IBM Applied Data Science Capstone

Final Assignment

PREDICTING CAR SEVERITY ACCIDENT

October 2020

Business Understanding

- According to NHTSA, total number of fatalities in car accident crashes increased from 26 to 36,560 starting from year 1899 to 2018
- Objective: to develop a model that could predict the severity of car accidents given by the factors affecting the collision in Seattle city

Data Understanding

- All types of collisions:
  - displayed at the intersection or mid-block of a segment
- Timeframe:
  - From January 2004 to May 2020.
- Data source:
  - Seattle Police Department (SPD) and Traffic Records group
- Original dataset:
  - 194,673 rows and 38 columns (22 attributes are object data type 16 attributes are integer or float)

Statistical Analysis

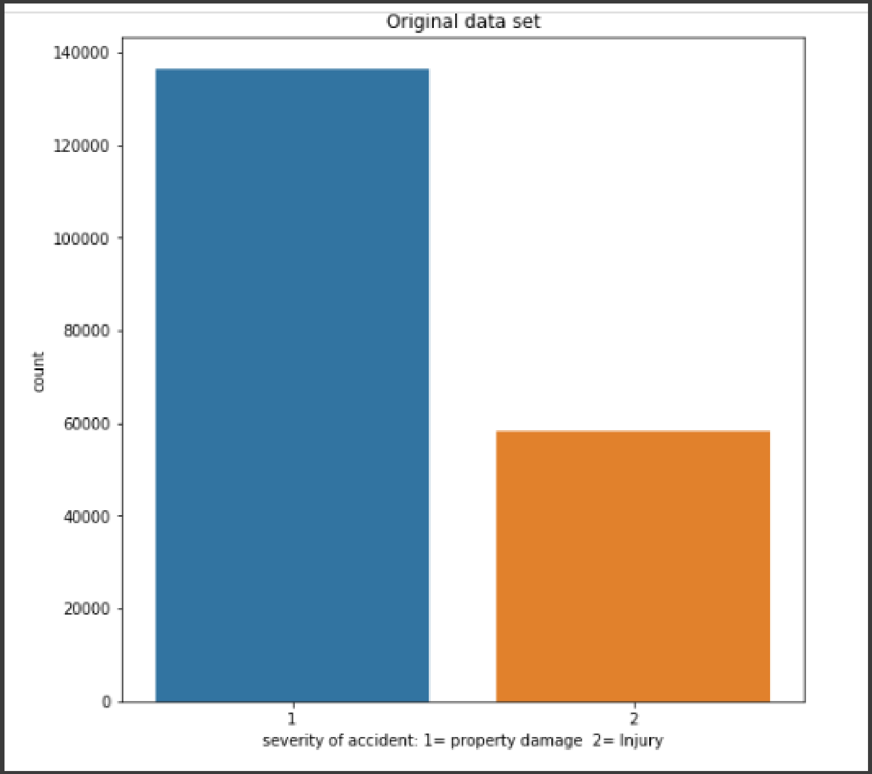
- Highly correlated attributes need to be excluded
  - “OBJECT ID”, “INCKEY” and “COLDETKEY”, SDOTCOLNUM are unique key: no impact in analysis
  - SEVERITYCODE.1” is a duplicate of “SEVERITYCODE”

	SEVERITYCODE	X	Y	OBJECTID	INCKEY	COLDETKEY	INTKEY	SEVERITYCODE.1	PERSONCOUNT	PEDCOUNT	PEDCYLCOUNT	VEHCOUNT	SDOT_COLCODE	SDOTCC
SEVERITYCODE	1	0.01	0.018	0.02	0.022	0.022	0.0066	1	0.13	0.25	0.21	-0.055	0.19	
X	0.01	1	-0.16	0.01	0.01	0.01	0.12	0.01	0.013	0.011	-0.0018	-0.012	0.011	
Y	0.018	-0.16	1	-0.024	-0.027	-0.027	-0.11	0.018	-0.014	0.01	0.026	0.017	-0.02	
OBJECTID	0.02	0.01	-0.024	1	0.95	0.95	0.047	0.02	-0.062	0.025	0.034	-0.094	-0.037	
INCKEY	0.022	0.01	-0.027	0.95	1	1	0.049	0.022	-0.062	0.025	0.031	-0.11	-0.028	
COLDETKEY	0.022	0.01	-0.027	0.95	1	1	0.048	0.022	-0.061	0.025	0.031	-0.11	-0.027	
INTKEY	0.0066	0.12	-0.11	0.047	0.049	0.048	1	0.0066	0.0019	-0.0048	0.00053	-0.013	0.0071	
SEVERITYCODE.1	1	0.01	0.018	0.02	0.022	0.022	0.0066	1	0.13	0.25	0.21	-0.055	0.19	
PERSONCOUNT	0.13	0.013	-0.014	-0.062	-0.062	-0.061	0.0019	0.13	1	-0.023	-0.039	0.38	-0.13	
PEDCOUNT	0.25	0.011	0.01	0.025	0.025	0.025	-0.0048	0.25	-0.023	1	-0.017	-0.26	0.26	
PEDCYLCOUNT	0.21	-0.0018	0.026	0.034	0.031	0.031	0.00053	0.21	-0.039	-0.017	1	-0.25	0.38	
VEHCOUNT	-0.055	-0.012	0.017	-0.094	-0.11	-0.11	-0.013	-0.055	0.38	-0.26	-0.25	1	-0.37	
SDOT_COLCODE	0.19	0.011	-0.02	-0.037	-0.028	-0.027	0.0071	0.19	-0.13	0.26	0.38	-0.37	1	
SDOTCOLNUM	0.0042	-0.001	-0.007	0.97	0.99	0.99	0.033	0.0042	0.012	0.021	0.035	-0.024	-0.041	
SEGLANEKEY	0.1	-0.0016	0.0046	0.028	0.02	0.02	-0.011	0.1	-0.021	0.0018	0.45	-0.12	0.21	
CROSSWALKKEY	0.18	0.014	0.0095	0.056	0.048	0.048	0.018	0.18	-0.032	0.57	0.11	-0.2	0.19	

