Traffic Accidents Analysis, Prediction and Visualization

Senbiao Fang, Hengwei Zhang, Xingyu Zhu
Electrical Engineering Department
Columbia University
sf2977@columbia.edu, hz2620@columbia.edu, xz2871@columbia.edu

***Abstract*-- Every year, there are thousands of people die in the fatal traffic accidents around the world. As the life quality and medical technique keep increasing these years, the traffic accident have become the major cause of death around the world. Therefore, to avoid the loss of life and property, the government and people should pay more attention to the cause of traffic accidents, and find more efficient methods to predict and prevent the accidents from happening. Based on this background, we decide to do some research and analysis about it. We use the UK traffic accident dataset which is published by the UK government, and use data analysis methods on them. Firstly, after data cleaning, we do EDA and several kinds of visualizations on the data to see the relationship between different factors and traffic accidents, having a direct feeling of the accidents distribution on the map. Secondly, we use several machine learning models like random forest to predict the severity of the traffic accident, compare their performance and choose the best one as the prediction model. Thirdly, we do the data visualization on the map, use k-means method on the location of most serious accidents data and find their centers, which is the best location for the medical stations. Finally, we embed all the elements and results on our web application to do the visualization, prediction, and suggestion.**

***Keywords*--Big Data Analysis, Traffic Accidents, Exploratory Data Analysis (EDA), Data Visualization, K-means, Random Forest Regression**

I. INTRODUCTION

Nowadays, the life quality, medical technique and social safety are keeping increasing around the world, the number of death caused by war, starve and disease has decreased a lot. However, the amount of people die in the traffic accidents is still very large around the world. Every year, there are thousands of people die in the fatal traffic accidents around the world, the traffic accidents have become the major cause of death around the world. The World Health Organization also claimed that the traffic accident has become top 10 causes of death in the world. Therefore, to avoid the loss of life and property, Both the people and government should pay more attention to the cause of traffic accidents, and find more efficient methods to predict and prevent the accidents from happening. Based on this background, we decide to do some research and analysis about traffic accident.

Actually, each traffic accident case has a specific accident condition, corresponding to a lot of objective and subjective factors such as road type, vehicle type, weather condition, light condition, air condition, age of drivers and so on, based on these factors, we can do some data mining, explore the relationships between factors and build an appropriate prediction model based on these factors to predict the traffic accident severity or probability, which can be used as a precaution for the drivers and management department. There were some researches previously which are also focused on the traffic accidents in UK from different perspectives, for instance, paper [1] try to find some trends and patterns about the number of traffic accidents during time, it explores the number of traffic accidents in different time scales, try to analyze the relationships between the traffic accidents and time. Paper [2] also do some data mining on the traffic accidents dataset, and explore the important factors that will cause the fatal accidents. Paper [3] explore the traffic accidents from geological perspective, focus on the Geographical distribution of fatal traffic accidents in England and Wales. These researches mainly focus on the data processing, visualization and analysis, and most of them just focus on one or two perspectives to do the analysis.

In this project, we also do data processing, analysis, and visualization, but we try to do the analysis from a more comprehensive perspective, rather than focus on one or two specific factors. What's more, we use several machine learning methods to build the prediction model, which can predict the traffic accidents severity according to the corresponding factors. We use the UK traffic accident dataset which is published by the UK government, which has a huge scale and valid data, and we use several data analysis methods on them. Firstly, we do data cleaning on the dataset, remove rows in the dataset which has many empty value and remove some irrelevant factors, and we also do EDA and several kinds of visualizations on the data to explore the relationship between different factors and traffic accidents, having a direct feeling of the accidents distribution on the map. Secondly, we use several machine learning methods like random forest to predict the severity of the traffic accident, train the models to do the prediction, compare their performance and choose the best one as our prediction model, a visualization website to predict the accident severity is then built on top of our predictive model. Thirdly, we do the data visualization on the map, use k-means method on the location of most serious accidents data and find their centers, which is the best location for the medical stations. Finally, we embed all the elements and results on our web application to do the visualization, prediction, and suggestion.

II. RELATED WORKS

For the prediction model, over the years, numerous studies have applied a number of methodological techniques to explore the relationship between accident severity and its contributing factors. These methods can be divided into two categories: the statistical-learning method and the machine learning methods. For the statistical methods have been widely used in the traffic accident's severity prediction. For example, logit models[4] , binary probit model [5], ordered logit model [6], ordered probit model [7] are widely used when the outcome has two discrete levels and multinomial

logit [8] for output with more than two levels. [9]has shown that most regression models have assumptions and predefined basic relationships between independent and dependent variables. When these assumptions are violated, the severity of the accident will be inaccurately predicted by a large probability. For the machine learning methods, they have been widely used in traffic prediction problem due to their efficiency, ability to handle multi-dimensional data, flexibility in implementation and strong predictive power. Performance of methods like artificial neural network [10], Bayesian [11], genetic algorithm[12] have been invested. [13]shows a comprehensive reviews of these models. In general, not one method can be identified as the best algorithm concerning various scenario when performing the prediction task. Each modelling technique has its own limitations and characteristics. For our prediction, we want to try various machine learning techniques and measure their performance against each other in order to determine which will be the best algorithm to perform our prediction task regarding both model metrics and training speed.

