avg       0.54      0.54      0.54       333
weighted avg       0.55      0.55      0.55       333

__Support Vector RBF Classifier__
Model: SVC(C=1, cache_size=200, class_weight=None, coef0=0.0,
  decision_function_shape='ovr', degree=3, gamma=0.1, kernel='rbf',
  max_iter=-1, probability=False, random_state=None, shrinking=True,
  tol=0.001, verbose=False)
Average accuracy score for SVC classifier: 0.561
___Classification Report___
              precision    recall  f1-score   support

           0       0.60      0.14      0.23       182
           1       0.46      0.89      0.61       151

   micro avg       0.48      0.48      0.48       333
   macro avg       0.53      0.52      0.42       333
weighted avg       0.54      0.48      0.40       333

__LinearSVC__
Model: LinearSVC(C=1000, class_weight=None, dual=True, fit_intercept=True,
     intercept_scaling=1, loss='squared_hinge', max_iter=5000,
     multi_class='ovr', penalty='l2', random_state=None, tol=0.0001,
     verbose=0)
Average accuracy score for LinearSVC classifier: 0.503
___Classification Report___
              precision    recall  f1-score   support

           0       0.00      0.00      0.00       182
           1       0.45      1.00      0.62       151

   micro avg       0.45      0.45      0.45       333
   macro avg       0.23      0.50      0.31       333
weighted avg       0.21      0.45      0.28       333
```

Fig. 9. Classification Reports for the best parameter models [7].

TABLE II. AUC SCORES FOR EACH SUPERVISED CLASSIFICATION MODEL

| Model Type | AUC Score |
|---|---|
| Logistic Regression | 0.609 |
| Multinomial Naïve Bayes | 0.581 |
| SVC with RBF Kernel | 0.545 |
| LinearSVC | 0.590 |

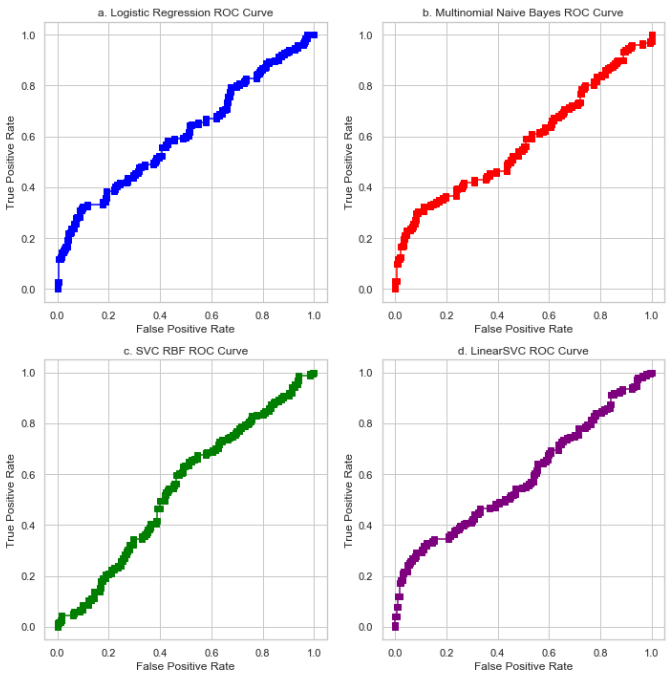


Fig. 10. a) ROC Curve for logistic regression model (blue). The AUC for the logistic regression model is 0.609. b) ROC Curve for Multinomial Naïve Bayes model (red). The AUC for this model is 0.581. c) ROC Curve for SVC having an RBF kernel (green). The AUC for this model is 0.545. d) ROC Curve for LinearSVC model (purple). The AUC for this model is 0.590.

### D. *The Final Model*

Based on these findings shown in Section III.B, the logistic regression model is the best model to use to analyze the influence the features have on accidents that are a result of cell phone usage. The model was created with the parameters shown in Fig. 11 for the logistic regression model. The logistic regression loss function used in this model is shown in (1).