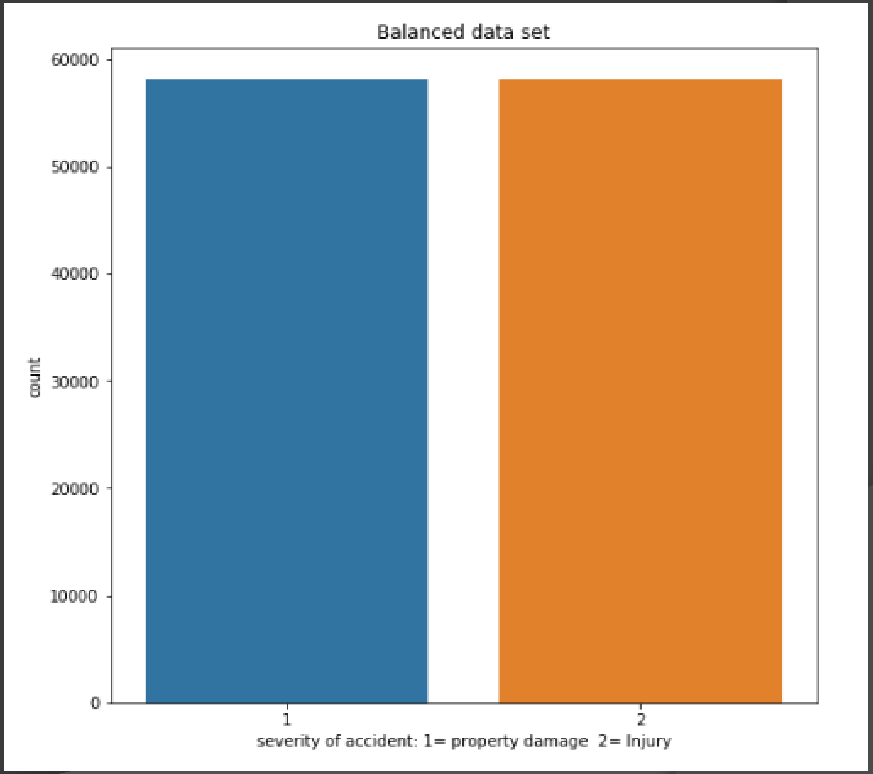
Data Preparation: Balancing data set

Class label: SEVERITYCODE

Values: Property damage (1) and Injury (2)



Before balancing data



After balancing data

Data Preparation: Handling missing values

- ROADCOND”: condition of road at the time collision
  - less than 0.1% missing values
  - Missing rows are excluded from analysis

- “LIGHTCOND”: light condition at the time collision
  - 0.1% missing values
  - Missing rows are excluded from analysis

- “SPEEDING”: missing values provide no meaning to analysis
  - Attribute is excluded from analysis

