PREDICTIVE MODELING OF ACTUARIAL LOSS PREDICTION

Using Tree-based models

- Riwaj Pokhrel

Dataset Description

→90,000 realistic, synthetically generated worker compensation insurance policies, all of which have had an accident. Source: Kaggle

→For each record there is demographic and worker related information.

→For each record there is a text description of the accident.

Sneak peek of Data fields

	0	1	2	3	4	5	6	7
ClaimNumber	WC8285054	WC6982224	WC5481426	WC9775968	WC2634037	WC6828422	WC8058150	WC7539849
DateTimeOfAccident	2002-04-09T07:00:00Z	1999-01-07T11:00:00Z	1996-03-25T00:00:00Z	2005-06-22T13:00:00Z	1990-08-29T08:00:00Z	1999-06-21T11:00:00Z	2001-07-13T11:00:00Z	2000-03-09T09:00:00Z
DateReported	2002-07-05T00:00:00Z	1999-01-20T00:00:00Z	1996-04-14T00:00:00Z	2005-07-22T00:00:00Z	1990-09-27T00:00:00Z	1999-09-09T00:00:00Z	2001-07-23T00:00:00Z	2000-04-16T00:00:00Z
Age	48	43	30	41	36	50	39	56
Gender	M	F	M	M	M	M	M	M
MaritalStatus	M	M	U	S	M	M	M	M
DependentChildren	0	0	0	0	0	0	0	0
DependentsOther	0	0	0	0	0	0	0	0
WeeklyWages	500.0	509.34	709.1	555.46	377.1	200.0	200.0	200.0
PartTimeFullTime	F	F	F	F	F	F	F	F
HoursWorkedPerWeek	38.0	37.5	38.0	38.0	38.0	38.0	38.0	40.0
DaysWorkedPerWeek	5	5	5	5	5	5	5	5
ClaimDescription	LIFTING TYRE INJURY TO RIGHT ARM AND WRIST INJURY	STEPPED AROUND CRATES AND TRUCK TRAY FRACTURE ...	CUT ON SHARP EDGE CUT LEFT THUMB	DIGGING LOWER BACK LOWER BACK STRAIN	REACHING ABOVE SHOULDER LEVEL ACUTE MUSCLE STR...	STRUCK HEAD ON HEAD LACERATED HEAD	FINGER BRUISED AND SWOLLEN LEFT ARM	CLEANING LEFT SHOULDER SPLINTER LEFT HAND
InitialIncurredCalimsCost	1500	5500	1700	15000	2800	500	500	500
UltimateIncurredClaimCost	4748.203388	6326.285819	2293.949087	17786.48717	4014.002925	598.762315	279.068178	1877.172243

Missing value imputation

How to deal with missing values?

i) Exclude rows with missing values (only if a few values are missing)

ii) Imputation of missing values with some operation of existing values

iii) "Smart" imputing by using a trained model to guess missing values

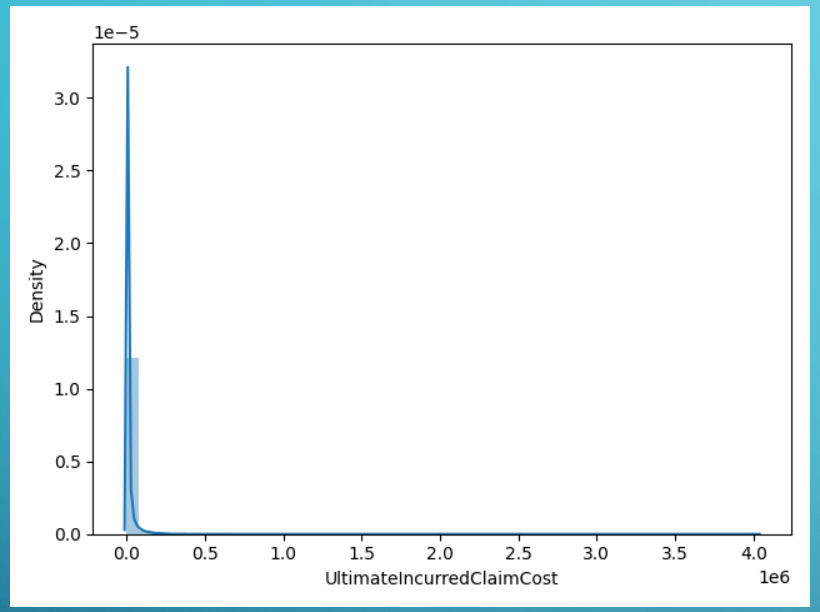
iV) Use multiple-copula imputation.

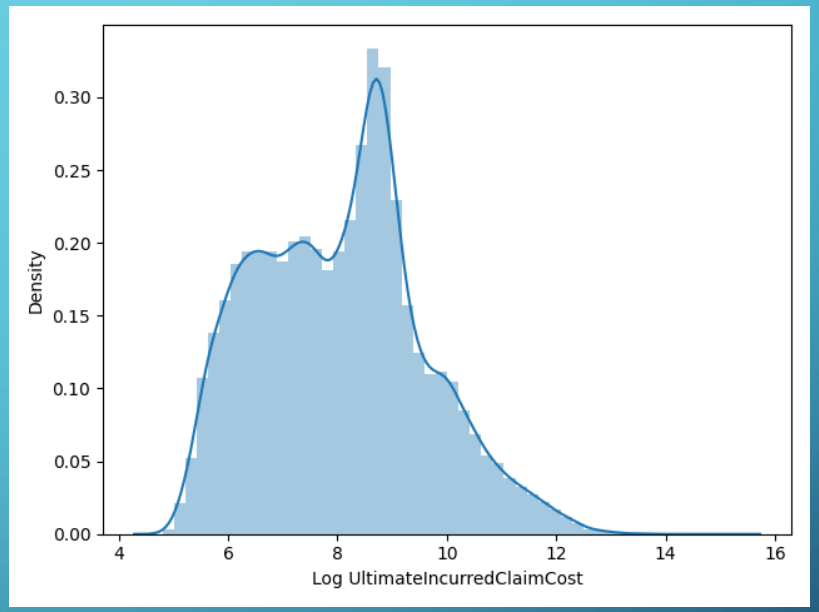
Missing values in “MaritalStatus” column:

- Imputation with mode value.

