Predicting accident severity

STEVEN WAMBUA

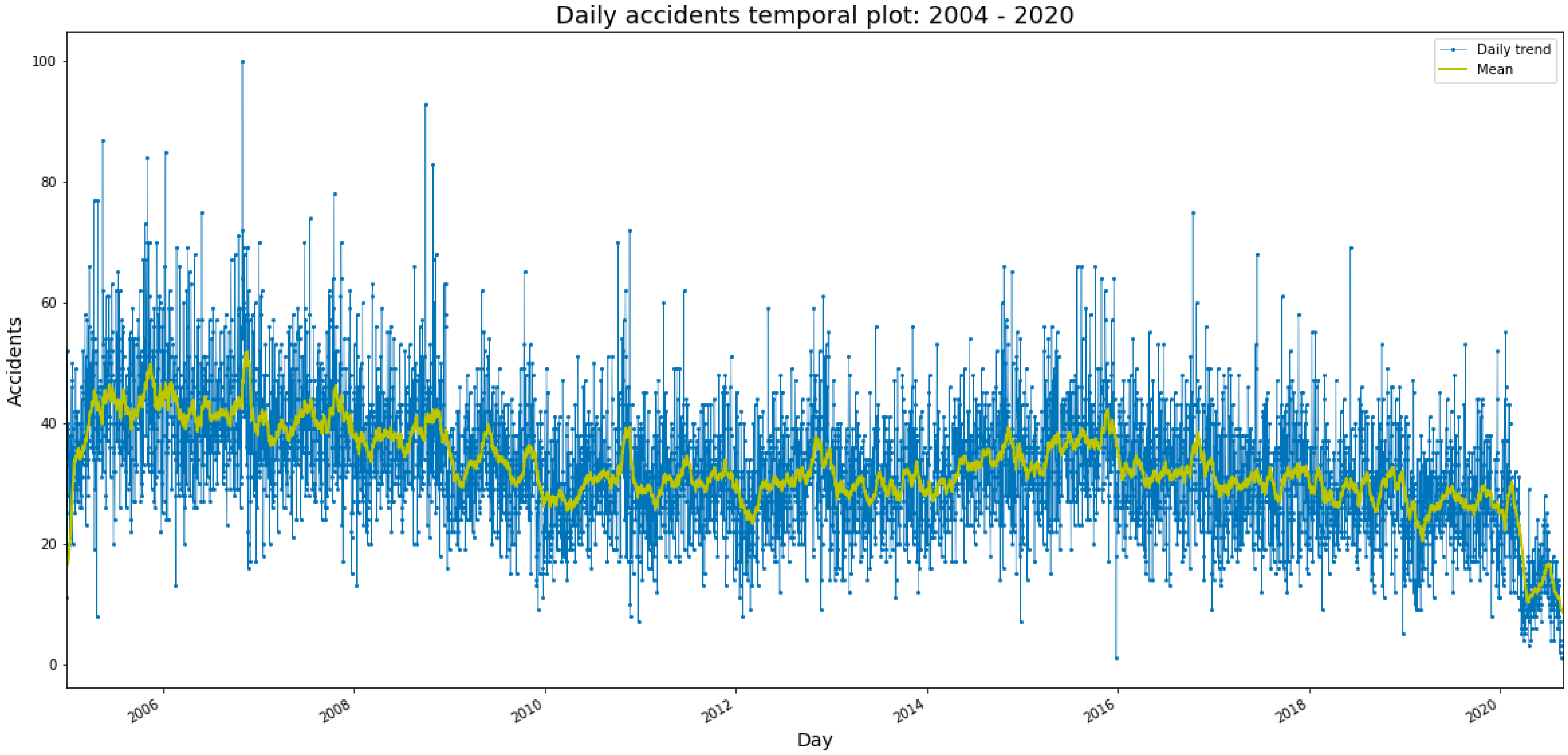
Background

- The deaths and injuries from traffic accidents are now a world phenomenon
- Every year the lives of approximately 1.35 million people are cut short as a result of a road traffic crash (WHO)
- Motorists, people who use public transport, police and medical personnel are usually inconvenienced greatly when accidents occur and especially severe accidents.
- It would be possible to build a machine learning algorithm to predict severe accidents using various factors that are determined to be related to a severe accident occurring and help people plan accordingly