III. DATA

We use "1.6 million UK traffic accidents" dataset, which contains detailed information about over 1.6million traffic accidents across United Kingdom, collected and published by UK government. This dataset contains large amount of information, the features include but are not limited to, weather conditions, type of roads, type of vehicles, number of casualties, police force, and accident severity. There are two tables in our dataset, the first one is Accident information table, which has 34 columns, and the second one is vehicle information table, which has 24 columns. To take a close look of the data, we firstly do some data analysis and visualization.

We drew a figure to show distribution of traffic accidents by hours in days. From the result shown in the graph, we could easily find that distributions in five weekdays are similar, with accidents concentrating in 8 AM and 5 PM, while accidents in two days of weekend are also similar but accidents concentrate on period between 12 PM to 5PM.

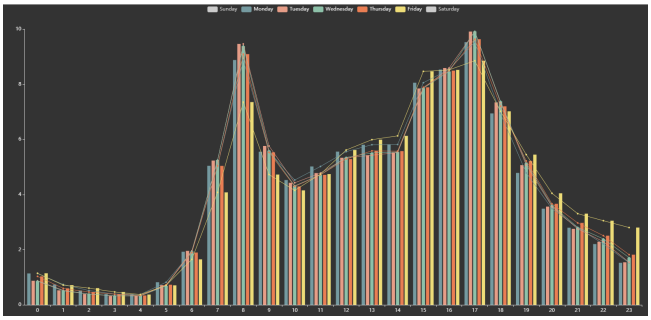

Fig. 1. distribution of traffic accidents in weekdays

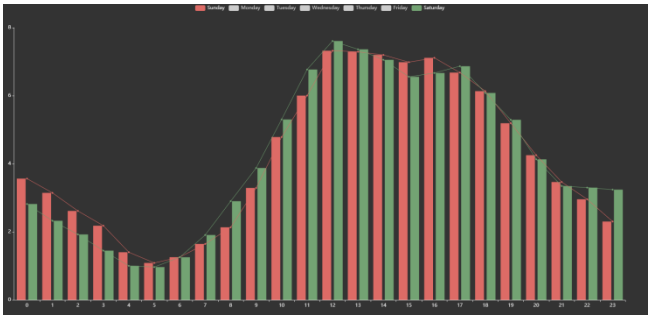

Fig. 2. distribution of traffic accidents in weekend

We picked out the features that could be foreknown before accidents and drew charts to analyze how different conditions in a single factor could influence the severity levels. The features include "Road Type", "Light Condition", "Weather Condition", "Road Surface Condition" and "Speed Limit".

●Road type. From figure 3 and 4 of road type, we could see that "single carriageway" is the road type where most traffic accidents happened from 2005 to 2014, accounted for 74.87%, followed by "dual carriageway's" 14.77%. According to figure 5 of road type, "single carriageway" causes the highest proportion of serious and fatal accidents among all road types.