Data Preparation: Encoding

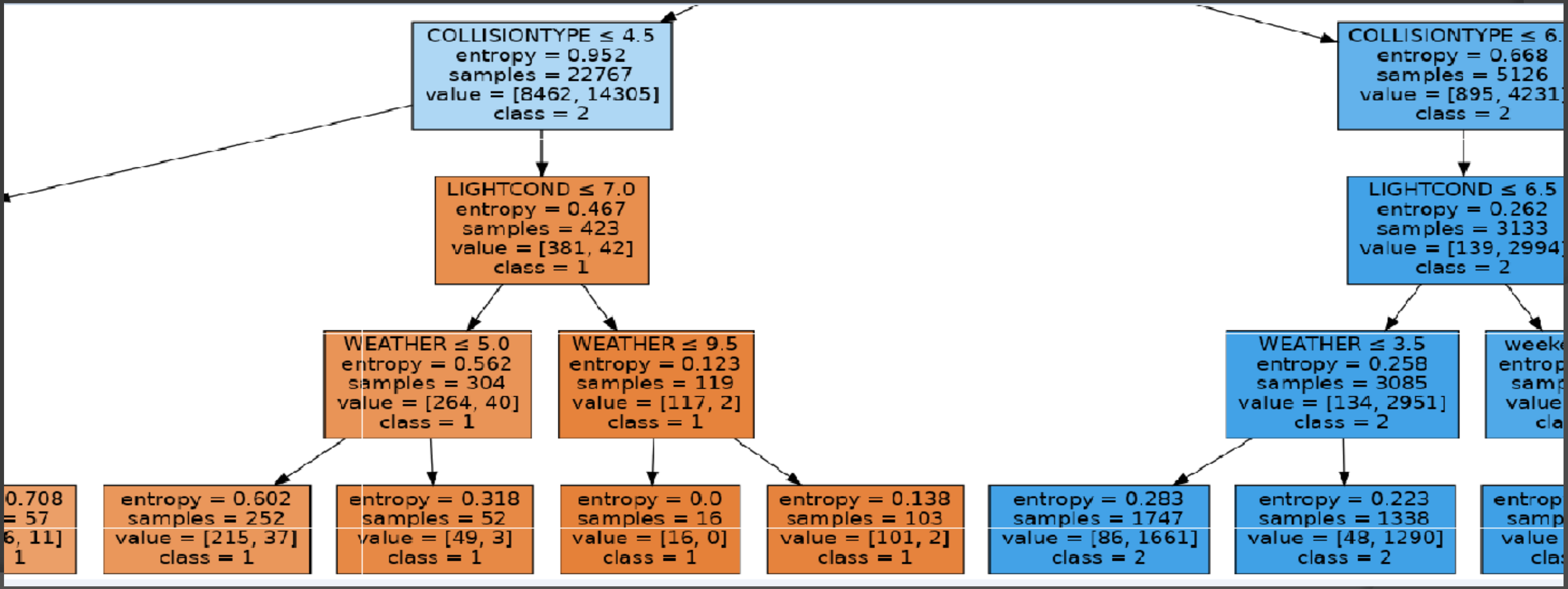
- New attributes are encoded and added to dataset
  - “weekend”
  - “dayofweek”
- “UNDERINFL”: whether or not a driver involved was under influence of drugs or alcohol
  - Encoded to “Y” and “N”
- “INATTENTIONIND”: whether or not collision was due to inattention
  - Missing values are encoded to “N”

Modeling

- Decision Tree

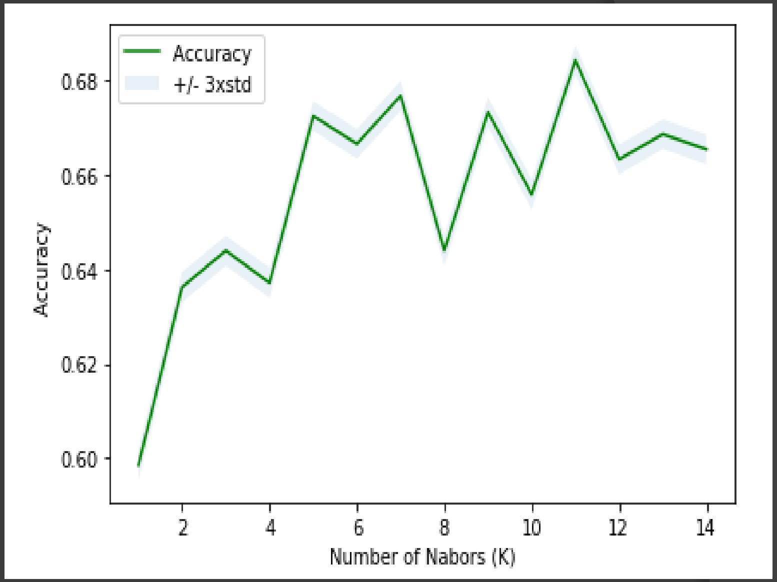
Maximum depth	3	4	5	6
Accuracy	68.4%	69.92%	70.2%	70.3%

