Statistical Learning

Fatalities in car accidents, a case of study in Canada

José Chacón
Alejandra Cruces
Mario Tapia

Introduction

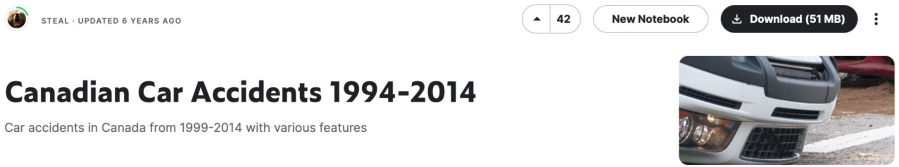
- This report presents the results of the Statistical Learning course project that aimed to analyze and predict car accidents in Canada.
- The main objective of the project was to identify the most effective way to allocate resources for traffic surveillance in Canada, by identifying the factors that influence the likelihood and severity of car accidents.

Dataset

- Canadian Car Accidents from 1994 to 2014, which was constructed by Transport Canada, a federal institution.
- The dataset is downloaded from Kaggle.
- The dataset consists of 22 categorical variables and 5,860,405 observations.
- Each observation corresponds to a person who was involved in a car crash.
- The variables can be grouped into three main categories: Accident level, Vehicle level and Person level.

Transport Canada

Canada



Dataset

ACCIDENT LEVEL

Year

Month

Day of week

Collision hour

Collision severity

Number of vehicles involved in collision

Collision configuration

Roadway configuration

Weather condition

Road surface

Road Alignment

Traffic Control

VEHICLE LEVEL

Vehicle sequence number

Vehicle type

Vehicle model year

PERSON LEVEL

Person sequence number

Person sex

Person age

Person position

Medical treatment required

Safety device used

Road user class

First look at the data

Head of the Dataset (columns 1 to 11)

C_YEAR	C_MNTH	C_WDAY	C_HOUR	C_SEV	C_VEHS	C_CONF	C_RCFG	C_WTHR	C_RSUR	C_RALN
1999	01	1	20	2	02	34	UU	1	5	3
1999	01	1	20	2	02	34	UU	1	5	3
1999	01	1	20	2	02	34	UU	1	5	3
1999	01	1	08	2	01	01	UU	5	3	6
1999	01	1	08	2	01	01	UU	5	3	6
1999	01	1	17	2	03	QQ	QQ	1	2	1

Head of the Dataset (columns 12 to 22)

C_TRAF	V_ID	V_TYPE	V_YEAR	P_ID	P_SEX	P_AGE	P_PSN	P_ISEV	P_SAFE	P_USER
03	01	06	1990	01	M	41	11	1	UU	1
03	02	01	1987	01	M	19	11	1	UU	1
03	02	01	1987	02	F	20	13	2	02	2
18	01	01	1986	01	M	46	11	1	UU	1
18	99	NN	NNNN	01	M	05	99	2	UU	3
01	01	01	1984	01	M	28	11	1	UU	1

Cleaning and Filtering of the data

Looking for **null values**

```
na_count <-sapply(cars, function(y) sum(is.na(y)))
print(na_count)
```

```
C_YEAR C_MNTH C_WDAY C_HOUR  C_SEV C_VEHS C_CONF C_RCFG C_WTHR C_RSUR C_RALN C_TRAF   V_ID V_TYPE V_YEAR   P_ID  P_SEX  P_AGE  P_PSN P_ISEV P_SAFE P_USER
     0      0      0      0      0      0      0      0      0      0      0      0      0      0      0      0      0      0      0      0      0      0
```

Looking for **blank spaces**

```
na_count <-sapply(cars, function(y) sum(y == ""))
print(na_count)
```

```
C_YEAR C_MNTH C_WDAY C_HOUR  C_SEV C_VEHS C_CONF C_RCFG C_WTHR C_RSUR C_RALN C_TRAF   V_ID V_TYPE V_YEAR   P_ID  P_SEX  P_AGE  P_PSN P_ISEV P_SAFE P_USER
     0      0      0      0      0      3      0      0      0      0      0      0      0      0      0      0      0      0      0      0      0      0
```

Cleaning and Filtering of the data

Looking at these 3 rows with blank spaces in column C_VEHS. We observed that these rows includes special characters in other columns.

C_YEAR <int>	C_MNTH <chr>	C_WDAY <chr>	C_HOUR <chr>	C_SEV <int>	C_VEHS <chr>	C_CONF <chr>	C_RCFG <chr>
2013	07	3	16	2		QQ	01
2013	07	3	16	2		QQ	01
2013	10	6	15	2		QQ	UU

Cleaning and Filtering of the data

There are some categorical variables that indicate missing information , these are:

- U/UU/UUUU : Unknown
- NN/NNNN : Data element is not applicable
- QQ : Choice is other than the preceding values
- XXXX : Jurisdiction does not provide this data element

Cleaning and Filtering of the data

We filtered the original dataframe to not include rows with these values, leaving the dataframe with 3,497,249 observations:

```
```{r replacemissing}
cars = cars %>% filter(C_YEAR != 'U' & C_YEAR != 'UU' & C_MNTH != 'U' & C_MNTH != 'UU' &
                       C_WDAY != 'U' & C_WDAY != 'UU' & C_HOUR != 'U' & C_HOUR != 'UU' &
                       C_VEHS != 'U' & C_VEHS != 'UU' & C_CONF != 'U' & C_CONF != 'UU' &
                       C_RCFG != 'U' & C_RCFG != 'UU' & C_RCFG != 'QQ' &
                       C_MNTH != 'U' & C_MNTH != 'UU' & C_WTHR != 'U' & C_WTHR != 'UU' &
                       C_RSUR != 'U' & C_RSUR != 'UU' & C_RALN != 'U' & C_RALN != 'UU' &
                       C_RALN != 'Q' &  C_TRAF != 'U' & C_TRAF != 'UU' & C_TRAF != 'QQ' &
                       P_PSN != 'NN' & P_PSN != 'QQ' & V_ID != 'U' & V_ID != 'UU' &
                       V_TYPE != 'U' & V_TYPE != 'UU' & V_YEAR != 'U' & V_YEAR != 'UU' &
                       V_YEAR != 'NNNN' & V_YEAR != 'UUUU' & V_YEAR != 'XXXX' &
                       P_ID != 'U' & P_ID != 'UU' & P_SEX != 'U' & P_SEX != 'UU' &
                       P_AGE != 'U' & P_AGE != 'UU' & P_AGE != 'NN' & P_PSN != 'U' &
                       P_PSN != 'UU' & P_ISEV != 'U' & P_ISEV != 'UU' & P_ISEV != 'N' &
                       P_SAFE != 'U' & P_SAFE != 'UU' & P_SAFE != 'NN' & P_SAFE != 'QQ' &
                       P_USER != 'U' & P_USER != 'UU')
```
```

Cleaning and Filtering of the data

We decided to focus primarily on the features at accident level. Otherwise, we would have duplicates of an accident per every individual involved.