EXPLORATORY DATA ANALYSIS

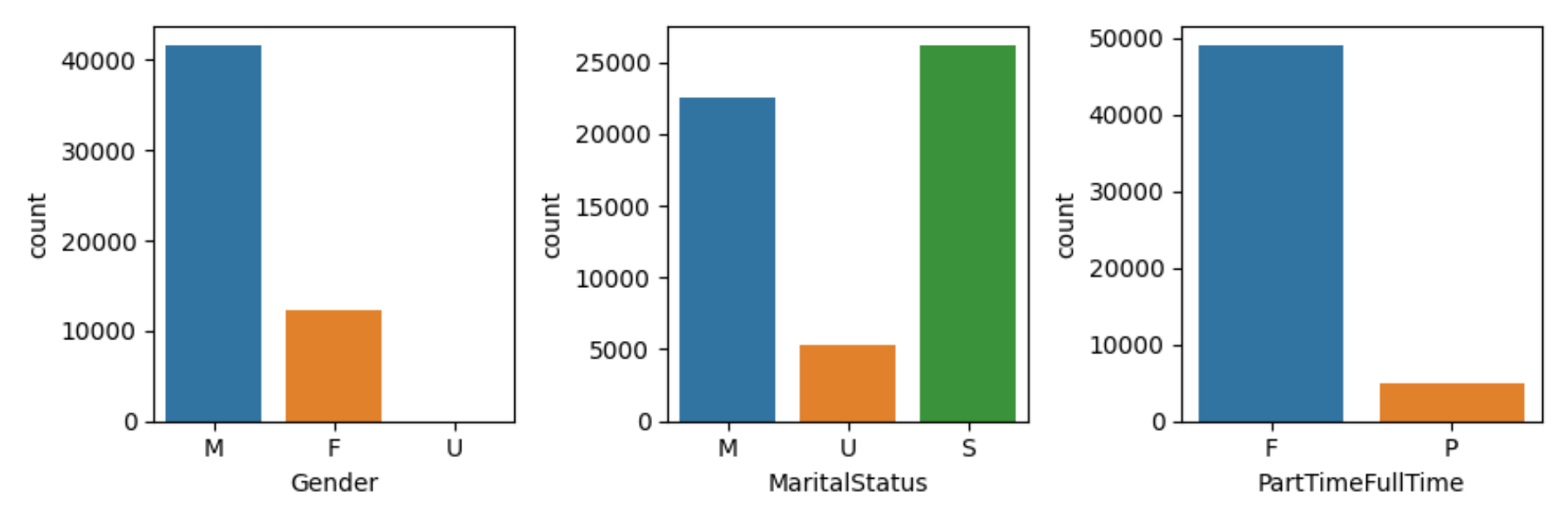
Target variable: UltimateIncurredClaimCost