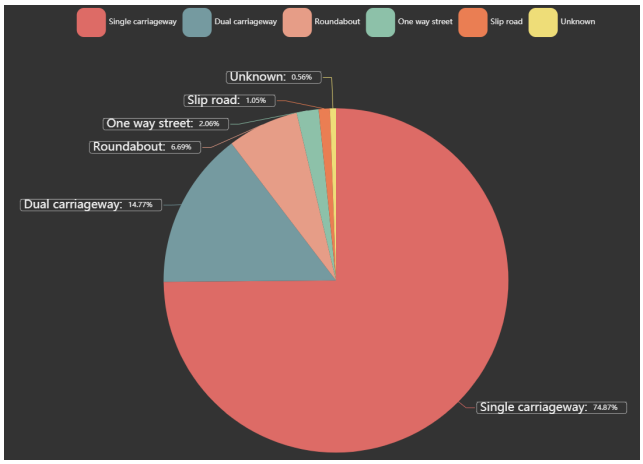

Fig. 3. relationship between road type and traffic accidents

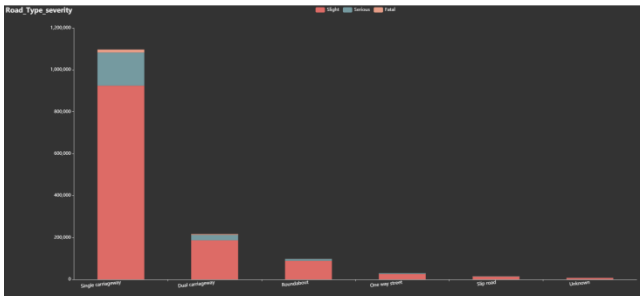

Fig. 4. distribution of traffic accidents on different road types

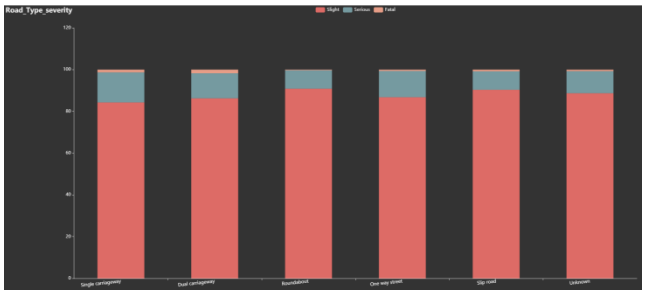

Fig.5.proportion of severity level in different road types

●For light condition, from figure 6,7,8, the highest number of accidents occurring under “daylight: street light present”, accounted for 73.29%, but “darkness: no street lighting” leads to highest proportion of serious and fatal accidents among all light conditions.

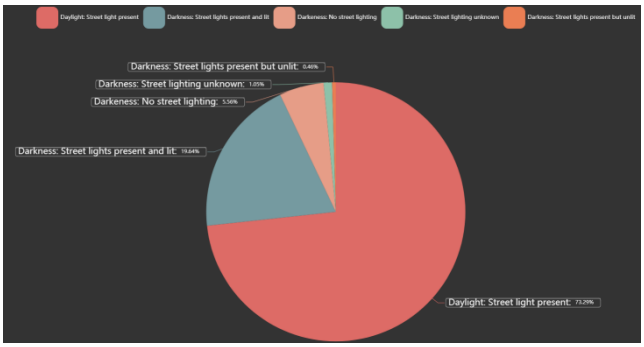

Fig.6.relationship between weather condition and traffic accidents

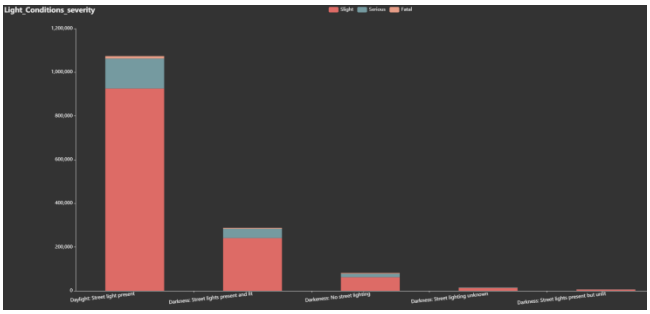

Fig.7.distribution of traffic accidents in different weather condition

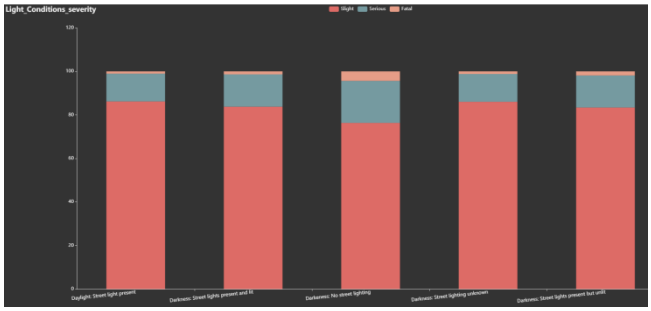

Fig.8.proportion of severity level in different weather condition

Similarly, we can see how other features may cause the most traffic accidents and affect the severity levels of accidents. We filled in a form according to the data and figures that look similar with those figures of road type we performed above, and analyzed the result in the form.

●For weather condition, most accidents happened under “fine weather with no wind”, accounted for 80.13%, and “fine with high wind” takes the highest ratio of serious and fatal levels among all weather conditions, closely followed by “fog or mist”.

●For road surface condition, “dry” road causes the most traffic accidents while accidents on “flood (over 3cm water” are more likely serious and fatal.

●For speed limit, accidents happened in a 30mph speed limit are the most. Accidents happened in a 60mph speed limit have the highest ratio of serious and fatal levels.

Feature	Most Accidents	Highest Ratio of Serious & Fatal
Road Type	Single Carriageway	Single Carriageway
Light Condition	Daylight: Street Light Present	Darkness: No Street Lighting
Weather Condition	Fine without High Winds	Fine with High Winds
Road Surface Condition	Dry	Flood (over 3cm of Water)
Speed Limit	30	60

IV. SYSTEM OVERVIEW

In our project, after we obtained our data, our system performs Exploratory Data Analysis (EDA) to gain insights about the data. Then we built a predictive model, containing a process from data preprocessing to feature engineering, to modeling, to evaluation, and an interactive map.