With the creation of a unique ID per accident, we removed duplicate entries.

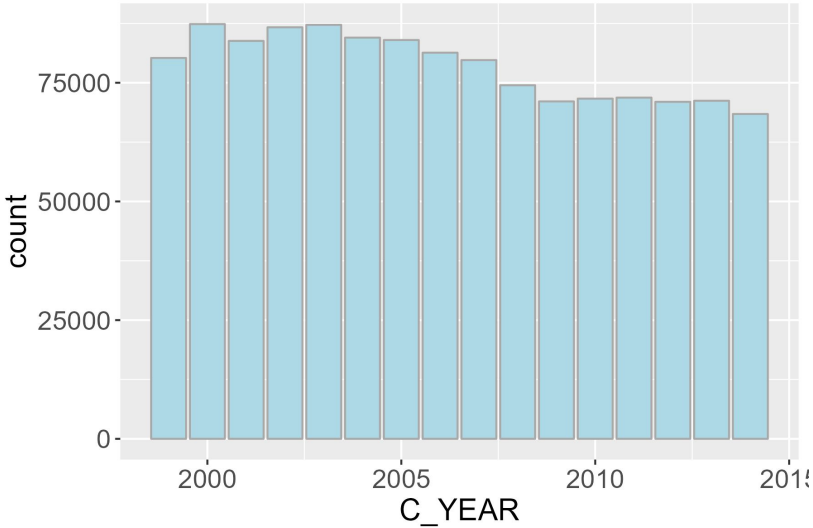
```
cars["ID"] <- paste(cars$C_YEAR, cars$C_MNTH, cars$C_WDAY, cars$C_HOUR, cars$C_SEV,
                    cars$C_VEHS, cars$C_CONF, cars$C_RCFG, cars$C_WTHR, cars$C_RSUR,
                    cars$C_RALN, cars$C_TRAF, sep = "")

#Keep one accident by ID, eliminate duplicate accidents
cars <- cars[!duplicated(cars[ , "ID"]), ]
```