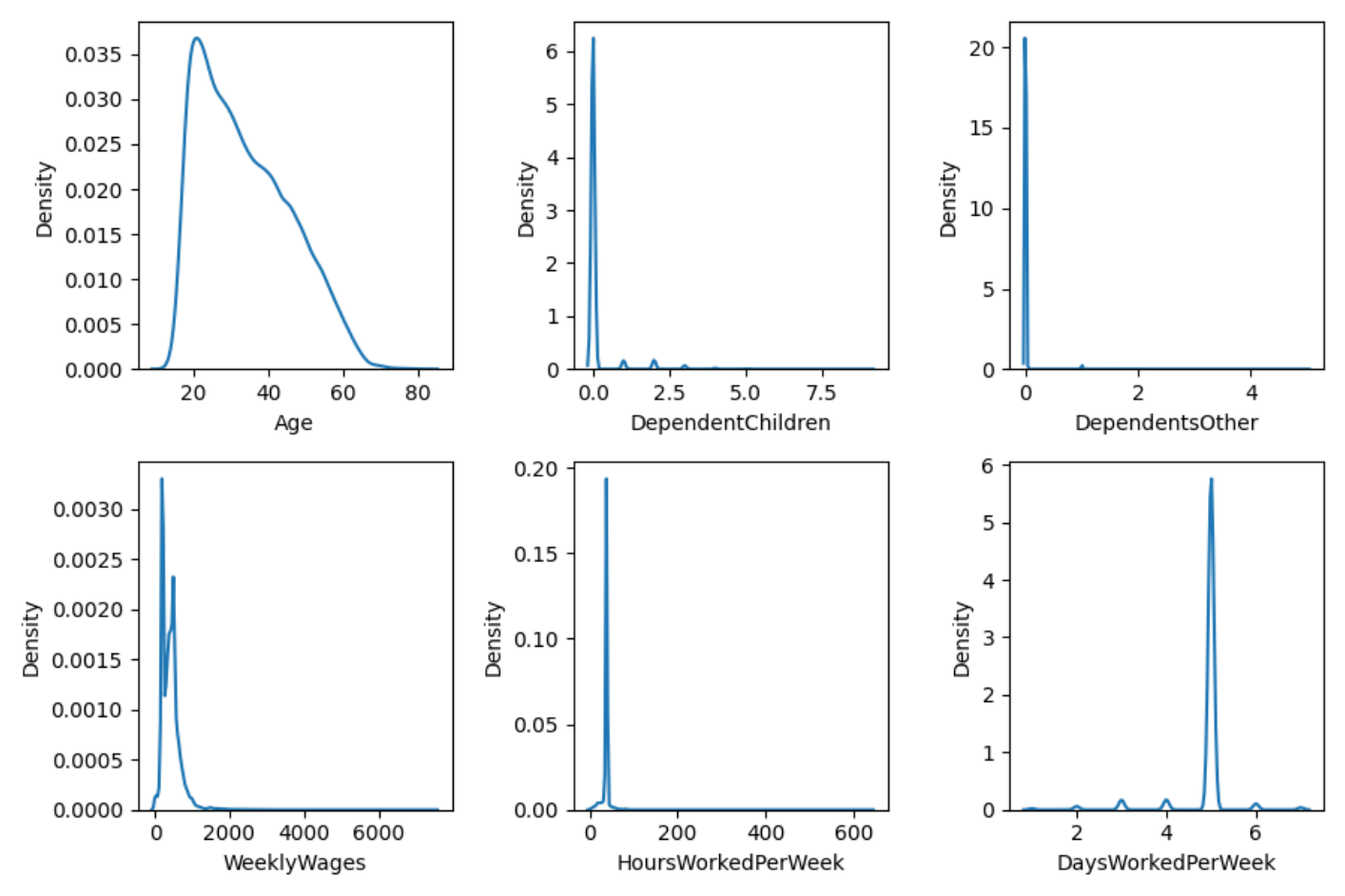
EXPLORATORY DATA ANALYSIS

Categorical variables

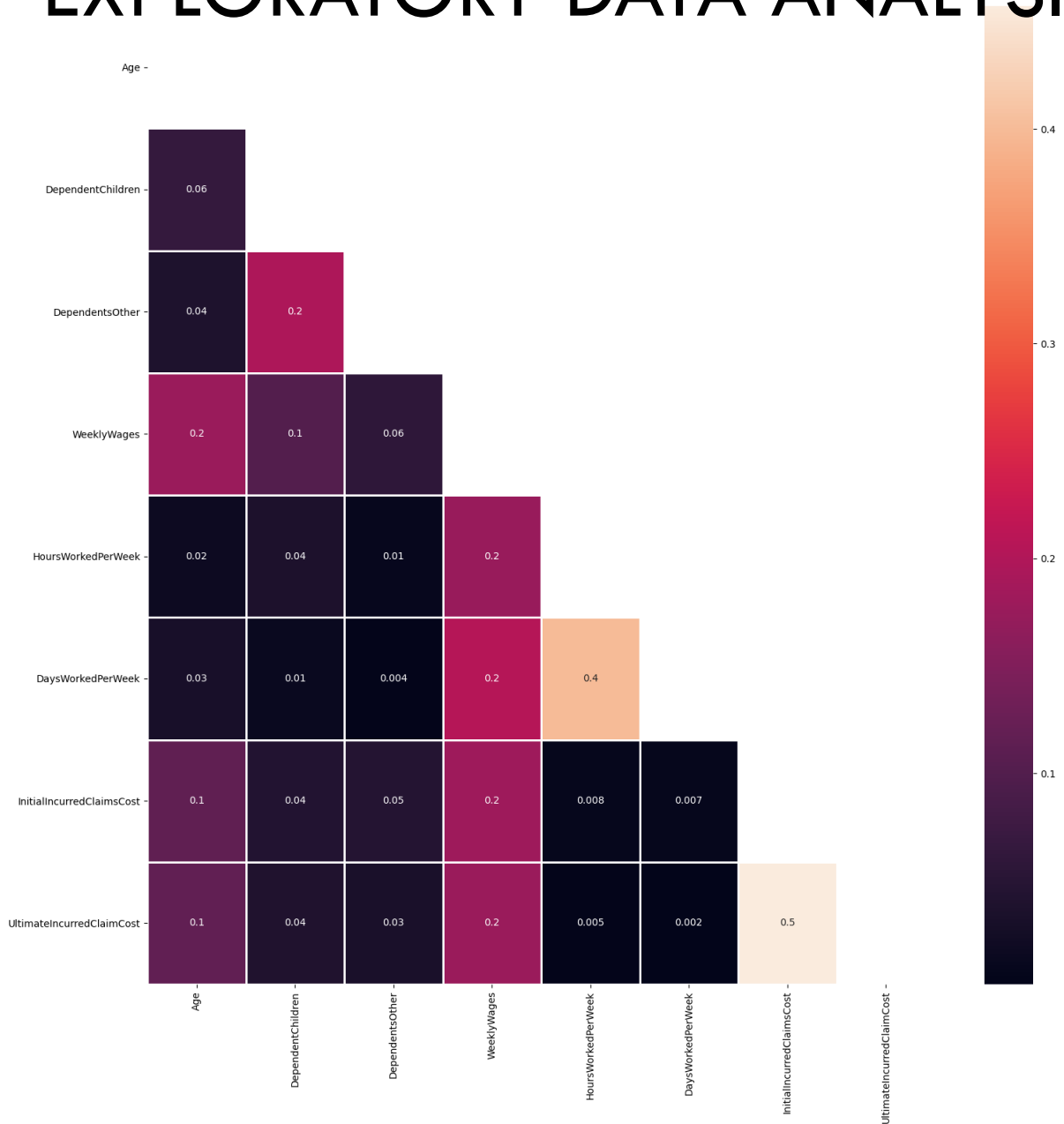