A flow chart of our system overview：

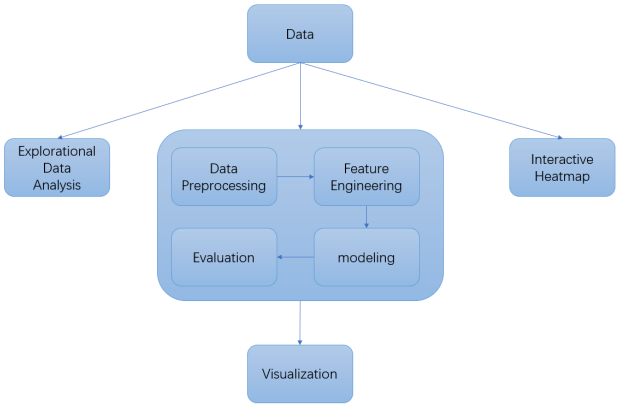

Fig.9. architecture flowchart

1. Predictive model:

Packages:
Scikit-Learn, Pandas, Flask, HTML, CSS, JavaScript, Ajax, Jquery

Having finished data clean up, we tried several machine learning techniques including random forest, logistic regression, XGBoost, LightGBM, AdaBoost, naive Bayes classifiers and MLP to build a predictive model trying to forecast the severity of an traffic accident and evaluates their performances with respect to the model metrics we used to measure our model. And next a webpage to visualize our model is deployed using flask as the back end.

An overview of the website:

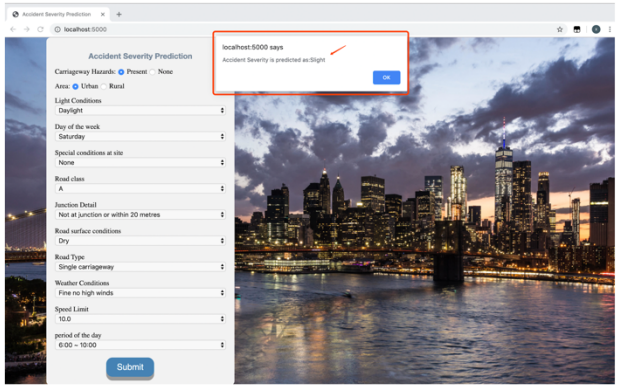

Fig.10. Prediction website overview

Basically, you can choose various conditions on these available variables to predict the accident severity.

For another part of our data visualization, we developed an interactive map to give user an intuitive feeling of how different features affect the number and severity level of traffic accidents. We also deployed a web page for prediction model, which will be posted in the following part.

2.Interactive map:

Package:
dash, pandas

Architecture:

Since Dash is written on the top of Flask, Plotly.js and React.js, we could easily use "html.(label)" and "dcc.(component)" to realize the same effect as labels and components do in HTML. The changes of components trigger callback functions to achieve data processing and data visualization. In our web page, a division, containing all components like checklist, dropdown as input, collects values of different features. The division includes features in format "list" (severity levels, days of a week and hours of a day) and features in format "string" (urban or rural, weather condition, light condition, road surface condition, road type and speed limit). Another division encompasses a graph where the interactive map is drawn as the output. Callback function is composed of data processing and graph drawing. The former one receives values from all components and uses them as filter criteria to extract coordinates fit the conditions we set. The latter one adds the coordinates into a list and pin them on the map.

Usage:

The web page consists of a filter box and a interactive map. Users could first upload a CSV file containing traffic accident data in different years, and then change the value of different features to construct a specific situation. Each time you change a single feature, a process of filtering and fetching data and pinning locations on the map newly will be executed. In this way, we could figure how different combinations of various features might contribute to traffic accidents in diverse levels intuitively.

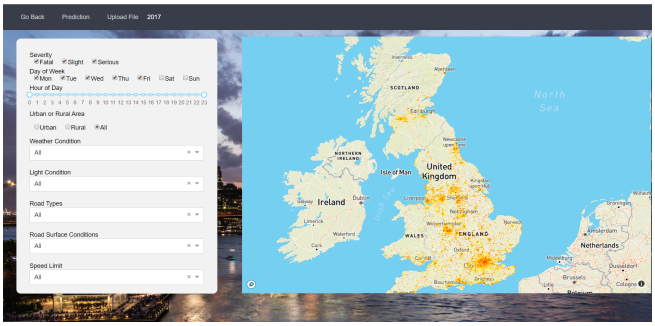

Fig.11. distribution of accidents in weekdays

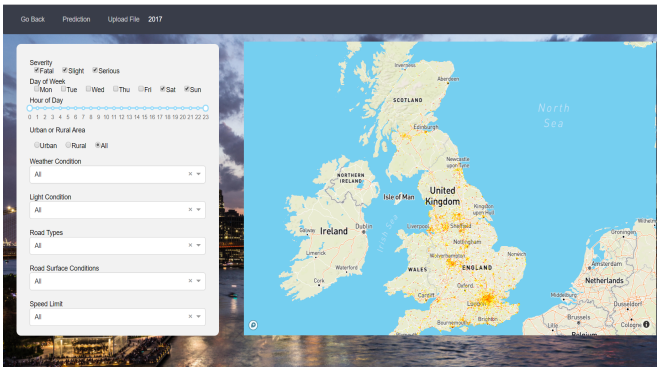

Fig.12. distribution of accidents in weekends

The biggest bottleneck we faced was that the features that could be foreknown before accidents are few. If we want to improve our model further, we might need datasets containing more information about conditions similar to weather conditions, light conditions and so on.