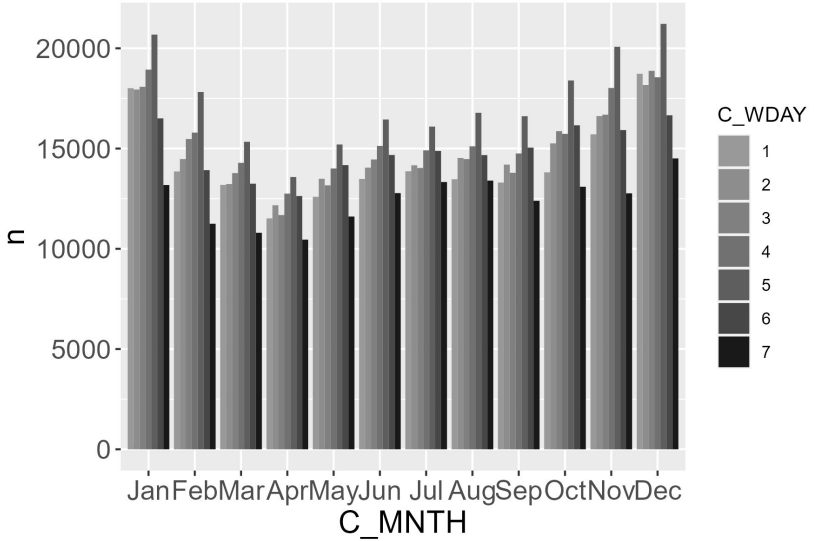
Exploration of the data

Findings

Rate of accidents over the years

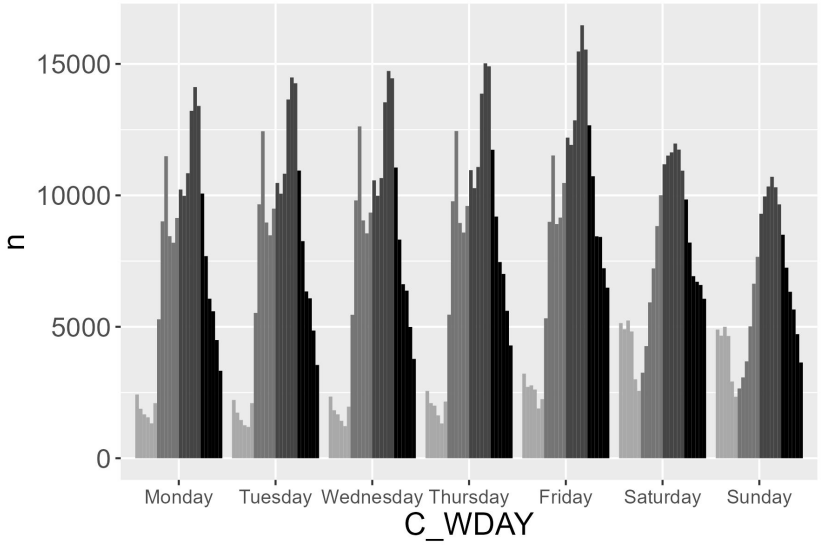


Accidents across months and days of the week

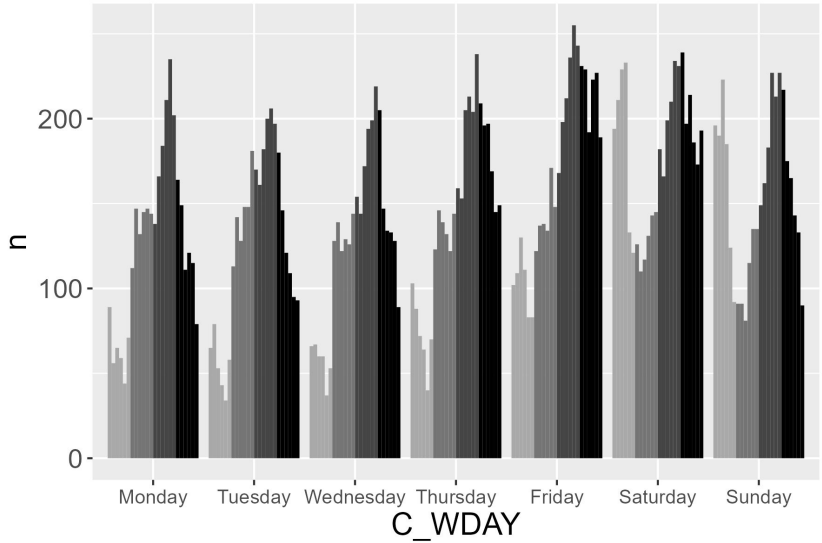


Findings

Accidents across days of the week and parts of the day



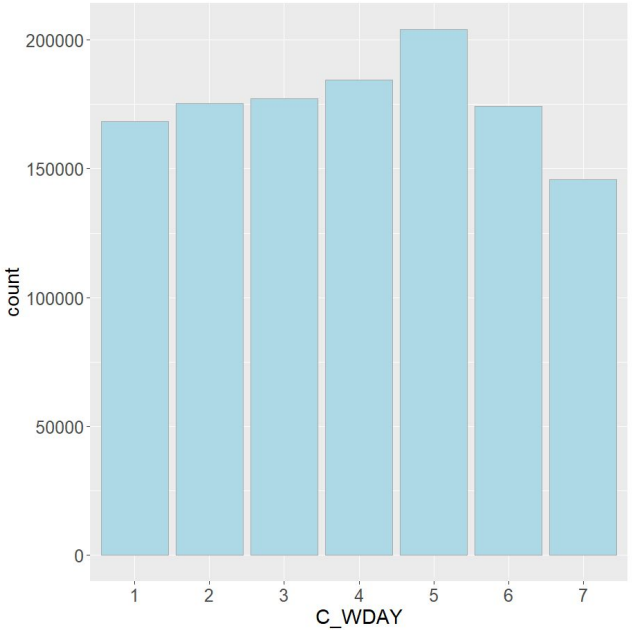
Fatal accidents across days of the week and parts of the day