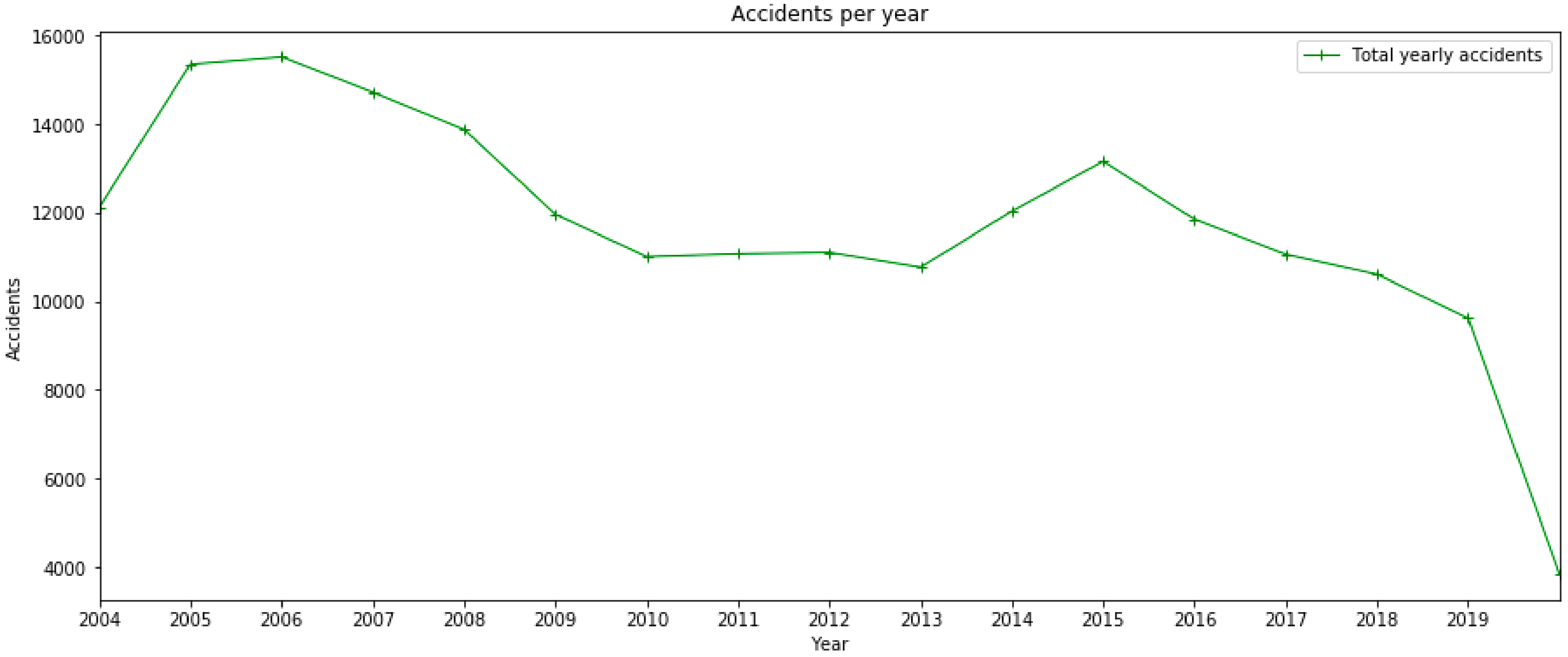
Data

- A dataset by SDOT Traffic Management Divisions, Traffic Records Group in Seattle was downloaded using this link.
- Data is updated weekly and captures all types of collisions since 2004 to present.
- A record of severity of the collision, "SEVERITYCODE" is provided alongside other 39 variables related to a given collision
- The dataset has a record of 221266 collisions.