EXPLORATORY DATA ANALYSIS

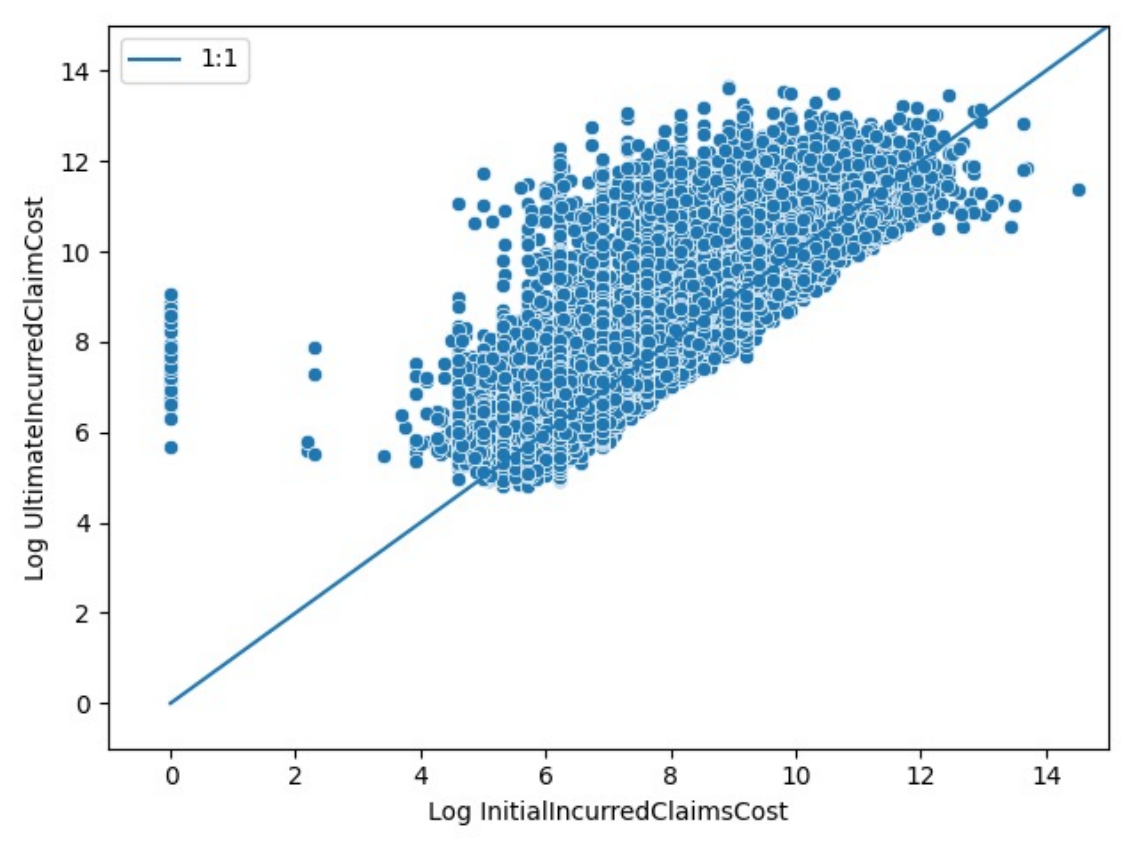
Non-categorical variables

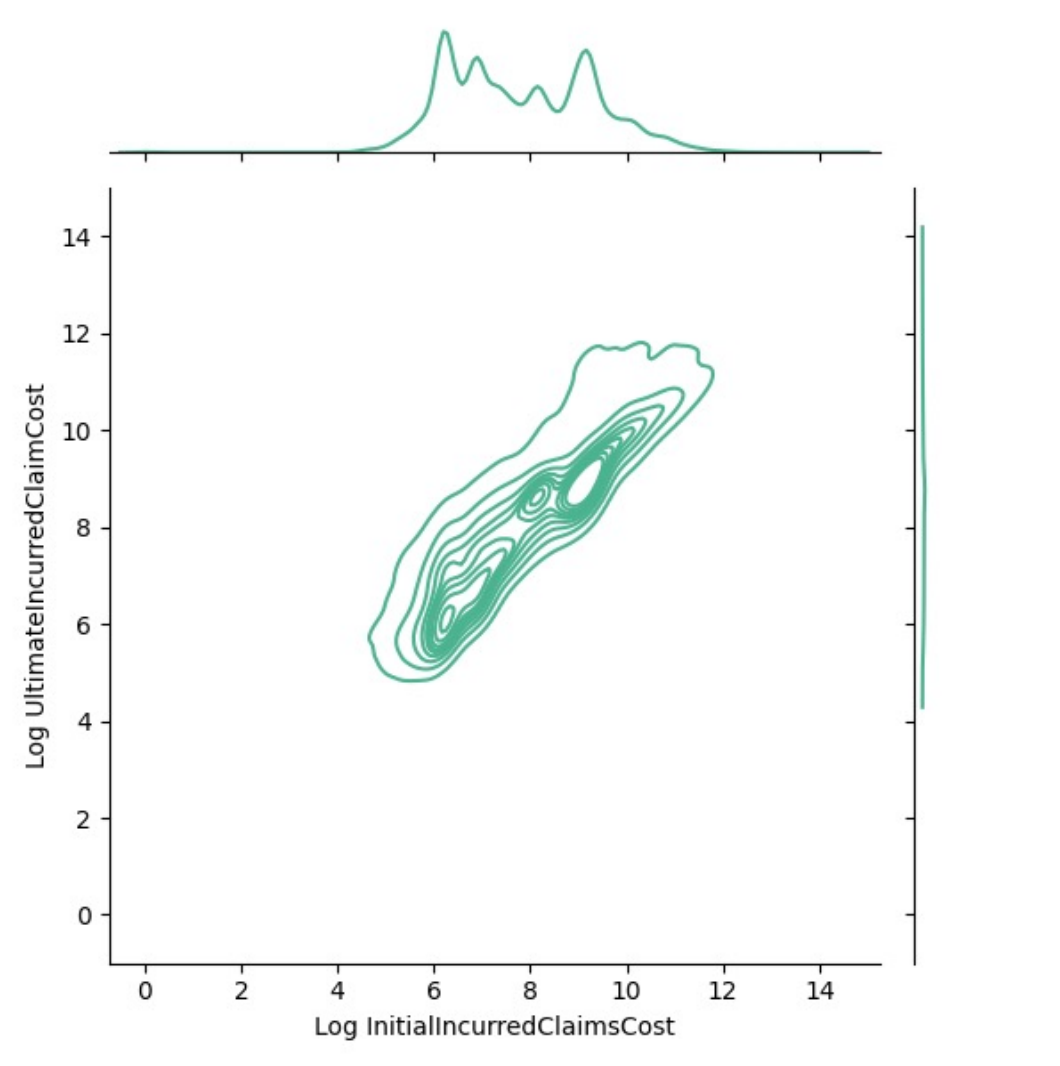


EXPLORATORY DATA ANALYSIS

