McKinsey & Company Challenge

Accident Severity detection

12/10/2019

HackUPC

Team Spacers

Animesh Bajpai

Hossein Yousefi

Jonathan Harel

Matteo Buitrago

Overview

On exploratory analysis:

- It was evident that the target class was skewed i.e there were 5 times more rows of data for non-severe/fatal accidents than for their severe counter parts.
- Visualizations were used to understand the impact of features on classes
- On further deep diving it was observed that the features that would lead to an intuitive inference e.g weather and road conditions were not a part of the most important features
- Multiple different models were tested (MLP, Random Forest, Logistic Regression) but the results obtained were unsatisfactory, so a decision to boosting algorithms was made.
- Catboost performed significantly better than other algorithms out of the box.
- On first benchmark modeling, it was observed that a high accuracy (>80%) can be obtained with an unimpressive F1 score of 0.03, this can be attributed and inferred from the Data Skew
- Additional Data preprocessing and feature engineering improved the F1 score to 0.72(Best result)

Detailed Technical Description

Exploratory Analysis

- To begin, a bar plot with the number of each samples was plotted to understand the distribution of the data between the two target variables.
- Stacked bar plots of features failed to reveal any relevant information due to the huge skew (5x).
- To gain a deeper understanding of the features the dataset was separated into the two target Dataframes i.e a dataframe for severe accidents and one for non-severe.
- Individual features were plotted with respect to the target variables to understand what kind of impact they may have on the classification
- Insights were obtained and intuitions were debunked about what may be important, this was vital in understanding and determining the next steps
- Key points: Location based features e.g OSGR and latitude and longitude turned out to be more important than weather and road conditions.

Feature Engineering

- To understand the significance of the columns and build a feature matrix, a correlation matrix was plotted and visualised using a heatmap. This provided a clear understanding of what columns were to be dropped. E.g road numbers, pedestrian crossings etc.
- Just removing the uninformative features helped improve the F1 score by a small amount (0.20).
- To counter the Data skew, we used downsampling of the non-severe accident class(target column = 0) and this significantly improved the F1 score (0.53) but not close to being considered a good model.
- A use of relative features was also made to decrease the size of the feature vector while improving the information gain e.g speed limit was squared to increase its importance, casualties was divided by the number of vehicles involved etc.
- Merging the tables for vehicle data and the accidents was the task that required some attention, this was due to the existence of multiple references to the same accident id, to combat that, the column values were one-hot encoded and aggregated by summing to produce an input feature vector.

Using this merged table of vehicle data and accident data, a satisfactory top F1 score of 0.72 achieved. Average reproducible F1 score of 0.70..

Business Problem Definition

Problem Statement

- The model is based on historical data and is used to predict the severity of an accident that has already happened, which is not exactly useful since such is evident after an accident.
- What the model can do is provide interpretability and a deep understanding of what actually makes the difference between a severe accident and a normal accident
- This understanding can provide valuable, actionable intelligence which will provide an insight into predicting and preventing future events and in turn saving lives and money.
- The actions proposed are related to providing information to relevant governmental authorities, to help in city planning and urban development of roads and passing of new bills and laws relating to road safety and automotive manufacturing.
- The above actions are based on the idea that McKinsey & Company provides consulting services for multiple governments across the globe.

Tentative Business Plan

- This plan is an add on McKinsey's existing consulting services in the public sector.
- It requires building a platform where the government can obtain actionable information and intelligence in order to better the public sector and help put a stopper on unwarranted and preventable deaths.
- The financial impact of the product can be substantial , in the 2012 Annual Report: Reported Road Casualties in Great Britain, it is stated the the value for prevention of road accidents is 34.4 Billion GBP
- The Global costs incurred due to Road Accidents is around 518 Billion USD.
- There is a growing opportunity to prevent these incidents, reduce costs incurred for countries and build a stable platform which can be profitable and providing it on a global level.
- The platform would include the use of heatmaps to alert governmental agencies to update the infrastructure at specific locations, build counter laws to unsafe road practices.

IT Infrastructure

- Hadoop cluster with a replication factor of 3
- 5-10 Nodes (small cluster) - scaled as per requirements
- Replication factor - 3
- Data growth assumption per day - 1 GB per day therefore
- Space growth per day is 3Gb per day
- Spark ML on the cluster to allow parallel computing abilities
- Hive for distributed Data processing

This cluster will allow us to build scalable Machine learning models and will make sure that we will always have enough space to add more data

Cluster Specifications

Resource	Requirement
Processor	Hex/Octa Core processor
Memory	64 gb expandable to 512 gb ECC Ram
Storage	12-34, 1 TB SATA disks
Network	Gigabit ethernet with link aggregation

Apart from the cluster, There will also be a requirement for a development server for the reporting platform, which will be built using Flask backend(python) with Angular for frontend.

Future and other uses

- This technology provides details that may be helpful in building better insurance plans for insurance companies
- It may also be used to provide real time updates based on driver information, vehicle information, location data, weather data and road conditions to provide real time risk detection in order to make sure the driver stays alert to avoid mishaps.

Key Takeaways and Conclusions

- Nintendo Switch Lite (Hopefully).
- More experience in feature engineering.
- Knowledge of the CatBoost model (First time users of this architecture).
- Better understanding of using traditional Machine learning algorithms beyond digital metrics.
- Road accidents are a product of human decision making and city planning and can be prevented by understanding the real errors of city planning and provide intervention at the right time.
- Accident prevention is a multi billion dollar industry