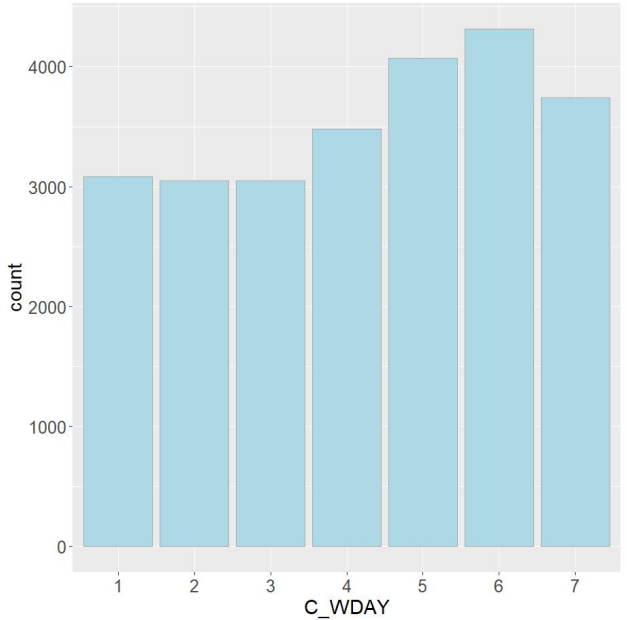
Findings

Weekday

Non-fatal accidents



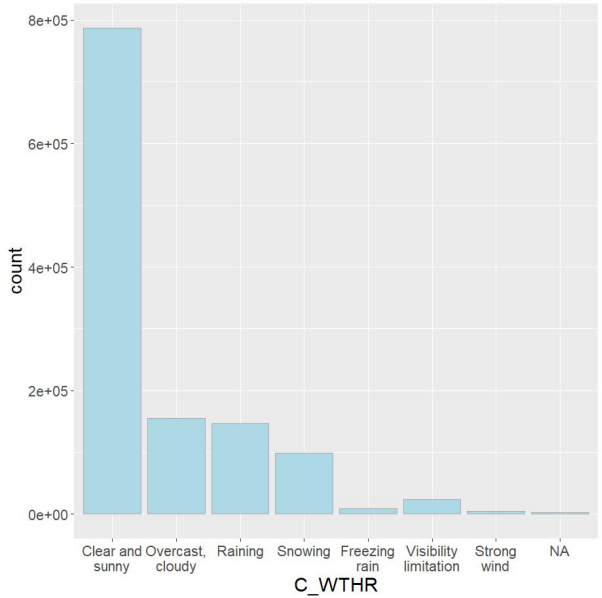
Fatal accidents



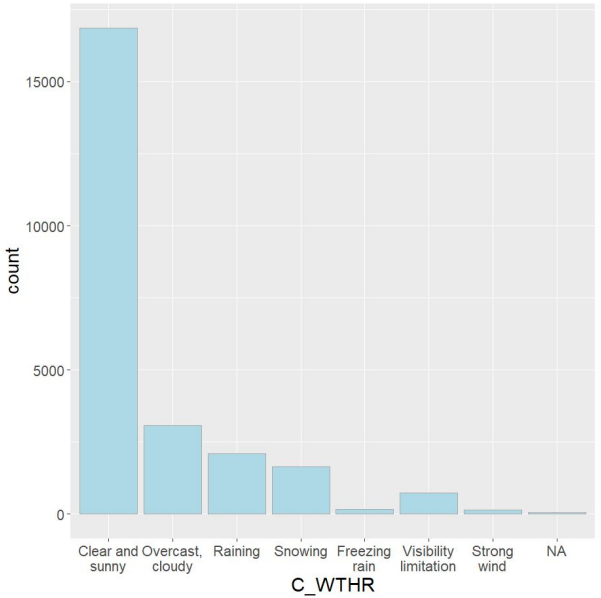
Findings

Roadway Configuration

Non-fatal accidents



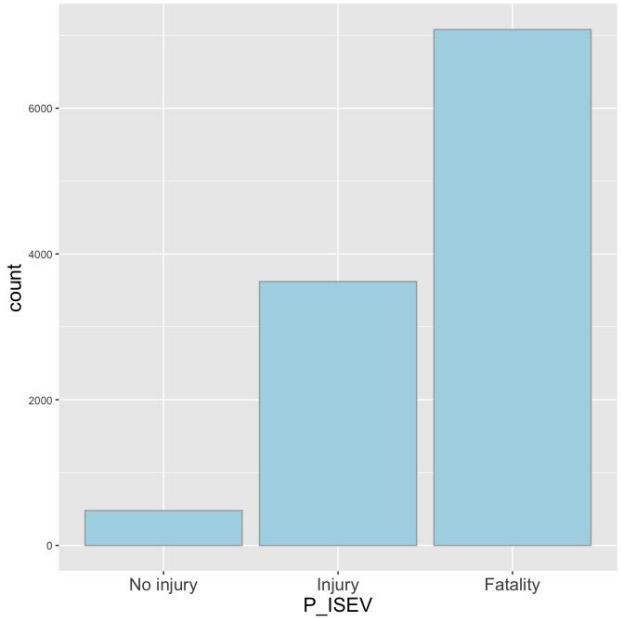
Fatal accidents



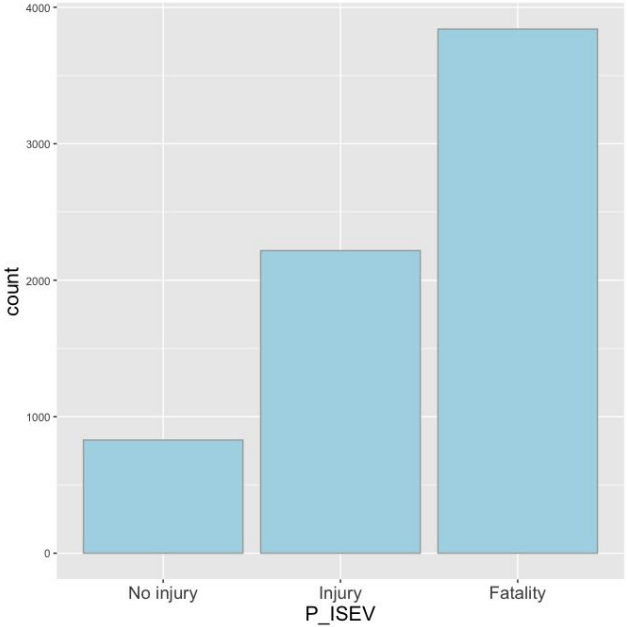
Findings

Medical treatment for person in a fatal accident by the use of safety device

Without safety device used



With safety device used



Data Science Questions

- Which time and day of the week is more probable that an accident with fatalities occurs?

- What kind of weather condition is most prevalent during car accidents with fatalities?

- Is there a correlation between the road configuration and the severity of the accident?

Data Science Questions

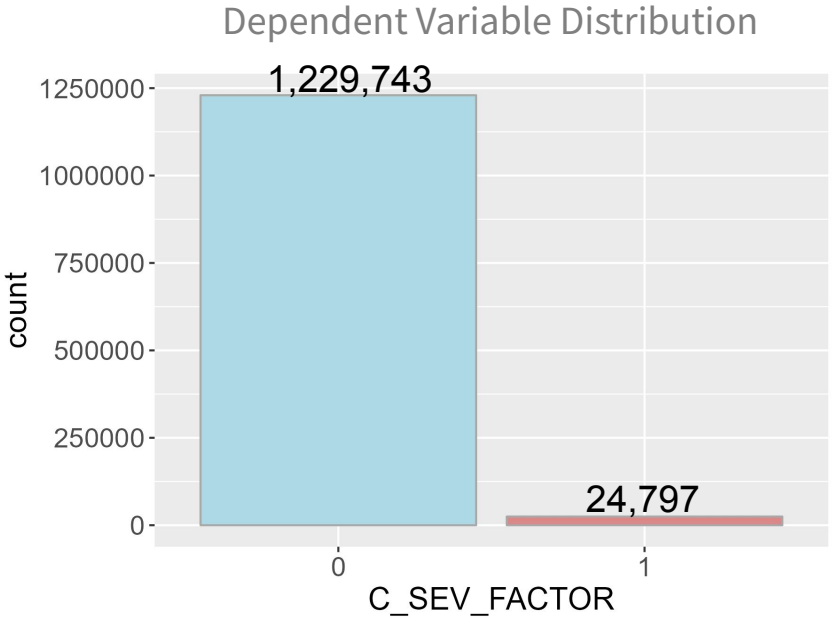
- Is there a correlation between the road alignment and the severity of the accident?

- How does the safety device affect the likelihood of casualties in traffic accidents?

- When there is elderly and underage people involved in a car accident, is it more probable to have fatalities reported?

Choice of dependent variable

We decided to use C_SEV as the outcome variable, which represents the presence or absence of a fatality in an accident.



Dataset at accident level

ACCIDENT LEVEL

Year

Month

Day of week

Collision hour

Collision severity

Number of vehicles involved in collision

Collision configuration

Roadway configuration

Weather condition

Road surface

Road Alignment

Traffic Control

VEHICLE LEVEL

Vehicle sequence number

Vehicle type

Vehicle model year

PERSON LEVEL

Person sequence number

Person sex

Person age

Person position

Medical treatment required

Safety device used

Road user class

Dataset at accident level

ACCIDENT LEVEL

~~Year~~

Month (12)

Day of week (7)

Collision hour (24)

Collision severity (2)

Number of vehicles involved in collision (54)

Collision configuration (20)

Roadway configuration (12)

Weather condition (9)

Road surface (11)

Road Alignment (8)

Traffic Control (19)

VEHICLE LEVEL

~~Vehicle sequence number~~

~~Vehicle type~~

~~Vehicle model year~~

PERSON LEVEL

~~Person sequence number~~

~~Person sex~~

~~Person age~~

~~Person position~~

~~Medical treatment required~~

~~Safety device used~~

~~Road user class~~

Feature Engineering

Dimensionality reduction

Time of the day

Early Morning (0-5), Morning (6-11), Afternoon (12-17), Evening (18-23)

Number of cars involved in an accident

1, 2, +3

Collision configuration

Categories related to the number of vehicles part of a collision (1, 2)

Roadway Configurations, Traffic control

Factors with low frequencies were grouped as a others

Variable creation

It was crucial to preserve information at person level to measure possible effects on the response variable.