V. METHODS

Algorithms:
Several classification algorithms are used and evaluated in the task of predicting accident severity. For this project, we've tried Random Forest, Logistic Regression, XGBoost classifier, Naïve Bayes, Multi-Layer Perceptron, AdaBoost classifier and LightGBM classifier. After fitting these models, their performances are compared with each other.

●Random forest is based on constructing a forest, e.g. a set of diverse and accurate classification trees, using bagging resampling technique and combining the predictions of the individual trees using a voting strategy[14]. The 'random' term refers to the step when constructing each tree, a random subset of variables are considered instead of the whole set of variables to prevent overfitting.

●Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable. In the logistic regression model, the expected value of response variable is given by[15]:

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$$

where x denotes a value of the independent variable, and β_i values denote the model parameters. The transformation of the $\pi(x)$ logistic function is known as the logit transformation:

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

The parameter is estimated by the maximum likelihood method.

●XGBoost and LightGBM are decision-tree-based ensemble Machine Learning algorithms that use a gradient boosting framework. AdaBoost and Gradient boosting are all ensemble methods, which ensemble multiple weak learners, typically decision trees to produce a strong learner. AdaBoost iteratively fits weak classifiers to the weighted data set, then calculates weights for the classifier, updates weights for each points of which data points difficult to classify are assigned higher weights whereas those easy to classify are assigned lower weights. At the end of iteration, the boosted classifier is the weighted:

$$F(x) = \text{sign}\left(\sum_{m=1}^{M} \theta_m f_m(x)\right)$$

Where f_m is the m-th weak classifier, θ_m is the corresponding weight.

Similarly, Gradient boosting ensembles multiple weak classifiers as the boosted classifier, by starting with a model, consisting of a constant function $F_0(x)$ and incrementally expanding it in a greedy way:

$$F_0(x) = \arg_\gamma \min \sum_{i=1}^{n} L(y_i, \gamma)$$

$$F_m(x) = F_{m-1}(x) - \gamma_m \sum_{i=1}^{n} \nabla_{F_{m-1}} L\left(y_i, F_{m-1}(x_i)\right)$$

$$\gamma_m = \arg\min \sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) - \gamma \nabla_{F_{m-1}} L(y_i, F_{m-1}(x_i)))$$

Where $h_m \in H$ is a base learner function.

●naïve Bayes classifiers are a family of simple "probabilistic classifiers" based on applying Bayes' theorem with strong (naïve) independence assumptions between the features. A bayes classifier, is the function that assigns a class label $\hat{y} = C_k$ for some k as follows:

$$\hat{y} = \arg\max p(C_k) \prod_{i=1}^{n} p(x_i | C_k)$$

Where C_k is classes and x_i is the independent variables.

●A multilayer perceptron (MLP) is a class of feedforward artificial neural network made up of multiple layers of nodes in a directed graph, with each layer fully connected to the next one. An MLP consists of at least three layers of nodes: an input layer, a hidden layer and an output layer. Except for the input nodes, each node is a neuron that uses a nonlinear activation function. MLP utilizes a supervised learning technique called backpropagation for training[16]. Its multiple layers and non-linear activation distinguish MLP from a linear perceptron. It can distinguish data that is not linearly separable.

●K means method is a very popular clustering method in machine learning, it is a simple but very efficient method for clustering problem. In the process of k means method, at first, we need to decide the number of centers k of our clustering, and we calculate the distance of each node and the k centers (we can use the Euclidean distance or Manhattan Distance to calculate the distance), then we assign the node to the cluster i which is the nearest center to it, after that, we recalculate the center for each clusters, and do the above operation again, until the centers of each cluster become stable.

Preprocessing:
First, we want to perform feature selection to identify a subset of features that are likely to affect the accident severity. For feature selection methods, they can usually fall in one of the three main categories: wrappers, filters and embedded methods.

Wrapper method uses some specific prediction algorithm to find which subset of variables achieve the highest metric score using that algorithm. Each new subset is used to train a model, and performance is measured on a held-out test set. It is usually computationally intensive but results in a subset of features which performs best for that particular type of algorithm.

Filter methods use a proxy measure instead of the error rate to score a feature subset. This measure is fast to compute and can capture the usefulness of the feature set. Filters are a more general method which don't require a specific type of predicting algorithm, so it usually result in feature subset which isn't tuned for a particular type of predictive model.

Embedded methods are algorithms which embeds a mechanism to perform feature selection as part of the model construction process. Lasso is an example of this method. During the construction of a linear model, it penalizes the regression coefficients with some kind of penalty, shrinking them to zero to obtain a subset of features.

