**INTRODUCTION TO BUSINESS PROBLEM**

This essay details the data preparation and modeling work that uses a publicly available dataset on vehicle accidents registered by the Seattle City Department of Transportation. This work benefits the American public by identifying those factors conducive to various types of traffic incidents and using this information to minimize this risk.

**ABOUT THE DATA**

In predicting the target variable of accident severity, there are 37 key input variables to be assessed for their value. Those variables and value-adding descriptors are summarized below:

1. X and Y appear to be coordinates, focused on one particular area
2. ObjectID may suggest what is involved in incident - it's a unique identifier.
3. Inckey is unique key for incident.
4. Intkey is unique key for intersection where incident happens.
5. Coldetkey corresponds to collision detail
6. Seglane and Crosswalk have keys reflecting where incident happened
7. Hit Parked Car reflects whether this happened
8. Collison code and description given by State Dept of Transportation
9. Whether speeding was factor and whether Pedestrian Right Of Way Was Not Granted
10. Lighting and road conditions
11. Type of Junction, whether or not driver was DUI
12. Whether inattention or DUI was a leading factor
13. Date and time of incident whether collision was due to lack of attention

**METHODOLOGY**

After importing the data, a review of the first few rows showed 37 features – some numerical and continuous, others alphabetical and discrete. This suggested the *need for encoding*, either ordinal or one-hot.

A review of the correlations using Seaborn’s heatmap didn’t reveal any particularly leveraging features that were numerical – neither did a review of the categorical features using Seaborn’s relplot. This confirmed that no one or few variables would be telling.

This review of the features and their value to predicting traffic severity led to a series of feature engineering and selection steps, followed by model selection and training.

FEATURE ENGINEERING AND SELECTION STEPS, WITH RATIONALE

1. Remove Report # and location since these do not give insight into the risk
2. Remove incKey – not insightful, just classification
3. Remove status
4. One-hot encode ObjectID
5. One-hot encoding of address type
6. Drop intkey -step 5 tells us if intersection
7. Crosswalkkey, seglanekey coldetkey are continuous; leave it in as-is for now - DONE
8. I'd think that road and lighting conditions would benefit from ordinal encoding but for now, sticking with one-hot
9. Drop coldesc and one-hot encoding of ST\_COLCODE – dropping repetitive detail
10. One-hot encoding of hit parked car
11. Remove coordinates – focused on particular area and location along is unhelpful
12. One-hot encode Collision key, remove Collision Description -
13. Delete rows with entries in 'EXCEPTRSNCODE', 'EXCEPTRSNDESC', then remove these columns -
14. Remove SeverCode.1, SeverityDesc(for now) – repetitive detail. We already have the desired severity.
15. One-hot encoding junction type – to incorporate this detail
16. Remove Column – not
17. Use INCDTTM to extract month --> season in year, and time --> morn,afternoon,evening. Then encode
18. One-hot-encoding weather - DONE
19. Leave'PERSONCOUNT', 'PEDCOUNT', 'PEDCYLCOUNT', 'VEHCOUNT' as-is

MACHINE LEARNING MODEL TRAINING

After completing steps above the data was split into training and testing sets.

The following models were selected: XGBoost, GradientBoost, RandomForest and DecisionTree, to fit on the training set. The model’s F1 score was then calculated on the test set. The table in the Results section shows their performance, as well as a Dummy classifier’s performance to understand baseline.

**RESULTS**

|  |  |
| --- | --- |
| CLASSIFIER | F1-SCORE ON TEST DATA |
| Dummy | 0.69 |
| XGBoost | 0.84 |
| Decision Tree | 0.79 |
| Gradient Boosting | 0.84 |
| Random Forest | 0.81 |

Based on performance above, XGBoost was selected for further hyperparameter optimization using GridSearchCV. It’s also telling that baseline performance – with the DummyClassifier – is 0.69, indicating a higher-than-normal threshold.

GridSearchCV took upwards of an hour and was taxing on the author’s computer, with minimal improvement in performance.

**DISCUSSION**

Key take-aways for Seattle administration on the variables that influence accident severity and **resulting recommendations to minimize traffic impact:**

* the month of the year is the second-most influential variable when predicting accident severity. Makes sense that certain months see poorer weather conditions, less focus by drivers. **The City of Seattle should consider safety programs, increased traffic accident avoidance efforts in problematic times of the year**
* the number of people and vehicles involved in an accident is obviously influential on severity. **The City of Seattle should consider options to limit passengers in a vehicle during certain times of the month and day, to minimize the magnitude of risk.**
* Seglane key and crosswalk key represent the area of incident – understanding these specific areas mean that the **city can install more cameras, more signage to discourage reckless behavior at these specific intersections.**
* Inattention is significantly less impactful than those above, but one that the City can still influence. **The City of Seattle should consider increasing fines for distracted driving and putting out more PSAs to discourage this behavior.**
* type of collision has the greatest influence in predicting severity. This is not surprising but not actionable by itself

**CONCLUSION**

This report has summarized approach taken in predicting traffic issues, and has delivered suggestions to minimize traffic impacts. Also listed are options to evaluate, further improve model performance.