```
{'C': 100,
 'class_weight': None,
 'dual': False,
 'fit_intercept': True,
 'intercept_scaling': 1,
 'max_iter': 100,
 'multi_class': 'warn',
 'n_jobs': None,
 'penalty': 'l2',
 'random_state': None,
 'solver': 'liblinear',
 'tol': 0.0001,
 'verbose': 0,
 'warm_start': False}
```

Fig. 11. Parameters for the logistic regression estimator.

$$l(y) = \log[P(\vec{x})] = \sum_{i=1}^{N} \log\left(1 + e^{-\bar{y}_i(\vec{a}_i \cdot \vec{x}_i)}\right) \qquad (1)$$

10-fold cross validation was used to produce the confusion table below Table III. Based on this table it is understandable why the evaluation metric scores are low. This model on average is predicting 547 true positives and 469 true negatives out of the 1,660 data points. The logistic regression model was saved as a pickle file to preserve the model at its current state from this project [8].

TABLE III. LOGISTIC REGRESSION CONFUSION TABLE

| **Confusion Table** | | **Actual Values** | | |
|---|---|---|---|---|
| | | ***Positive*** | ***Negative*** | |
| **Predicted Values** | ***Positive*** | 547 | 283 | 830 |
| | ***Negative*** | 361 | 469 | 830 |
| | | 908 | 752 | 1,660 |

## IV. RESULTS

### A. *Model Results*

The resulting model produced the logistic equation shown in (2). The model equation is stored as a pickle file at [8]. The highest influencing features are Land Use and Number of Occupants in Vehicle. The lowest influencing appears to be Body Type and Day.

$$\begin{aligned} f(\text{Cellphone Use}) &= -2.8677e^{-05} + 0.0340 \cdot \text{Age} \\ &- 0.0004 \cdot \text{Body Type} - 0.0009 \cdot \text{Day} \\ &+ 0.2628 \cdot \text{Land Use} - 0.0057 \\ &\cdot \text{Model Year} + 0.0052 \cdot \text{Month} \\ &+ 0.2378 \cdot \text{Number of Occupants} \\ &- 0.0113 \cdot \text{State Number} - 0.0348 \\ &\cdot \text{Fatalities in Vehicle} + 0.0049 \cdot \text{Year} \end{aligned} \qquad (2)$$

### B. *Feature Influence*

Another way to analyze the feature's influence is by plotting each feature by their coefficient. Fig. 12 shows each feature and its coefficient in ascending order. It is color coded with negative coefficients being red and positive coefficients in blue. Fig. 12 indicates that Land Use and Number of Occupants have the greatest influence on the model. Age and Number of Fatalities in Vehicle (vfatcount) also have a slight influence on the model.

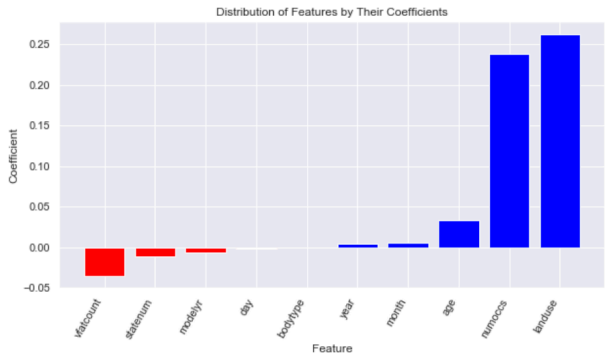


Fig. 12. Graph of each feature based on their coefficient in the logistic regression model. Red represents negative coefficients and blue represents positive coefficients. The idea for this graph came from [9].

## V. CONCLUSIONS

The results shown from the analysis of the undersampled data supports the claim younger drivers are more likely to use a cell phone while driving Section III.A. However, the model shows that the feature age does not have a strong influence on the classification of the data. It is unclear as to why age does not have a stronger influence.

The model could be used for predicting the probability of whether an accident involves cell phone usage or not. Once the 2017 data retrieval error is fixed from the NHTSA it could be used as a new testing group to better the model's prediction. It would be interesting to analyze if there is a more parsimonious model outside of what the PCA dimension reduction could offer. However, since this project was done in a limited amount of time this option was not explored.

## VI. Future Work

The data that can be queried from [2] can be used to answer a bunch of different questions. This data can be used and explored more in depth in terms of other features that were not included in this project. In the future, the Author would like to include the analysis of vehicle makes to analyze the safety results by manufacturers.

Outside of the context of the FARS dataset, the analysis of the social aspect to driving while operating cell phones should be done. Since the original Twitter dataset was gathered during Distracted Driving Awareness Month, it created a bias dataset. The Author would like to re-run the SFM during May and June to collect tweets from two different months. From there, the tweets will be analyzed through the same process used originally [3].

## Acknowledgment

The Author would like to thank Dr. David Broniatowski and Dian Hu for the guidance and time given throughout this semester. The Author would also like to thank Sherlock Warndorf for serving as a therapy dog during times of stress.

## References

[1] National Center for Statistics and Analysis, "Driver Electronic Device Use in 2017," Traffic Safety Facts Research Note, Jan-2019. [Online]. Available: https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812665. [Accessed: 29-Apr-2019].

[2] National Highway Traffic Safety Administration, "FARS Encyclopedia," FARS Encyclopedia. [Online]. Available: https://www-fars.nhtsa.dot.gov//QueryTool/QuerySection/SelectYear.aspx. [Accessed: 29-Apr-2019].

[3] M. M. Warndorf, "GitHub Repository that Houses Twitter MTurk Analysis." GitHub, GitHub, Apr-5AD. Available: https://github.com/madelinew/EMSE6992_Machine_Learning/tree/master/MMW_Twitter_Analysis

[4] National Highway Traffic Safety Administration National Center for Statistics and Analysis, FARS Analytical User's Manual 1975 – 2017. NCSA, 2018.

[5] M. M. Warndorf, "GitHub Repository that Houses R code and Associated Files." GitHub, GitHub, Apr-5AD. Available: https://github.com/madelinew/EMSE6992_Machine_Learning/tree/master/MLEMSE6992.

[6] M. M. Warndorf, "Final dataset that was produced from the raw data using R." GitHub, GitHub, Apr-5AD. Available: https://github.com/madelinew/EMSE6992_Machine_Learning/blob/master/MLEMSE6992/fulldataset1516.csv.

[7] M. M. Warndorf, "Jupyter file containing Python code." GitHub, GitHub, Apr-5AD. Available: https://github.com/madelinew/EMSE6992_Machine_Learning/blob/master/MMW_Final_Project.ipynb.

[8] M. M. Warndorf, "GitHub Repository for MMW_Final_Project." GitHub, GitHub, Apr-5AD. Available: https://github.com/madelinew/EMSE6992_Machine_Learning

[9] A. Bakharia, "Visualising Top Features in Linear SVM with Scikit Learn and Matplotlib," Medium, 31-Jan-2016. [Online]. Available: https://medium.com/@aneesha/visualising-top-features-in-linear-svm-with-scikit-learn-and-matotlib-3454ab18a14d. [Accessed: 29-Apr-2019].