2017 Travelers Case Competition

Team Da(ta) Duo
Joon An
Tae Park

Contents

- Methods, first vs final model comparison
- Preparation & Exploratory data analysis
- Model fitting & Selection
- Model evaluation
- Overall Explanation on the Significant Variables
- Additional variables to consider

Methods

- Linear/Quadratic Discriminant Analysis
- K-nearest neighbor (c-stat score on train: 0.6764)
- Generalized Boosted Models (~0.7340)
- Random Forest (~0.6944)
- Logistic Regression (0.7398)
  - Makes less assumptions
  - More safe from Overfitting
  - More intuitive interpretation
  - Consistent result

First vs Final model comparison

First Model (c-stat score: 0.713885)

- No interaction terms
  - Average 10-fold cross validation c-stat score on training set: 0.728
- Variable ‘year’
  - Assuming year 2017 is same as baseline for the variable year in the model
- No ‘Landlord’ treatment
  - Assuming ‘Landlord’ is same as baseline for the variable ‘dwelling.type’ in the test set

Final Model (c-stat score: 0.709756)

- Interaction terms that are sensible
  - Average 10-fold cross validation c-stat score on training set: 0.733
- No variable ‘year’
  - Ignored variable year since we don’t have information of year 2017
- Treatment for ‘Landlord’
  - Replaced ‘Landlord’ in test set with other dwelling type predicted by package ‘mice’

Preparation & Exploratory data analysis

Training Set

- NA or cancel = -1 -> Deleted
- age => 100 -> Deleted
- Len.at.res > age -> Deleted

Test Set

age => 100 -> NA

Len.at.res > age -> NA

Used R package ‘mice’ to fill up NA on Test Set

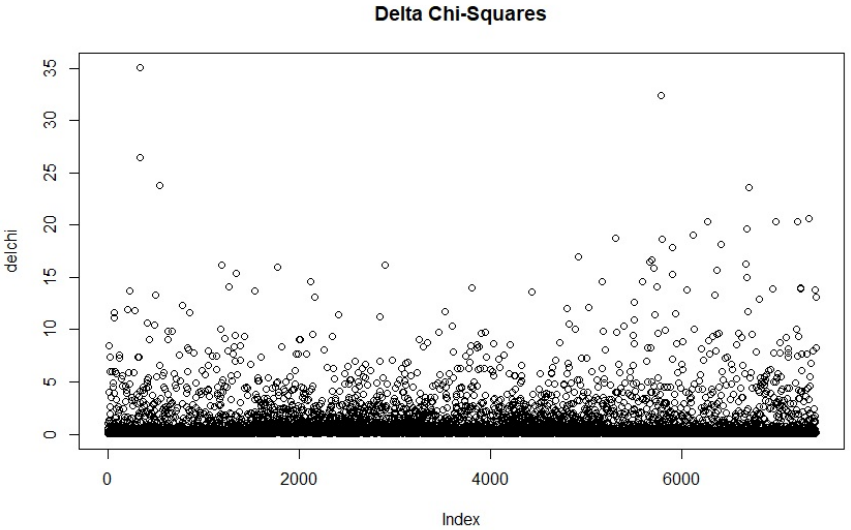
Exploratory data analysis

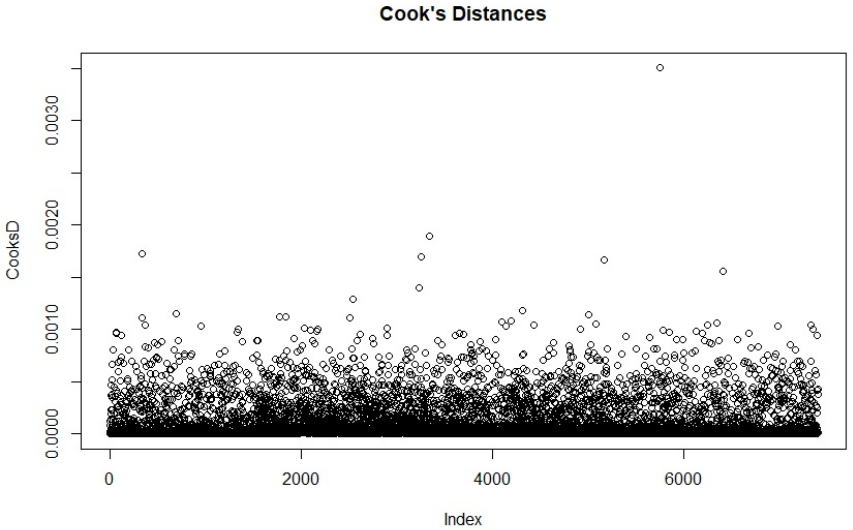
1. Created dummy variables
2. Checked correlations among variables
3. Tested categorical variables to reduce number of dummy variables