Portion of decision tree



Modeling

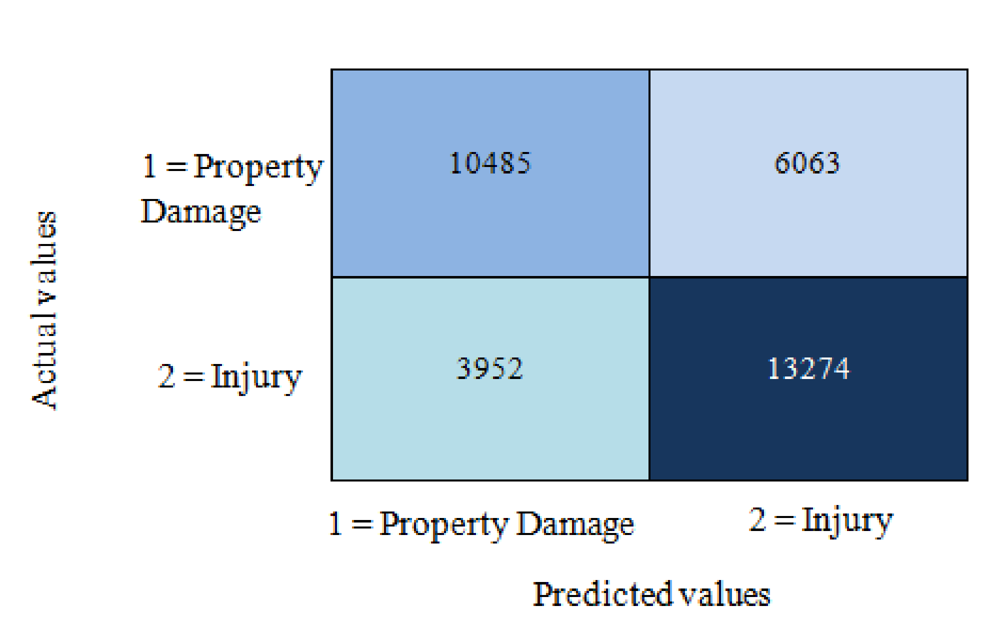
- K-Nearest Neighbor (KNN)
  - Highest accuracy: 68.4%
  - Best value of K: 11
  - Distance metric: Minkowski
- Logistic Regression
  - Algorithm: ‘Liblinear’: due to the size of dataset
  - Highest accuracy: 65.5%



KNN: Identifying values for K and respective accuracies

Result Evaluation

- Confusion Metrics
- e.g. among 10485 cases, the actual severity situation was property damage in test dataset
  - Classifier correctly predicted them as property damage
  - While the actual label of 6063 of cases were property damage, the classifier predicted as injury.
    - This is considered as the error of the model for property damage cases.



Class label	Precision	Recall	F1-score
Property damage	0.73	0.63	0.68
Injury	0.69	0.77	0.73

Result Evaluation

- Jaccard index
- F1-score
- Log loss

Algorithm	Jaccard Index	F1 score	Log loss
Decision tree	70.5%	70.2%	NA
K-Nearest Neighbour	68.4%	68.1%	NA
Logistic Regression	61.2%	61.2%	65.5%

Result Discussion

- Decision tree has performed better among all algorithms
  - highest accuracy of 70%

- Logistic regression model perform well when the training data is less, and there are large number of features
  - in this study the number of features have to be reduced due to the lack of expert knowledge

- KNN has presented slightly lower accuracy from decision tree
  - confirm that these algorithms have approximately performed the same considering same data

Conclusion

- Result shows that type of collision and location that collision occurred are the most effective factor for predicting both types of car accidents.
- Weather condition, road condition and attention of driver are the most influencing factors for accidents with injuries.
- Light condition, weather condition, and being under influence of drug or alcohol are predicted to be the most influencing factors for property damage accidents.

Conclusion (Cont.)

- Developed model can works as an assistant to help drivers in providing the required information about
  - road traffic
  - possibilities in getting into a car accident
  - identifying the severity of an accident
- Developed model gives option to drivers to either changing their travel time or the route.
- Results:
  - Reduced number of motor vehicle crashes
  - Reduced injury and fatality rate.

References

- [1] National Highway Traffic Safety Administration (NHTSA). (2020). Traffic Safetey Facts Annual Report. https://cdan.nhtsa.gov/tsftables/Fatalities%20and%20Fatality%20Rates.pdf
- [2] The CRISP-DM process model (1999), http://www.crisp-dm.org/
- [3] Everitt, Brian S.; Landau, Sabine; Leese, Morven; and Stahl, Daniel (2011) "Miscellaneous Clustering Methods", in Cluster Analysis, 5th Edition, John Wiley & Sons, Ltd., Chichester, UK
- [4] Scikit-learn. (n.d.). *-learn 0.22.2 documentation*. scikit-learn: machine learning in Python — scikit-learn 0.16.1 documentation. Retrieved October 5, 2020, from https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html#sklearn.linear_model.LogisticRegression