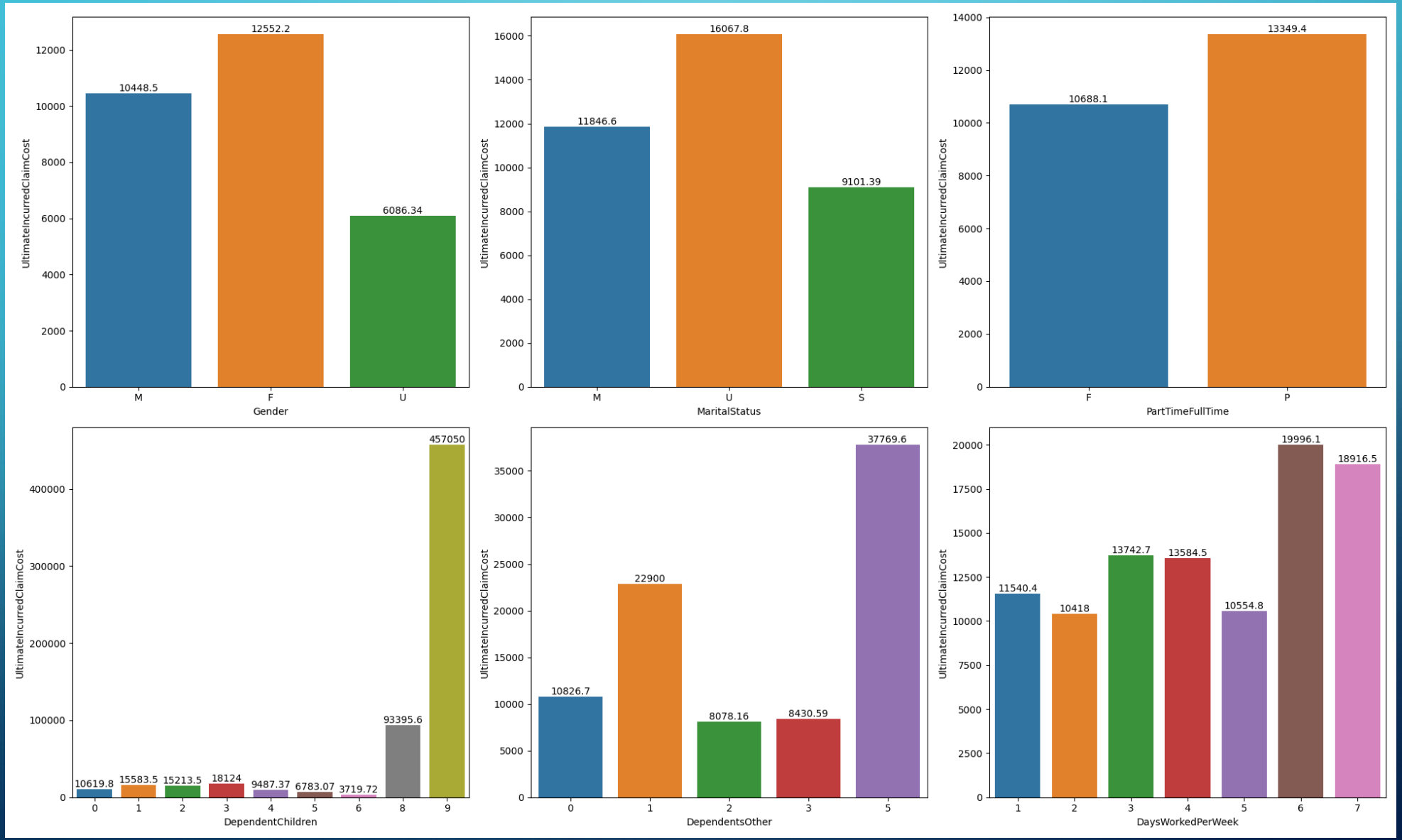


EXPLORATORY DATA ANALYSIS



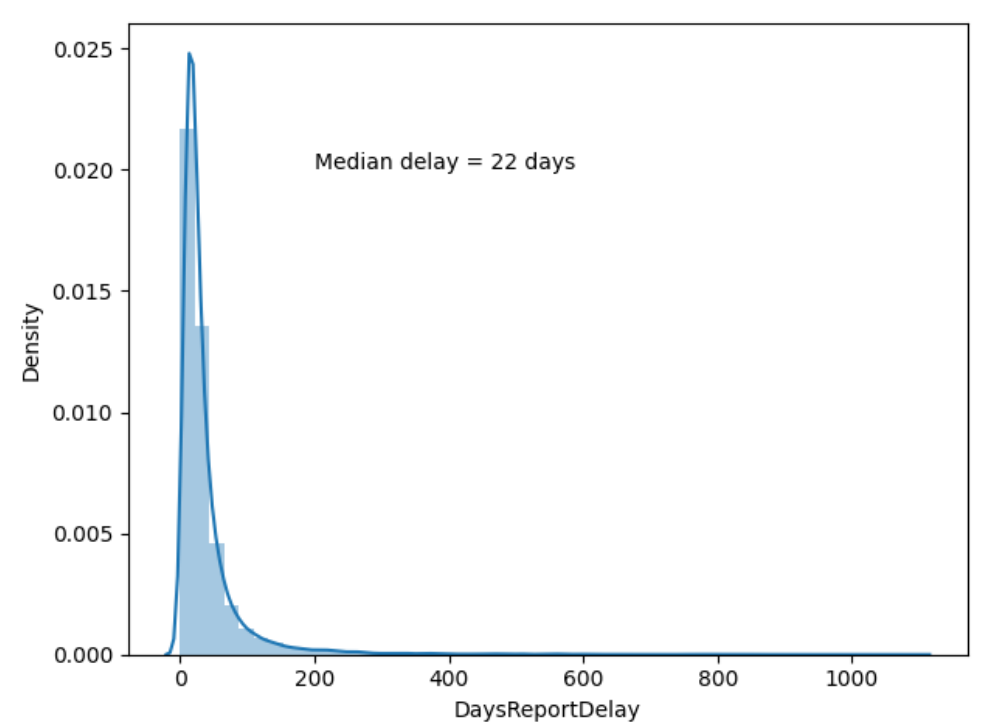


EXPLORATORY DATA ANALYSIS