Model fitting & Selection

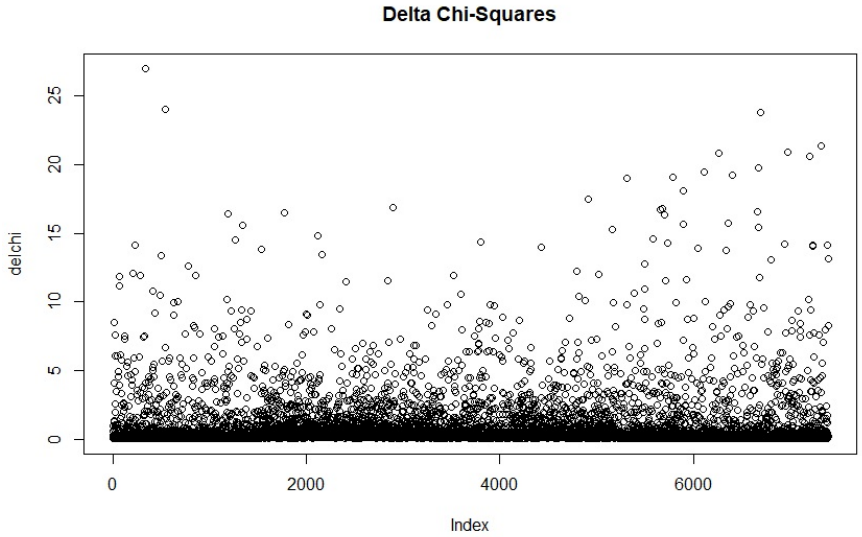
- Fitted data to a binary glm with logit link function
- R Package “My.stepwise”
  - Stepwise regression with alpha = 0.05 to go in and stay
  - Checks VIF every step
  - Provides AIC every step
- Selected the model at the final step in stepwise selection which has all significant variables at alpha = 0.05 and lowest AIC value

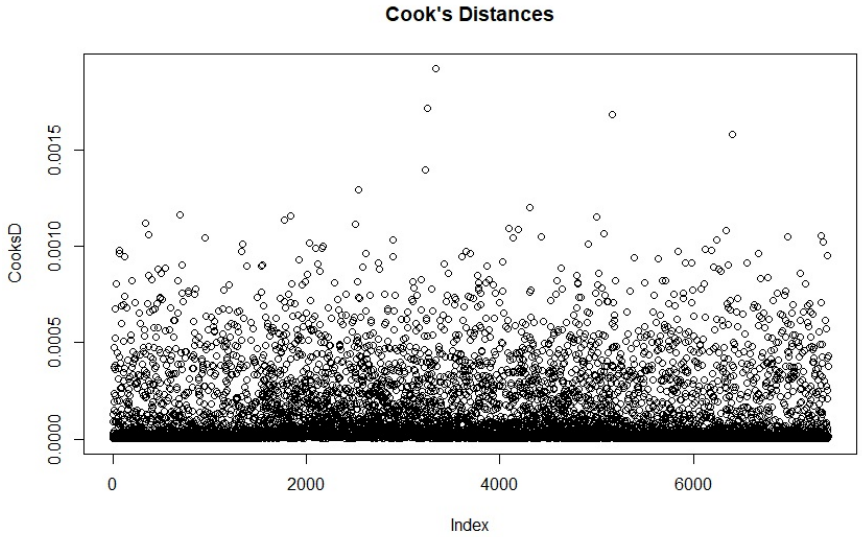
Model fitting & Selection (Diagnostics)



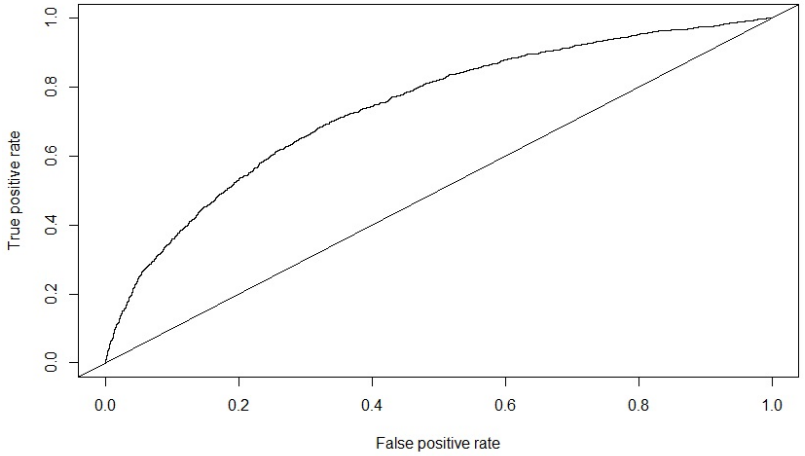


Model fitting & Selection (Diagnostics)





Model evaluation



- C-stat score: 0.7398
- Hosmer-Lemeshow goodness of fit Test: p-value > 0.05
- Average 10-fold cross validation c-stat score: 0.728
- C-stat score on test set: 0.713885

Overall Explanation on the Significant Variables

- The model:

$$
\begin{aligned}
\text{logit}(p) = \; & -1.390484 \\
& +0.820160 \times sales.channelNonBroker \\
& +0.919291 \times zip.codeDC \\
& +0.643669 \times year2014 \\
& +0.358731 \times claim.ind \\
& -0.021362 \times ni.age \\
& -0.317995 \times ni.marital.status \\
& +0.086865 \times n.adults \\
& +1.468463 \times creditlow \\
& +0.721136 \times creditmedium \\
& +0.129829 \times n.children \\
& -0.648579 \times zip.codePA \\
& +0.377083 \times year2015 \\
& -0.031081 \times len.at.res \\
& +0.026272 \times tenure \\
& -0.215616 \times zip.codeCO
\end{aligned}
$$

Additional variables to consider

zip.codeDC	0.9193
zip.codeCO	-0.2156
zip.codePA	-0.6486

Potential Variables :
Region (Urban, Suburban, Rural)

creditlow	1.4685
creditmedium	0.7211
n.adults	0.0869
n.children	0.1298

Potential Variables :

- Education Level
- Single Parent
- Occupation (Blue Collar, White Collar)
- Birth Order

sales.channelNonBroker	0.82016

Works Cited

- Greg Kaplan, "Moving Back Home: Insurance against Labor Market Risk," *Journal of Political Economy* 120, no. 3 (June 2012): 446-512.
- Sulloway, Frank J. “Birth Order and Intelligence.” *Science* 316.5832 (2007): 1711–1712. Web.