P_CHILD

Is there a child between the people involved in the accident?

P_ELD

Is there an elderly person between the people involved in the accident?

C_SAFE

Is there a person not using a safe device between the people involved in the accident?

Variable creation

```
cars_copy$P_AGE <- as.numeric(cars_copy$P_AGE)

cars_copy$P_CHILD <- case_when(cars_copy$P_AGE <= 18  ~ "1",
                          cars_copy$P_AGE > 18   ~ "0")

cars_copy$P_ELD <- case_when(cars_copy$P_AGE < 65  ~ "0",
                          cars_copy$P_AGE >= 65   ~ "1")

cars_copy$C_SAFE <- case_when(cars_copy$P_SAFE == "01" | cars_copy$P_SAFE == "13"  ~ "1",
                          cars_copy$P_SAFE != "01" & cars_copy$P_SAFE != "13"  ~ "0")
```

```
cars_2 <- cars_copy[,c('ID','P_CHILD','P_ELD', 'C_SAFE')] %>%
  group_by(ID) %>%
  summarise_all(max)

#Merge the new variables with our dataset
cars <- merge(cars,cars_2, by = "ID", all.x = TRUE, all.y = FALSE)

#Factor this new variables
cars$P_ELD  <- factor(cars$P_ELD)
cars$P_CHILD  <- factor(cars$P_CHILD)
cars$C_SAFE  <- factor(cars$C_SAFE)
```

Dataset at accident level

ACCIDENT LEVEL

~~Year~~

Month (12)

Day of week (7)

Collision hour (**24 → 4**)

Collision severity (2)

Number of vehicles involved in collision (**54 → 3**)

Collision configuration (**20 → 5**)

Roadway configuration (**12 → 3**)

Weather condition (**9 → 8**)

Road surface (**11 → 10**)

Road Alignment (**8 → 6**)

Traffic Control (**19 → 6**)

VEHICLE LEVEL

~~Vehicle sequence number~~

~~Vehicle type~~

~~Vehicle model year~~

PERSON LEVEL

~~Person sequence number~~

~~Person sex~~

~~Person age~~

~~Person position~~

~~Medical treatment required~~

Safety device used (**2**)

~~Road user class~~

Presence of Child (**2**)

Presence of Elderly (**2**)

Final Dataframe

C_SEV_FACTOR	C_MNTH	C_WDAY	C_HOUR_AGR	C_VEHS_AGR	C_CONF_AGR	C_TRAF_AGR
0	jan	mon	Early Morning	1	single	no control present
0	jan	mon	Early Morning	1	single	no control present
0	jan	mon	Early Morning	1	single	no control present
0	jan	mon	Early Morning	2	two same direction	traffic signals fully operational
0	jan	mon	Early Morning	2	two same direction	traffic signals fully operational
0	jan	mon	Early Morning	2	two different direction	traffic signals fully operational

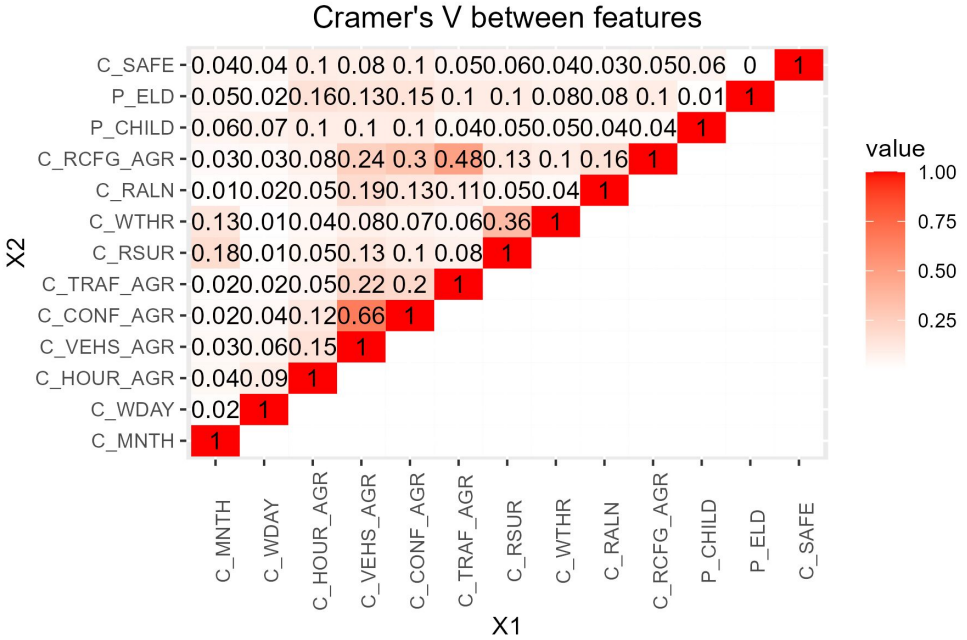
C_RSUR	C_WTHR	C_RALN	C_RCFG_AGR	P_CHILD	P_ELD	C_SAFE
slush	clear	1	non-intersection	1	0	0
icy	clear	1	non-intersection	0	0	1
dry	overcast	1	non-intersection	0	0	0
icy	clear	1	at an intersection	1	0	0
icy	snowing	1	at an intersection	0	0	0
snow	snowing	1	at an intersection	0	0	0

Models

- Logistic Regression.
- LR with L1 Regularization.
- KNN.