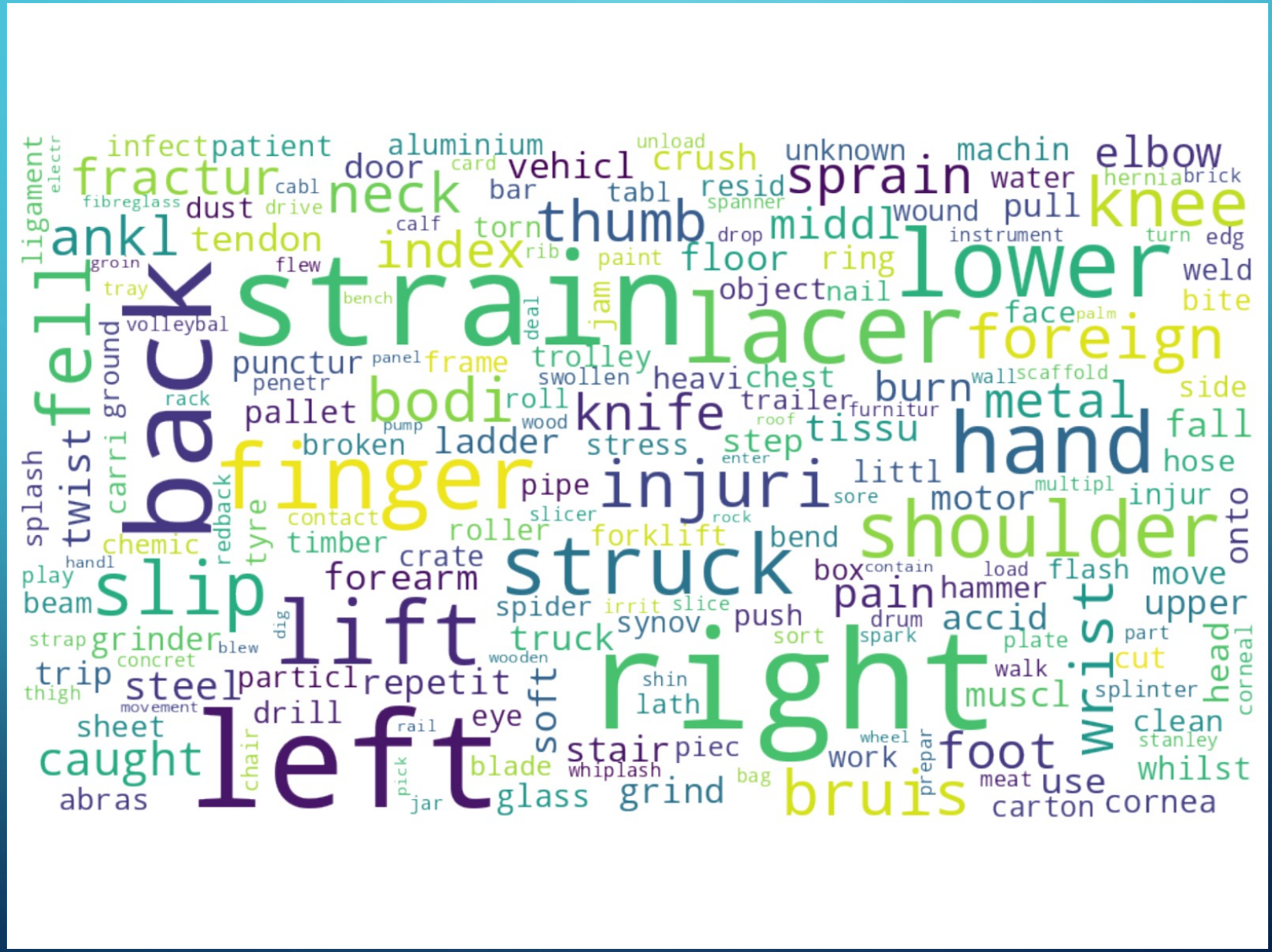
FEATURE ENGINEERING

- Difference between the time of accident and reported time could be important.