Filter method usually give a feature ranking instead of an explicit feature subset. It has also been used as a preprocessing step for wrapper methods, allowing a wrapper to be used on larger problems.

For our project, we've used a filter as part of the preprocessing step. We used Pearson's Correlation Coefficients and Cramer's V as our measure to identify variables that have "high" correlation with the label and select a subset of these variables and also some variables we are interested in as our independent variables in our predictive model.

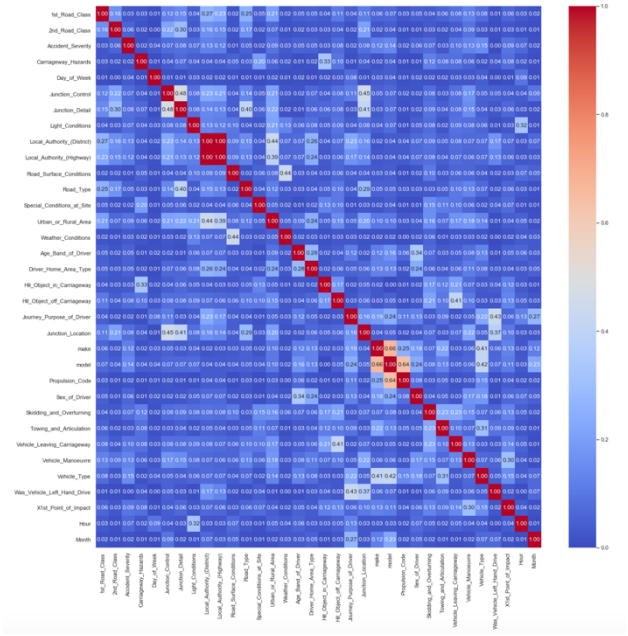

Fig.13. Cramer's V maxtrix for categorical features

Fig.14. Correlation maxtrix for numerical features

Then a wrapper method RFECV (Recursive Feature Elimination with Cross Validation) with Random Forest is performed to find the best subset of variables and we also tried an embedded method which is Logistic Regression with different kinds of penalties.

After choosing our variables, some further preprocessing steps are performed. For columns that contain missing values, impute with the mean for numerical columns and the most common occurring class for categorical columns when their proportions of the corresponding column are large (more than 5%) and cannot be simply filtered. If a column has only a small proportion of missing values, we simply filter these rows out.

For the encoding, we used one hot encoding for distance-based algorithms, and label encoding for tree-based algorithm. Also standardize normalization is performed for the data when fitted to distance-based algorithms. The label variable contains three classes: Fatal, serious, slight. To reframe this as a binary classification problem, we class fatal accident to serious accident.

After these steps, another issue to deal with is that this dataset is imbalanced, with the majority of the label class being 'slight' and only a small proportion of the label class being 'serious'. This problem arises when we're interested in predicting the minority class since the classifier will be biased towards the majority class which means they tend to minimize the errors on the majority class samples and the accuracy rates of predicting the minority class will be low. In predicting the accident severity, correctly identifying the accidents with a serious level is as important as identifying the slight samples. To address imbalance, usually some resampling techniques can be used such as oversampling, undersampling, SMOTE.

For this dataset, we adopted undersampling, which is to reduce the samples of the majority class to roughly the same number of samples of the minority class to achieve a uniform distribution. After that, some data binning is performed to minimize the effects of small observation errors. Also, for imbalanced dataset, other metrics should be used to measure our model performance instead of accuracy. We used roc-auc-score and f1-score as metrics.

Note that train-test-split should be performed before normalization and undersampling to prevent the leakage of information of the test dataset into the training dataset.

Feture engineering

For this dataset, we performed some simple feature engineering, such as creating new variables like hour and month from Time and Date.

Model fitting and hyper-parameter tuning

After preprocessing, our data are ready to be feed into our models, for the hyperparameter tunning, we utilize the tool provided by scikit-learn, use RandomizedSearchCV and GridSearchCV to tune the hyperparameters for each algorithm.

VI. EXPERIMENT RESULTS

We split the dataset into 70% training set and 30% validation set randomly.

A summary of the comparison of our model performance can be seen in the chart below:

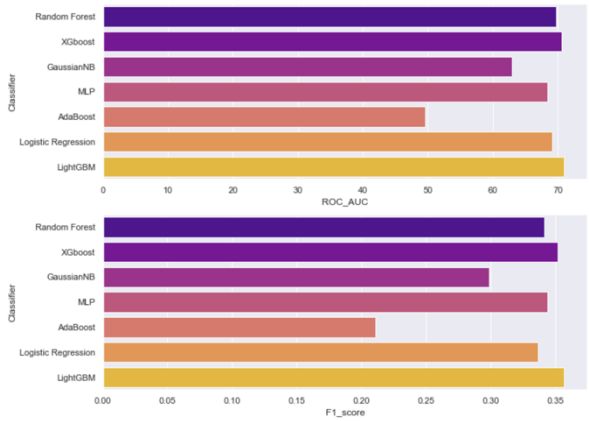

Fig.15. model metrics comparison

As shown in the table, Tree-based algorithm performed relatively similarly concerning the measuring metrics of our model, except for AdaBoost whose performance is inferior to other methods. This is may due to that hyper-parameter are not well tuned since we performed randomizedsearch for algorithms that have long training time. Naïve bayes performs also slightly worse, which is not surprising since it is assuming independence between variables. For this dataset, LightGBM achieves the best performance in both metrics and also is very fast in training speed. So this is adopted as our classifier when developing our website.