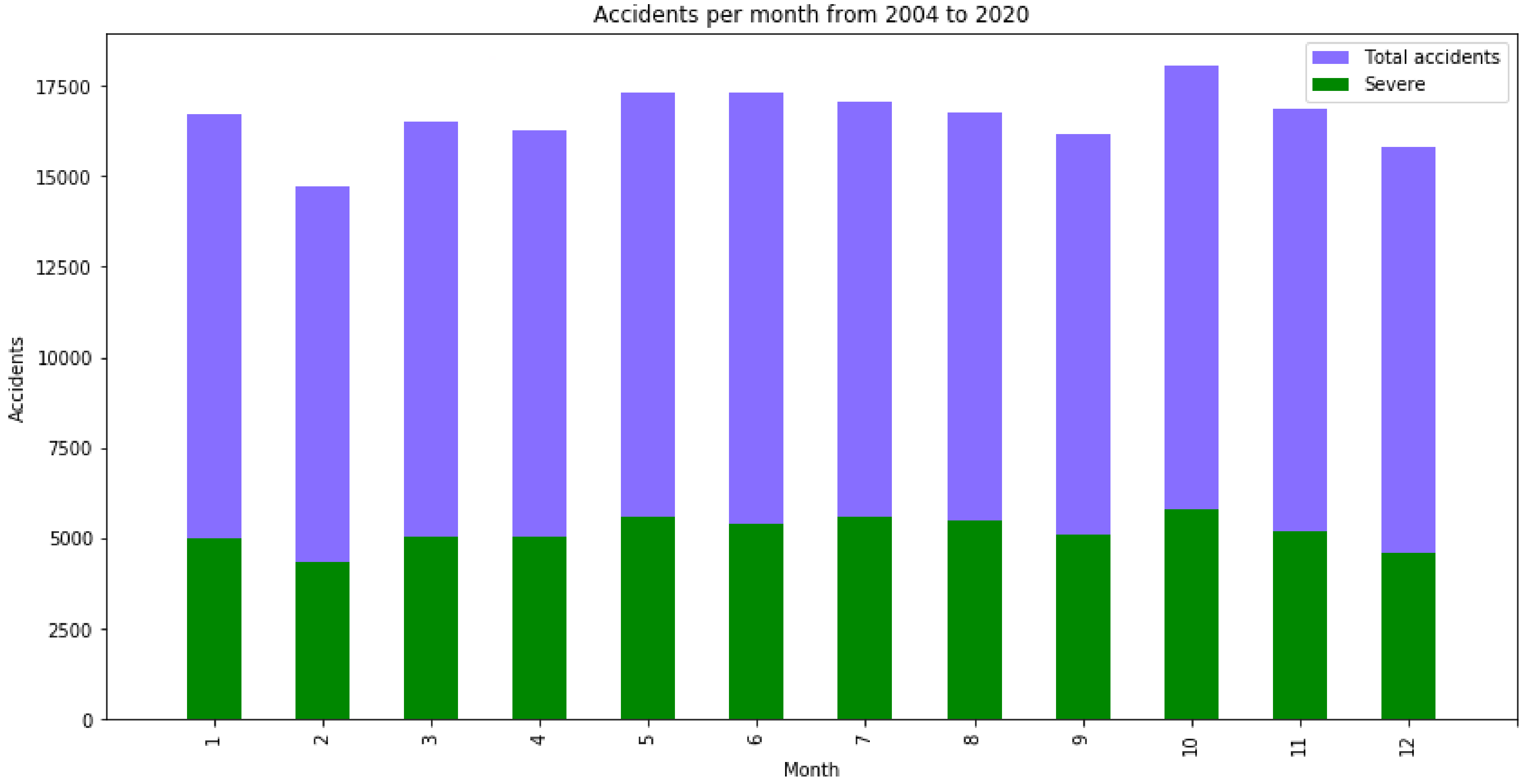
Temporal trend: Daily accidents



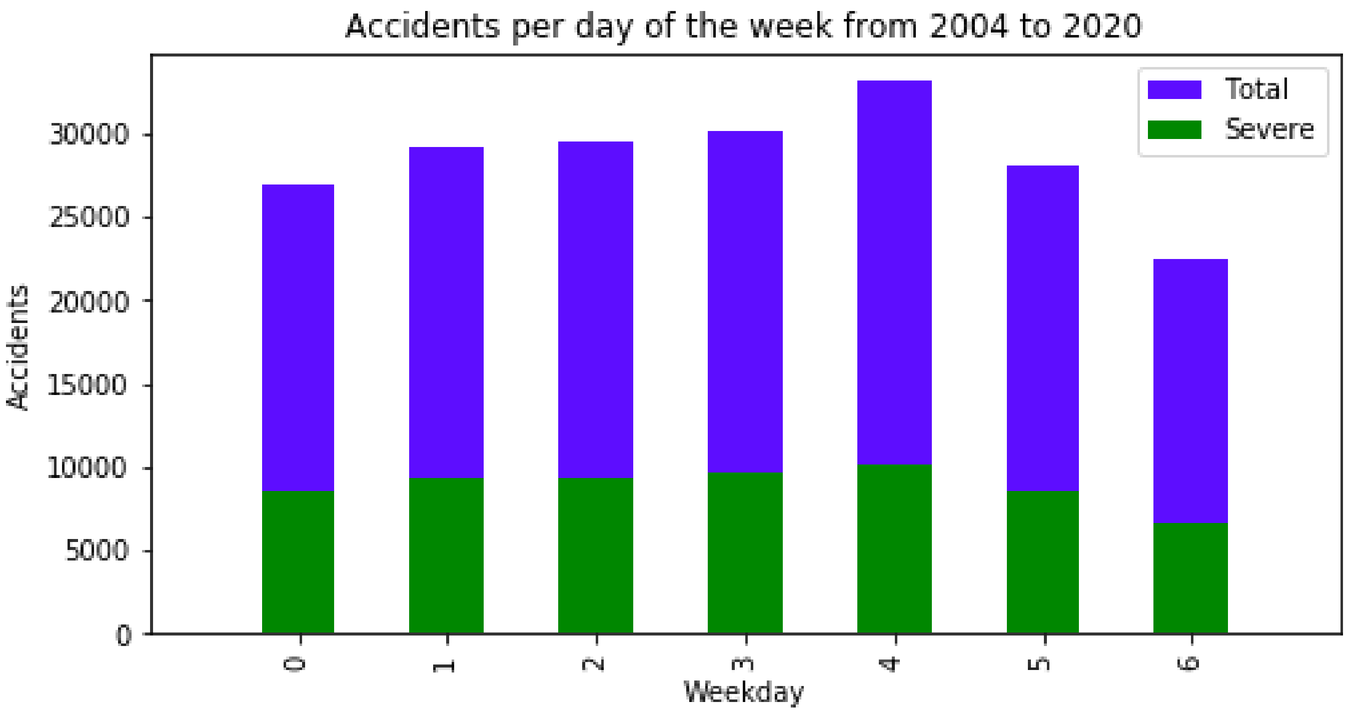
Temporal trend: Yearly accidents



Temporal trend: Monthly accidents



Temporal trend: Weekday accidents



Correlation between variables

	SEVERITYCODE	PERSONCOUNT	PEDCOUNT	PEDCYLCOUNT	VEHCOUNT	INJURIES	SERIOUSINJURIES	FATALITIES	SDOT_COLCODE	UNDERINFL	SEGLANEKEY	CROSSWAL
SEVERITYCODE	1	0.370575	0.260548	0.203206	0.38489	0.700391	0.280069	0.168462	0.311591	0.0966905	0.0974853	0.1
PERSONCOUNT	0.370575	1	0.0116041	-0.00916478	0.558807	0.319327	0.107266	0.0463413	0.008312	0.0538994	-0.00834905	-0.01
PEDCOUNT	0.260548	0.0116041	1	-0.0158227	-0.154356	0.167294	0.132543	0.0728061	0.260955	0.0319634	0.00126502	0.5
PEDCYLCOUNT	0.203206	-0.00916478	-0.0158227	1	-0.15074	0.122495	0.0620951	0.0110157	0.369044	-0.0126391	0.456218	0
VEHCOUNT	0.38489	0.558807	-0.154356	-0.15074	1	0.142673	-0.00306759	-0.0106176	-0.0785456	0.0507102	-0.0752901	-0.1
INJURIES	0.700391	0.319327	0.167294	0.122495	0.142673	1	0.279368	0.0671804	0.138528	0.0641258	0.0593988	0.1
SERIOUSINJURIES	0.280069	0.107266	0.132543	0.0620951	-0.00306759	0.279368	1	0.173007	0.0866685	0.0472939	0.0315772	0.05
FATALITIES	0.168462	0.0463413	0.0728061	0.0110157	-0.0106176	0.0671804	0.173007	1	0.0458338	0.0431549	0.00511154	0.03
SDOT_COLCODE	0.311591	0.008312	0.260955	0.369044	-0.0785456	0.138528	0.0866685	0.0458338	1	0.115384	0.202098	0.1