Logistic Regression

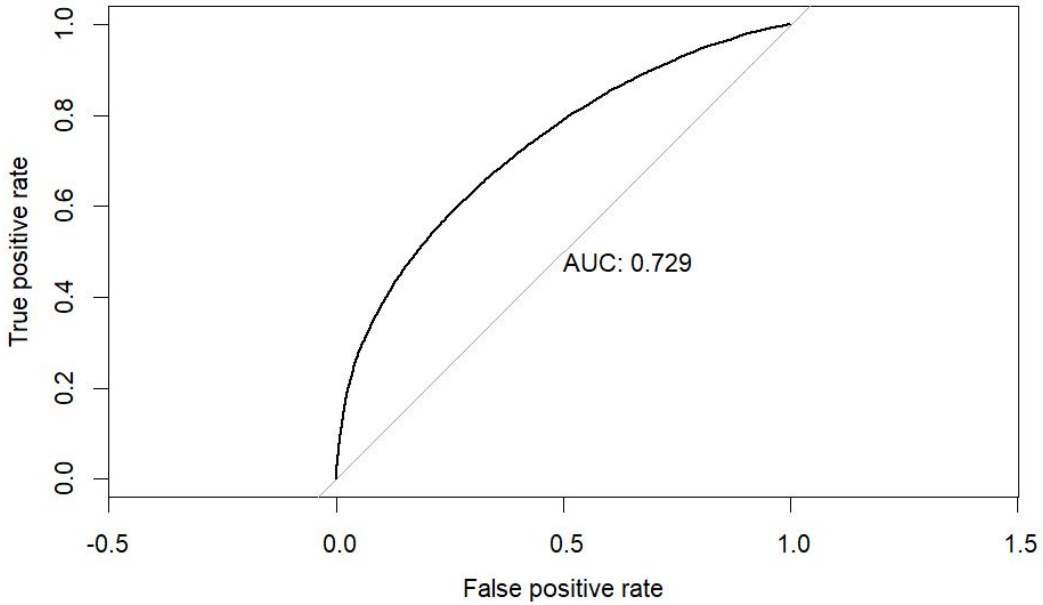
Multicollinearity



- Collision configuration **(C_CONF_AGR)** and Number of vehicles involved in collision **(C_VEHS_AGR)**: The number of vehicles determines the configuration of the collision.
- Weather conditions **(C_WTHR)** and Road surface (C_RSUR): The type of weather explains in most cases the characteristics of the surface.
- Traffic control (C_TRAF) and Roadway configuration **(C_RCFG_AGR)**: The type of roadway could explain the presence or absence of traffic signs.