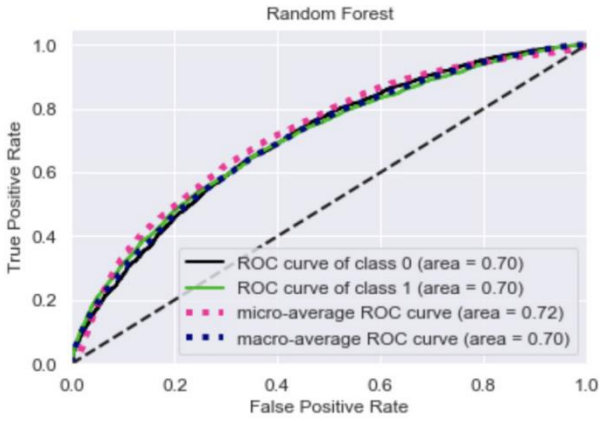

Fig.16. ROC curve of Random Forest

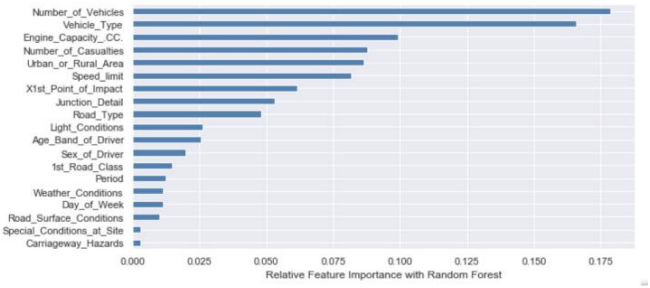

Fig.17. Feature importance by Random Forest

Random forest has a built-in mechanism to measure the feature importance, it's easy to visualize which variable contribute more to the seriousness of the accident severity as shown figure above. 'Number_of_Vehicles' ranks first in the feature importance whereas variable like 'Weather_Conditions', 'Road_Surface_Condition' which we initially suspect may server as contributing factors in determining the severity of an accident turned out to be not so strong variables.

REFCV with Random Forest:

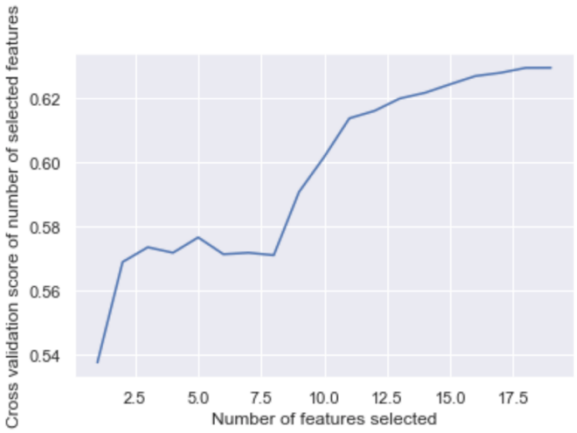

Fig.18. results of RFECV with RF

This is a wrapper method we performed trying to find the best subset of variables which achieves the best performance. For random forest and the set of features we chose, including all features actually result in the best performance.

We also use the k means method to find the best location to build the medical station (consider the minimum average distance). In the previous work, we have already shown the number of traffic accidents and the severity of the traffic accidents on the map, and then, we use the k means clustering method on the locations of most serious accidents and find the centers of them, which are the best locations for the government to build the medical stations. Since the k means method should set the number of clusters k at first, which in our model means we should are given the number of medical stations we want to build, and then we an give the best recommendation of locations for these k medical stations. Following figures show our result to give the best locations for 5 and 20 medical stations.

Fig.19. k-means results with k = 5

Fig.20. k-means results with k = 20

VII. CONCLUSION

This article gives details of our data preprocessing, analysis and visualization on UK traffic accidents dataset, and there are many useful conclusions for our projects.