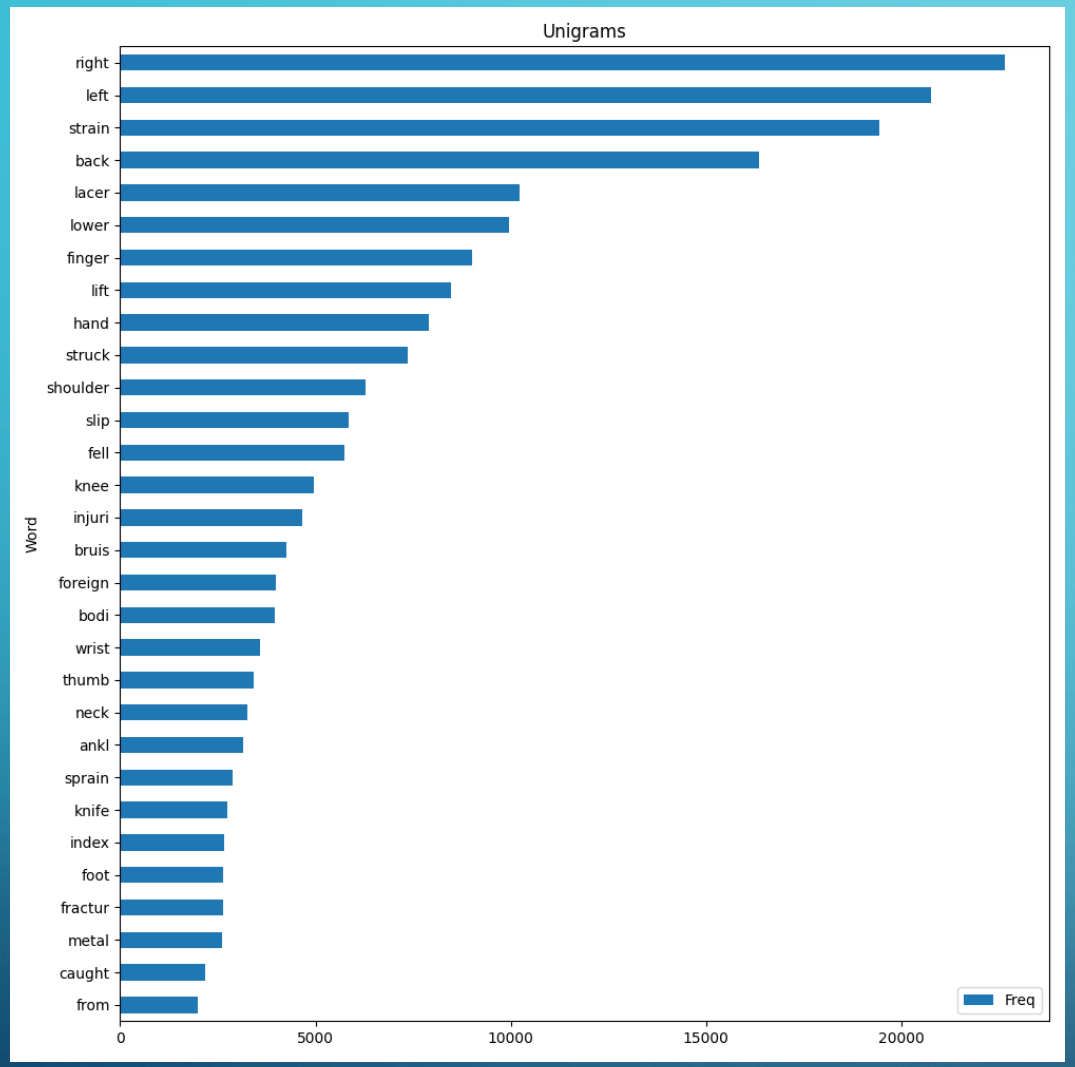


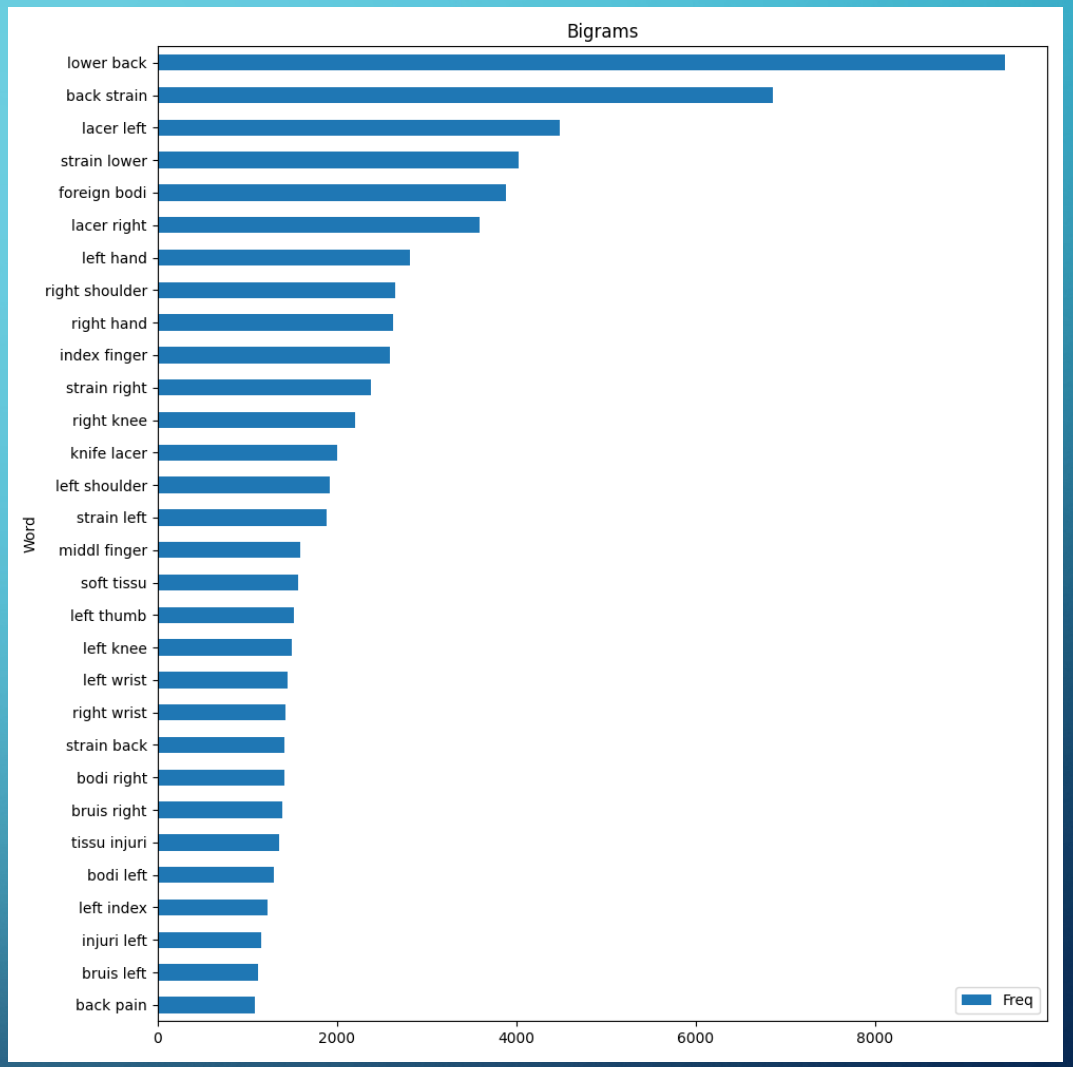
- FEATURE SCALING
- ONE-HOT-ENCODING

NATURAL LANGUAGE PROCESSING (NLP)



NATURAL LANGUAGE PROCESSING (NLP)





NLP model: Bag-of-Words

MODELS:

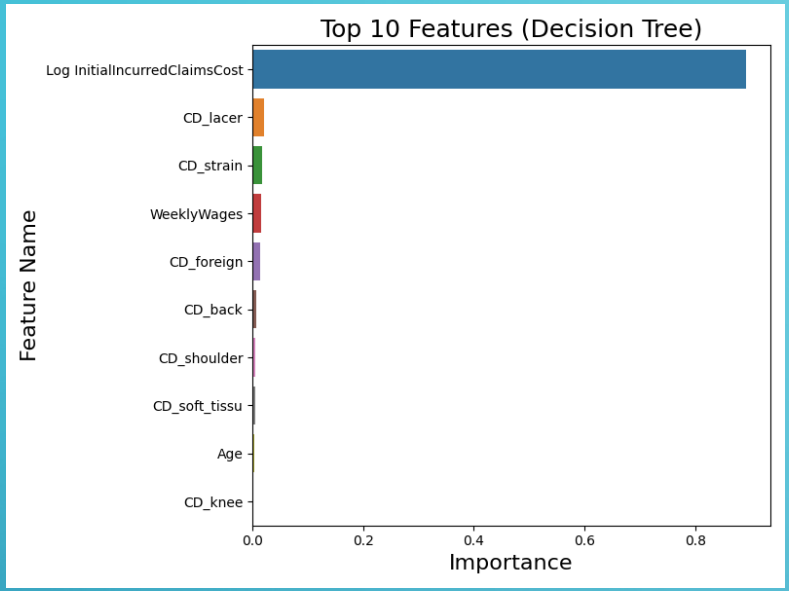
- Decision Tree
- Random Forest
- Gradient Boosting
- Stochastic GB
- AdaBoost
- XGBoost
- LightGBM
- CatBoost"

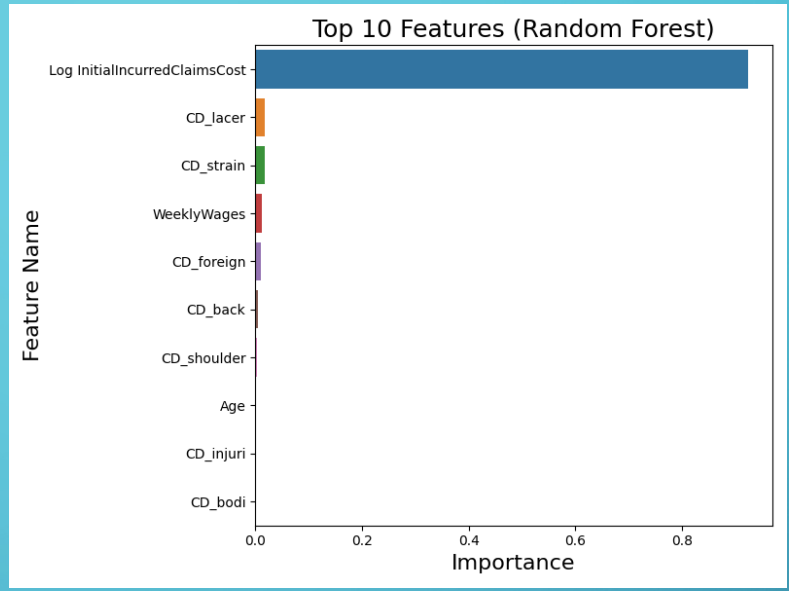
For each model:

