Traffic Violation Analysis

Introduction:
- Jinan, Shandong, China is a large city with heavy traffic problems.
- The objective of the project was to analyze the traffic violation data in Jinan and propose suggestions to the Traffic Management Department.
- The data used was 2015 Traffic Violation Data. The data has 2,897,543 records with 32 features including license number, driver information, violation site, violation behavior, violation time, etc.
- Preprocessing the data:
  - Selected the vehicle type 'cars' (2,332,085 records).
  - Focused on three features: violation time, violation site, violation behavior.
  - Violation time was labeled from 1 to 24; violation behaviors were re-categorized; violation sites were converted to latitude and longitude by using Baidu API.
  - The below is a heatmap showing the traffic violation situation in Jinan.

Method:
- Birch clustering was used to find the District.

- Statistical summary was applied to find main type of violations in each town of suburbs.
- Applied k-prototypes clustering on downtown data in every time range.
  - Silhouette analysis was used to choose optimal number of clustering for each time label.

Result:
- The right plot shows violation distribution from 17:00 pm to 18:00 pm.
  - The clustering centroid could be the center where the traffic police start their patrolling from.

Ames – Housing-Prediction

Introduction:
- The objective was to help a real estate investment firm in Ames, Iowa, a mid-west town in the United States, to analyze how housing features influence housing price and what types of real estate properties are good investments.
- The data used for this project was housing sales from 2006 to 2010 that contained information about the characteristics of the house (number of bedrooms, number of bathrooms, square footage, etc.) and the house’s sale price.

Exploratory data analysis:
- The original distribution of outcome variable: price was highly skewed below its mean. To obtain a normal distribution, log transformation was applied on price.
- Checked with correlation between continuous variables.
- Random forest and boosting were applied to measure feature importance. The below left plot shows top 10 features derived by random forest and the below right table is the result from boosting.

Model:
- Simple model (no interaction): OLS, OLS with BIC, random forest, boosting, ridge and lasso model were explored.
- Complex model (with interaction): GAM, boosting, random forest, BMA and Lasso model were explored.

Results:
- In terms of rmse, OLS with BIC is the best for simple model. The rmse for the training data is 15,493.6357; the rmse for the testing data is 15,735.0644.
- The below left is diagnostic plots for OLS with BIC, showing insufficient of the model.
  - Lasso was selected as the final model. The rmse for training data dropped to 14,965.164 and the rmse for the testing data dropped to 14,740.647. The right above is residual plot of Lasso Model.
- 10 overvalued and undervalued houses were found by using predicted price from the final model.

Acknowledgements:
- Joint work with Liyu Gong, Wenqi Cheng and Yilin Gao.

Hospital and Physician Quality Assessment

Introduction:
- The Centers for Medicare and Medicaid Services (CMS) is required by Congress to evaluate hospital performance. In addition, hospitals (and patients) have interest in ranking physicians.
- The data included physician name, hospital name, region, procedure type, number of cases, number of deaths, observed mortality rate and expected mortality rate.
- The objective was to build a model to evaluate performance of hospitals and physicians.

Model:
- Binomial Generalized Linear Mixed Model: Number of deaths and number of survived patients were treated as outcome variable; number of deaths with different physicians and hospitals within a procedure should be highly correlated, so GLMM with random intercept effect was applied.
  - The model used is the following:
  - Stage1: logit(outcome = survived) = β0 + βregion* + βprocedure* + βregion*+ βprocedure* + βhos* + βhos*
  - Stage2: fixed estimates: β0~unif (0,1); pβ =N(0,1), pβ =L, pβ =L; random effects: βhos*+ βhos*~N(0,Σ); Σ = half-t(3,0,10)

Result:
- The standard deviation produced for hospital random effects and physician random effects are showed at the below left table:
- Estimates of the ratio of the model-predicted mortality rate to the expected mortality rate are showed at the below right table:

Conclusion:
- Physicians had random effects; Error for sd of hospital random effects was large, so we kept doubt whether we should include hospital random effects.
- The model produced reports for physicians performance. Presby and advised the manager based on physicians performance.
- The model produced 5 high-achieving hospitals and 5 low-achieving hospitals based on predicted ratio.

- [Image of hospital and physician quality assessment]