AUC and ROC

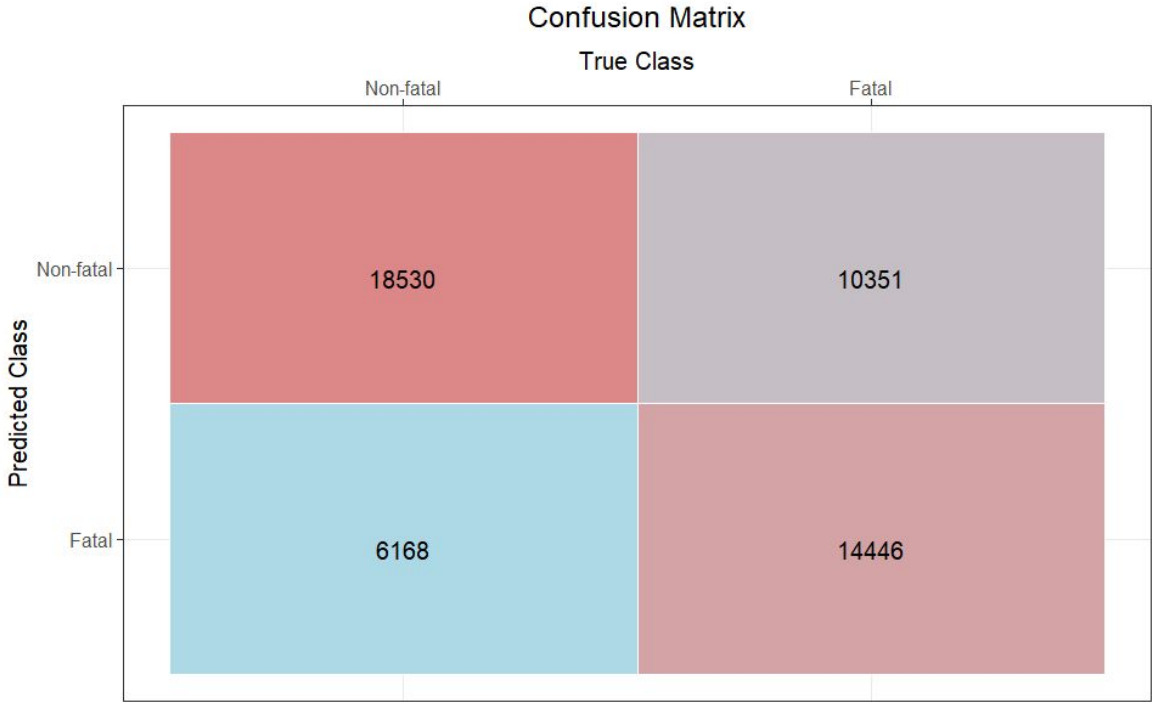
Chosen threshold: 0.5167



Metrics

AUC	0.7293
Specificity	0.7506
Sensitivity	0.5825

Confusion Matrix

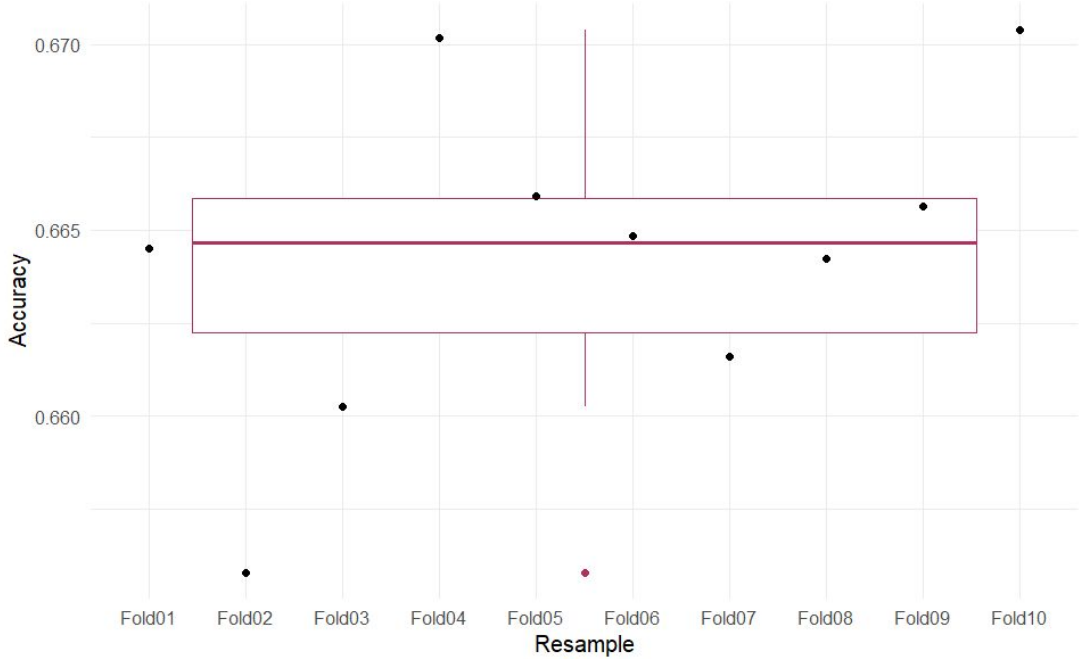


Metrics

Accuracy	0.6661
Specificity	0.7506
Sensitivity	0.5825

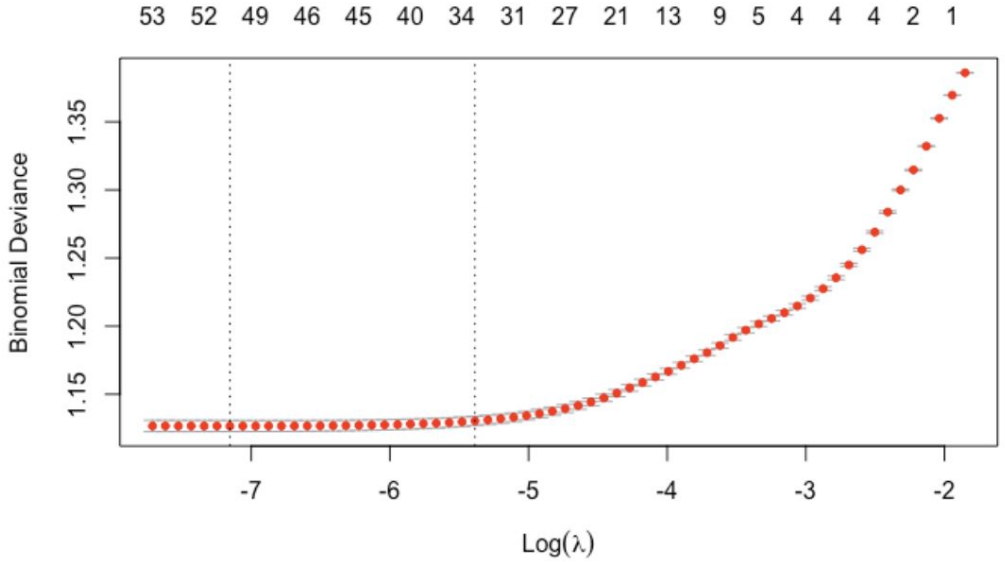
K-Fold Cross Validation

- Results for K = 10
- Our model performs well on the training set with K fold cross validation. It has an average accuracy of about 67% and a low variability.



Shrinkage regression

- We obtained a cross-validation error minimum at lambda 7.8144 ×10−4, which improves the accuracy by about 3 percentage points.
- This value for lambda obtained an accuracy of 0.6908 for the total dataset with undersampling.



Stability of parameters with undersampling

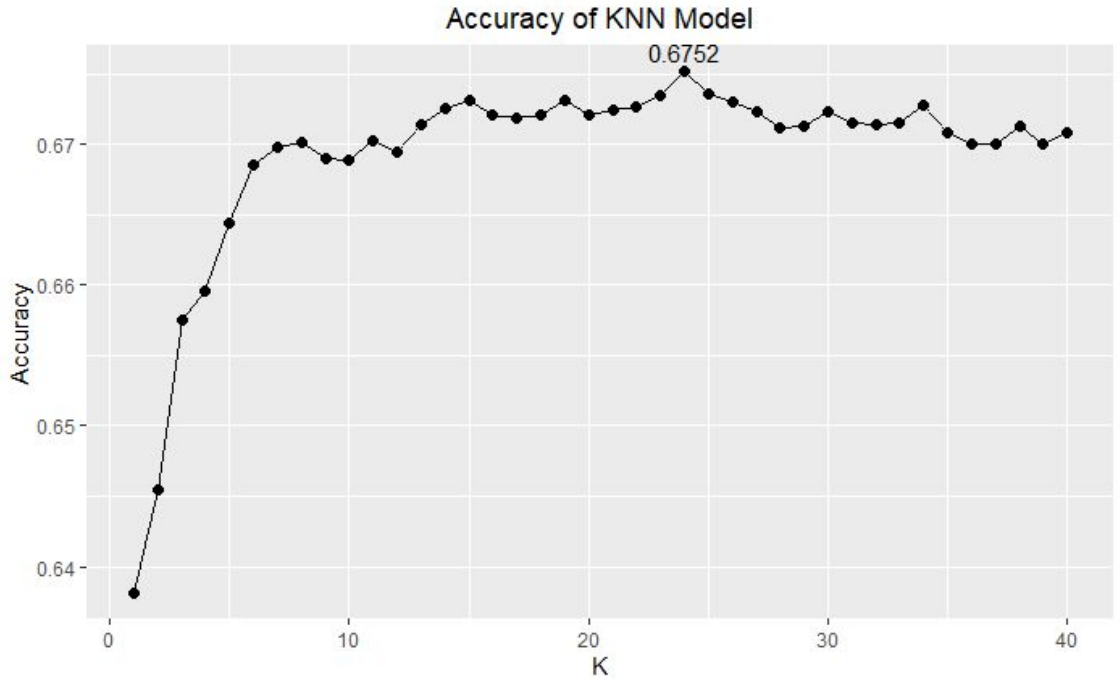
We undersampled the data set using 10 different seeds, fit each model using glm, and calculate the mean and standard deviation using each of the ten values for every level.

```
##                 coef_mean    coef_sd
## C_MNTHfeb      0.01456178 0.03550002
## C_MNTHmar      0.08454909 0.03608395
## C_MNTHapr      0.21182099 0.02307641
## C_MNTHmay      0.33358703 0.03147296
## C_MNTHjun      0.34846328 0.03063728
## C_MNTHjul      0.49792401 0.03378087
## C_MNTHaug      0.48369190 0.03333728
## C_MNTHsep      0.40551947 0.03229313
## C_MNTHoct      0.35285223 0.02420346
## C_MNTHnov      0.22460281 0.02777406
## C_MNTHdic      0.16174686 0.02699064
## C_WDAYtue     -0.02002679 0.02046616
```