UNDERINFL	0.0966905	0.0538994	0.0319634	-0.0126391	0.0507102	0.0641258	0.0472939	0.0431549	0.115384	1	-0.00595666	-0.004
SEGLANEKEY	0.0974853	-0.00834905	0.00126502	0.456218	-0.0752901	0.0593988	0.0315772	0.00511154	0.202098	-0.00595666	1	-0.003
CROSSWALKKEY	0.167778	-0.0102738	0.553888	0.10378	-0.120686	0.100689	0.0559026	0.0318511	0.187266	-0.00407634	-0.00353797	
year	-0.0257858	-0.0673016	0.0214149	0.0276288	-0.108892	-0.00442378	-0.00503714	-0.00087498	-0.0846856	-0.0139701	0.0231274	0.05
month	-0.00444333	-0.00767659	0.00431671	0.00518652	-0.0104186	0.000967585	-0.00069574	0.00441976	0.00671416	0.0010763	0.00364075	0.006
weekday	0.000308089	0.0552597	-0.0176523	-0.0235545	0.017723	0.00840898	0.00351161	0.0045957	0.0162656	0.0732051	-0.0134219	-0.01
LOCATION1	-0.0270454	-0.0123105	-0.0361607	-0.0189904	0.00448877	-0.0273793	-0.00324572	0.000387719	0.0306835	0.0069568	-0.011494	-0.04
COLLISIONTYPE1	-0.0617565	0.0375634	0.102328	-0.215409	0.113219	-0.104717	-0.0236771	-0.00183166	0.00892772	0.00761879	-0.0989727	0.03
JUNCTIONTYPE1	-0.264412	-0.15276	-0.128913	-0.0939488	-0.0859161	-0.193834	-0.03328	-0.00752107	-0.15698	0.0225128	-0.0408399	-0.1
LIGHTCOND1	-0.123226	-0.0858907	-0.0532116	0.00502127	-0.0688841	-0.0751511	-0.0299769	-0.0177534	-0.174937	-0.225657	0.00108349	-0.02
ROADCOND1	0.0973057	0.0770512	0.0197047	-0.0361814	0.173137	-0.00178518	-0.00593351	-0.00732513	0.018023	0.00582948	-0.0163792	0.01