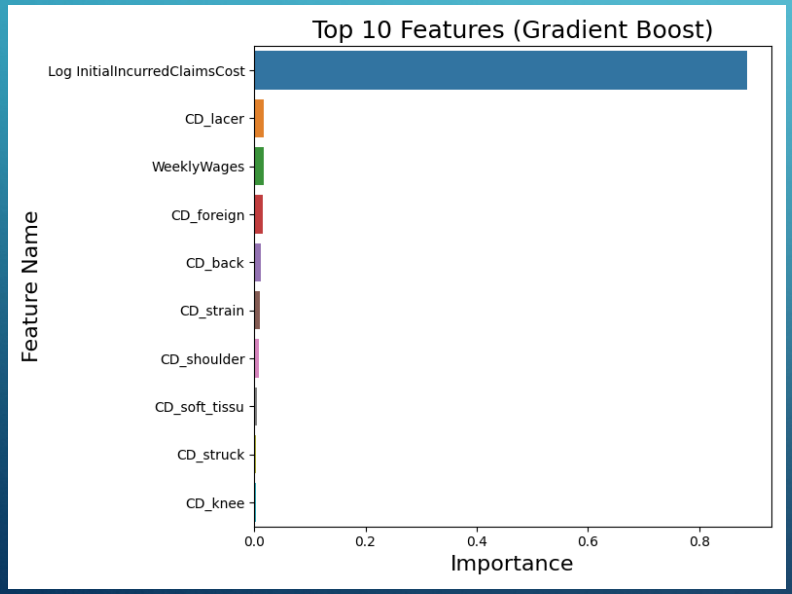
- 5-fold Cross-Validation
- Hyper-parameter Tuning with Grid Search
  - Classical Grid Search
  - Randomized Grid Search

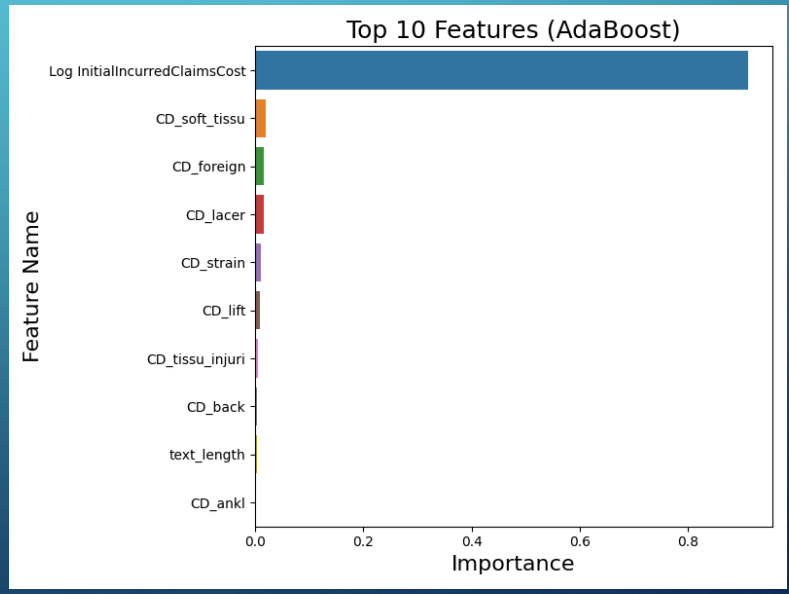
- Metric: RMSE

MOST IMPORTANT FEATURES FOR PREDICTION

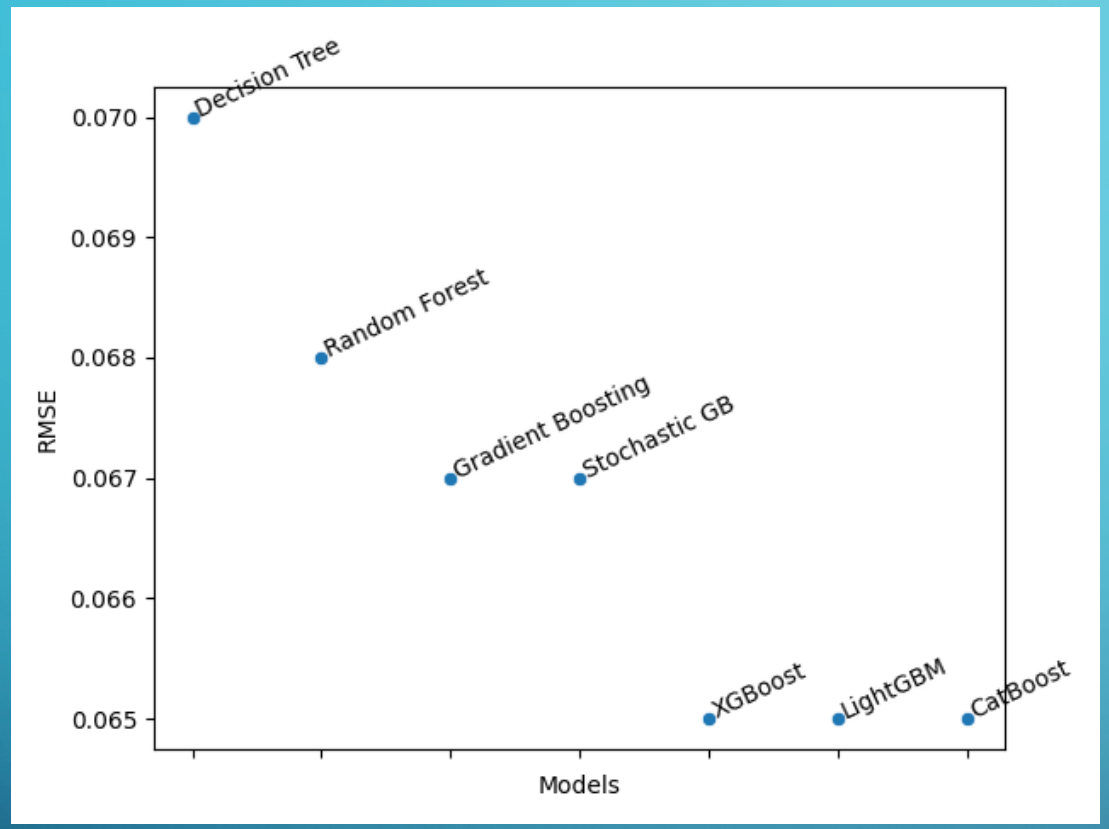








MODEL COMPARISON



CONCLUSION

XGBoost, LightGBM and CatBoost perform as the best predictive models for actuarial Loss prediction.