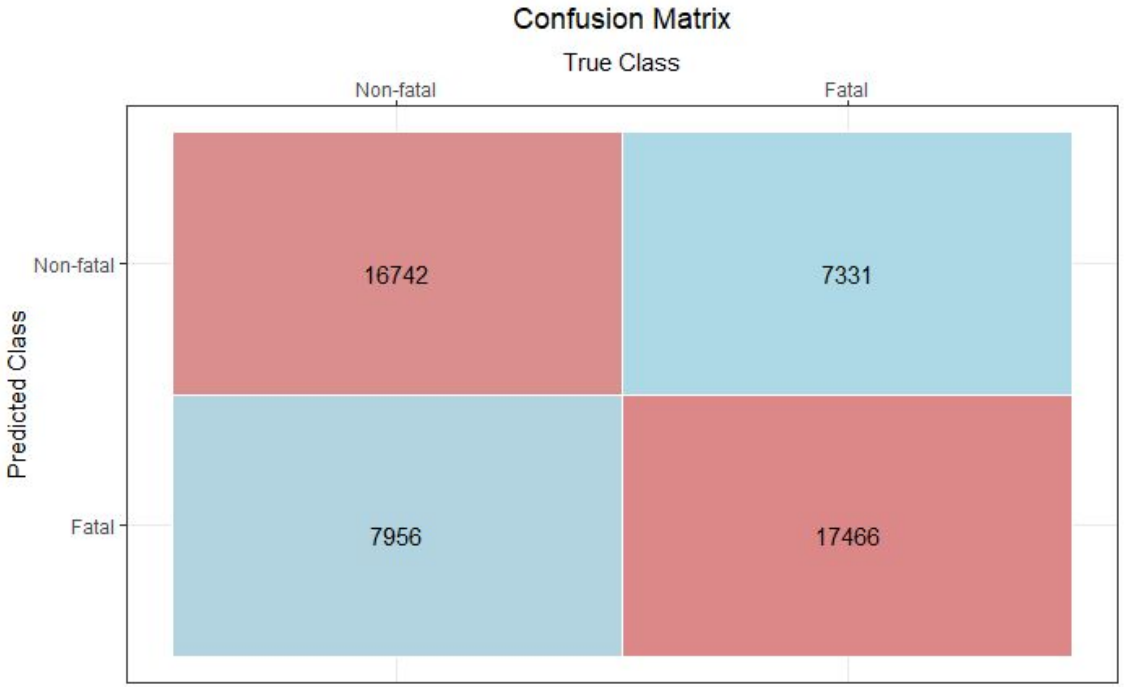
KNN

KNN

- Picking an optimal value for K



KNN



Metrics

Accuracy	0.6911
Specificity	0.7043
Sensitivity	0.6779

Results

Interpreting results to answer DS Questions

➢ Which time and day of the week is more probable that an accident with fatalities occurs?

- Compared to the Afternoon, Early morning and Evening have higher odds of fatal accidents by 1.63 and 1.28 times respectively.
- Compared to Mondays, Saturdays and Sundays have higher odds by 1.11 and 1.13 times respectively.
- Compared to January, July, August and September have the highest odds increases by 1.67, 1.59 and 1.47 times respectively.
- **Campaign**: Summer months, the weekends and the early morning hours.

Interpreting results to answer DS Questions

- What kind of weather condition is most prevalent during car accidents with fatalities?

  - Compared to clear and sunny days, the odds of at least one fatal in an accident are 1.52 and 1.49 times higher when there is visibility limitation or strong wind, respectively.
  - The odds of a fatal crash are lower when the weather is overcast, rainy, or snowy, by factors of 0.91, 0.71, and 0.85 respectively.
  - **Campaign**: Educate drivers on how to cope with low visibility or high wind situations.

Interpreting results to answer DS Questions

- Is there a correlation between the road configuration and the severity of the accident?
  - Compared to non-intersection accidents, the odds of having at least one fatality are 0.46 times lower for intersection accidents, and 0.51 times lower for other road configurations.
  - **Campaign**: More traffic signals should be installed in non-intersection roads.

- Is there a correlation between the road alignment and the severity of the accident?
  - Compared to straight flat roads, curved leveled roads have a 1.46 times higher risk, while curved roads with gradient have a 1.62 times higher risk.
  - **Campaign**: Raise awareness of the dangers of these roads and to install more traffic signs on them.

Interpreting results to answer DS Questions

- How does the safety device affect the likelihood of casualties in traffic accidents?

  - The odds of having at least a fatal outcome in accident increases by a factor of 5.02 when there is someone without a safety device.
  - **Campaign**: Promote the use of safety devices among people and companies.

Interpreting results to answer DS Questions

- When there is underage and elderly people involved in a car accident, is it more probable to have fatalities reported?

  - The analysis shows that the presence of underage people reduces the odds of a fatal outcome by a factor of 0.70 (idea: check age criteria)
  - The presence of elderly people increases the odds of a fatal outcome by a factor of 1.60.
  - **Campaign**: More attention and care should be given to elderly people when they use any mode of transportation.

Best Models Results

Model	Accuracy
Logistic Regression	0.6661
KNN	0.6911
LR with L1 Regularization	0.6866

Conclusions

- Applying Logistic regression, we conclude that the main variable that contributes to a casualty in a car accident is the Safety device.
- The promotion of safety devices, as seat belts, child restraints, helmet, among other, is very important to decrease the number of casualties in car accidents.
- Even though KNN has higher accuracy, the interpretability of the parameters obtained with Logistic Regression is more important for this project.

Statistical Learning

Fatalities in car accidents, a case of study in Canada

José Chacón
Alejandra Cruces
Mario Tapia