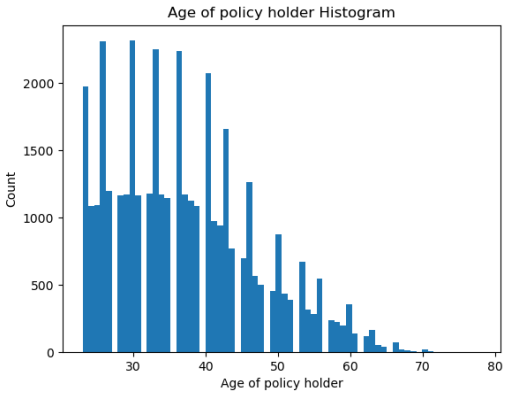
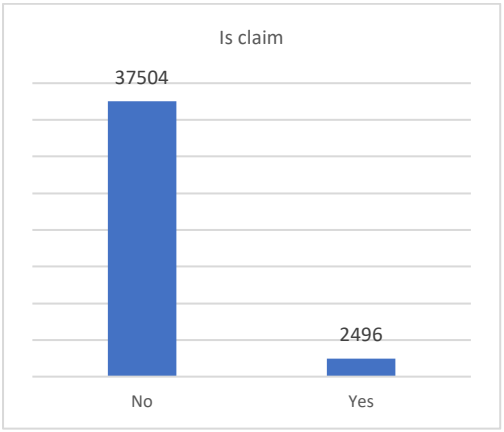
Data Exploratory Analysis and Processing

The dataset used in this study describe car insurance information for each record of car insurance policy. Majority of the columns, a total of 38, are information about the vehicle registered with the insurance policy, e.g. “age_of_car”, “make”, “model”, “gross_weight”, “ncap_rating”, and various car features. Only 3 columns, “age_of_policyholder”, “area_cluster”, “population_density”, are information about policy holder. Determination of policy holder age from this dataset is, therefore, difficult since not much information is known about each person. Determining insurance claim is less affected by the lack of information about policy holder. Additional features are artificially created in an attempt to combat this issue and will be discussed in data processing. Another issue discovered through data exploratory analysis is class imbalance. For regression target variable “age_of_policyholder”, age 47-78 are significantly underrepresented, less than half, in the dataset. For classification target variable, “is_claim”, is even more severely underrepresented with only 6.24% of the dataset being positive class. This class imbalanced is addressed through model selection and weight.

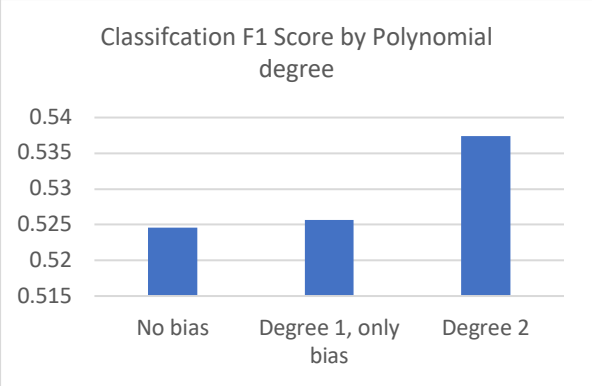


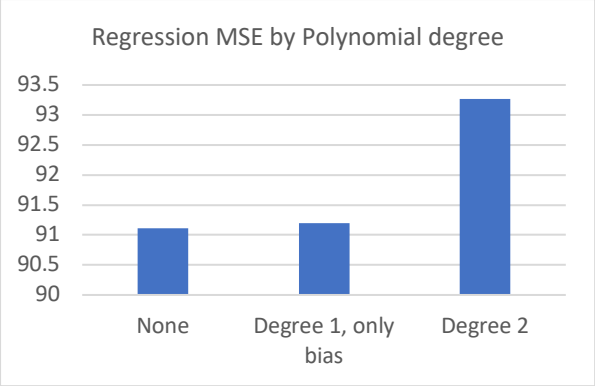


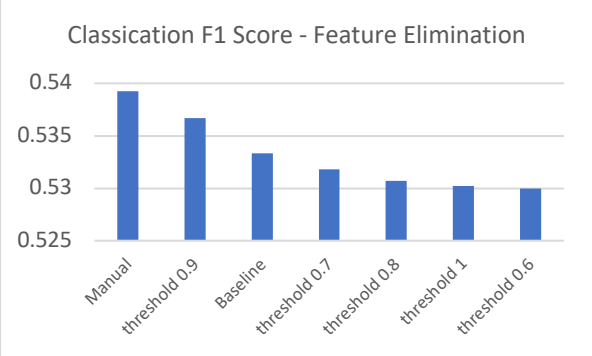
For data processing, input into machine learning models must be numerical. Non-categorical attributed with text value, including age_of_car, max_power, max_torque, are converted to numerical. “is” attributes with “Yes” or “No” value are converted to 1 and 0, respectively. One-hot encoding is used for categorical attributes with more than 2 categories. Since we cannot assume any order of categorical value would form relationship with target variable, simply assigning numerical value to each category would not be logical.