WEATHER1	0.0573449	0.0528063	0.00320965	-0.0393614	0.182701	-0.0490784	-0.0133418	-0.0081717	-0.0429878	-0.0226485	-0.0202497	0.003
HITPARKEDCAR1	-0.201678	-0.11906	-0.0437938	-0.0365627	-0.123492	-0.106276	-0.02004	-0.0090251	-0.158467	-0.00980709	-0.0183863	-0.03
Alley	-0.0225235	-0.0244881	0.0020348	-0.00694962	-0.0208688	-0.0227097	-0.00376006	-0.0023759	-0.0879492	0.0014004	-0.00506764	-0.008
Block	-0.193272	-0.0841446	-0.144518	-0.087084	1.69995e-05	-0.166582	-0.0341676	-0.00815805	-0.00487136	0.0335608	-0.0397057	
Intersection	0.196892	0.0876876	0.144696	0.0882867	0.0027786	0.170144	0.0347779	0.00850179	0.0166685	-0.0338531	0.0405085	0.1

- Features that were significantly associated with severity were taken forward for modelling.
- Clearly we see year, month, weekday, Location and Alley have a very small correlation with severity.

Results

Algorithm	Jaccard	F1-score	LogLoss
KNN	0.993153	0.993169	NA
Decision Tree	0.990374	0.990331	NA
SVM	0.991051	0.991137	NA
Logistic Regression	0.973154	0.973286	0.0914651

- Generally all the classification models performed well
- KNN performed better

Conclusions

- The data from Seattle helps in identifying useful factors that help in building a predictive model
- The classification models predicted severity of the accidents accurately and KNN classifier is the best
- More features could be tested to assess how well they predict severity of accidents
- Missing data in some features could be corrected by ensuring these features are also captured