- Firstly, during the process of EDA, we find that there are so many valuable relationships between different conditions and the number of traffic accidents and the severity of different traffic accidents. For instance, the number of traffic accidents according to the time in a single day have the same trend for the days in weekday which is shown in the above DATA part, and the number of traffic accidents according to the time in two days of weekends have their own specific trend, which implies that there are different modes for the distribution of the traffic accidents in weekdays and weekend. Also, we explored how different conditions in a single feature affect the amount of traffic accidents and the severity level of traffic accidents. For example, if we take a look at the light condition, we might find the accidents happened in daylight took up the biggest part of all traffic accidents. However, if we compare the ratio of serious and fatal traffic accidents, we may find that among all the light conditions, darkness: No street lighting holds the highest proportion of serious and fatal accidents.
- Secondly, we use many different machine learning methods to try to build the prediction model, we expect to build a model which can predict the severity of traffic accidents precisely. We use the models like logistic regression, Adaboost classifier, random forest, XGBoost classifier, Multi-Layer perception, Gaussian naive bayes, and light GBM classifier. We use our data to train each model and adjust the parameters and features of the models, at the end, the result shows the light GBM classifier has the best performance, and we use this model as our accident severity prediction model, and build a visualization model on our web application to do the prediction according to the given condition of features.
- Thirdly, we do visualization of locations and severity of traffic accidents and our prediction on the dynamic map, which can give us a direct feeling of the distribution of the traffic accidents, the map is also embedded in our web application. We also use a k means method on the location of most serious traffic accidents and find there k centers, which are the best locations to build medical stations.

Our article also provides thoughts and methods to do the analysis and prediction on traffic accidents dataset.

ACKNOWLEDGE

Sincerely thanks to our professor CHING-YUNG LIN and teaching assistant Juncai Liu, Hung-Yi Ou Yang, Tingyu Li, and Yunan Lu. During this semester, with the direction and help from professor and TAs, we have learned many useful skills and methods in the big data area. What's more, during the process of this project, they also give us many insightful advise and selfless help to fix difficult problems.

APPENDIX

Individual Contribution

Basically, each team member has the same contribution to this project.

●Senbiao Fang has been working on data visualization, build of web application and the k means recommendation model. (33.3% of the total work)

●Xingyu Zhu has been working on the build of prediction model, try different machine learning methods on the prediction models. (33.3% of the total work)

●Hengwei Zhang has been working on the data cleaning, data visualization and build of the interactive map. (33.3% of the total work)

RENFERENCE

[1] Peter Ljubič, Ljupčo Todorovski, Nada Lavrač, "Time-Series Analysis Of UK Traffic Accident Data "

[2] David D.Clarke, Patrick Ward, Craig Bartle, Wendy Truman, "Killer crashes: Fatal road traffic accidents in the UK "

[3] Robin Haynes, Andrew Jones, Ian Harvey, Tony Jewell, David Lea, "Geographical distribution of road traffic deaths in England and Wales: place of accident compared with place of residence"

[4] Kononen DW, Flannagan CAC, Wang SC. Identification and validation of a logistic regression model for predicting serious injuries associated with motor vehicle crashes. Accid Anal Prev. Pergamon; 2011;43: 112–122. pmid:21094304

[5] Haleem K, Abdel-Aty M. Examining traffic crash injury severity at unsignalized intersections. J Safety Res. Pergamon; 2010;41: 347–357. pmid:20846551

[6] Quddus MA, Wang C, Ison SG. Road Traffic Congestion and Crash Severity: Econometric Analysis Using Ordered Response Models. J Transp Eng. 2010;136: 424–435.

[7] Zhu X, Srinivasan S. A comprehensive analysis of factors influencing the injury severity of large-truck crashes. Accid Anal Prev. Elsevier Ltd; 2011;43: 49–57. pmid:21094296

[8] Malyshkina N V, Mannering FL. Markov switching multinomial logit model: An application to accident-injury severities. Accid Anal Prev. 2009;41: 829–838. https://doi.org/10.1016/j.aap.2009.04.006 pmid:19540973

[9] L.-Y. Chang and H.-W. Wang, ''Analysis of traffic injury severity: An application of non-parametric classification tree techniques,'' Accident Anal. Prevention, vol. 38, no. 5, pp. 1019–1027, Sep. 2006.

[10] REZAIE, MOGHADDAM F., Sh Afandizadeh, and M. Ziyadi. "Prediction of accident severity using artificial neural networks." (2011): 41-48.

[11] Zong, Fang, Hongguo Xu, and Huiyong Zhang. "Prediction for traffic accident severity: comparing the Bayesian network and regression models." Mathematical Problems in Engineering 2013 (2013).

[12] .Kunt, Mehmet Metin, Iman Aghayan, and Nima Noii. "Prediction for traffic accident severity: comparing the artificial neural network, genetic algorithm, combined genetic algorithm and pattern search methods." Transport 26.4 (2011): 353-366..

[13] Mujalli RO, Oña J De. Injury severity models for motor vehicle accidents: a review. Proc ICE—Transp. 2012; 1–16

[14] Breiman L (2001) Random forests. Mach Learn 45(1):5–32

[15] A. S. Al-Ghamdi, "Using logistic regression to estimate the influence of accident factors on accident severity," Accident Analysis and Prevention, vol. 34, no. 6, pp. 729–741, 2002.

[16] Hornik, Kurt, Maxwell Stinchcombe, and Halbert White. "Multilayer feedforward networks are universal approximators." *Neural networks* 2.5 (1989): 359-366.