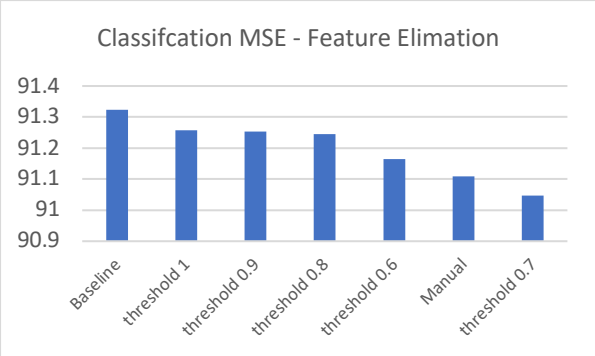
To address the lack of policy holder information, additional features are created generating polynomial and interaction features (sklearn.preprocessing.PolynomialFeatures). For is_claim classification, polynomial degree 2 produced the best result. However, for regression, polynomial feature generation added too much complexity for regression task and result in high MSE. As a result, no feature generation is done for regression task.

Feature elimination is performed to eliminate colinear features. Basic algorithm that iteratively remove colinear variable based on correlation coefficient threshold is devised and compared with manual feature elimination based on logical deduction. Manual elimination produced the best result for classification. Algorithm at 0.7 threshold yield the lowest MSE for regression. Sign of information loss is present when threshold is decreased, this might be due to the algorithm being too rudimentary that it removed feature that hold important information.









Machine Learning Models

Due to class imbalance issue and high number of features, Random Forest Regressor and Random Forest Classifier is chosen for both tasks. This is because random forest can handle a lot of input variables without variable deletion as well as being less sensitive to class imbalanced with bagging and weight balancing. We compare linear regressor and KNN classifier with random forest (regressor and classifier).

Regression models and results

Model	MSE	R2
Linear	91.4145274	0.05595664
Random Forest	91.1083188	0.05911887

Classification models and results

Model	Precision	Recall	Accuracy	F1	Conf matrix
KNN	0.51817269	0.50121213	0.933	0.48856196	[[4663 18] [317 2]]
Random Forest	0.54372054	0.57347986	0.855	0.55062587	[[4195 486] [239 80]]

For age_of_policyholder regression task, random forest regressor performed marginally better than linear regressor in both MSE and R2. For is_claim classification task, random forest achieved better F1 score at 0.55 but lower accuracy than KNN classifier. However, because of class imbalance, accuracy doesn't determine the model performance. Random forest is chosen for both regression and classification.

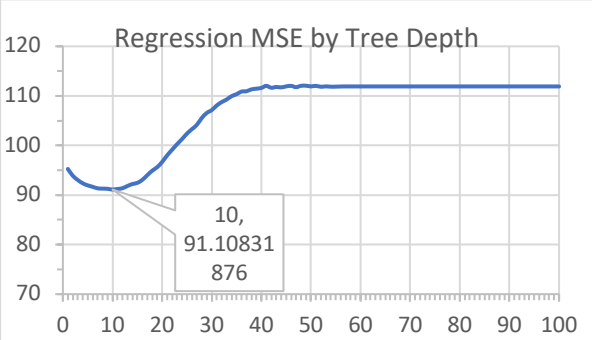
For random forest regressor, split criterion used is squared error to comply with the assessment metric of MSE. For classifier, split criterions are compared and gini performed the best. To combat class imbalance, weighting is employed. For random forest regressor, sample weighting was implemented but achieved poor performance due to overfitting and bias generated from minority class. The built-in class weighting for random forest regressor, however, produce the best F1 score with balanced subsample (balance weights for each bagging sample).

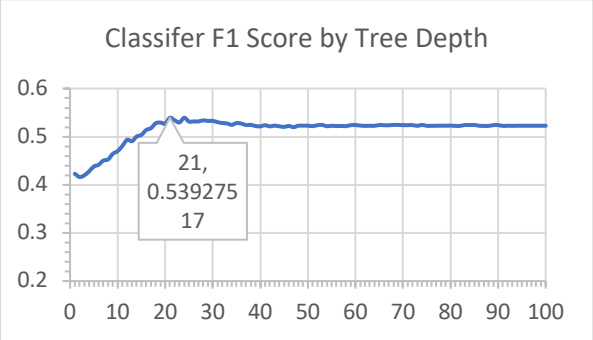
Split Criterion	Precison	Recall	Accuracy	F1
gini	0.53346605	0.54579608	0.8688	0.53738807
entropy	0.52829605	0.54376895	0.8568	0.53168602
logloss	0.52829605	0.54376895	0.8568	0.53168602

Regressor Sample weight	MSE	R2
None	91.1122947	0.05907781
Balanced	150.26623	-0.5518085

Classifier weight	Precison	Recall	Accuracy	F1
No class weight	0.53593348	0.50441892	0.9308	0.49610633
Balanced	0.5254208	0.50295833	0.9308	0.49336454
Balanced subsample	0.55180634	0.50619995	0.9314	0.49912017

The optimum number of features to draw from when splitting, max_features, is determined to be 12 for regressor and 10 for classifier. Optimal maximum tree depth is determined to be 10 for regressor and 21 for classifier. As depth increases, trees overfit. Final random forest regressor model achieved 91.1 MSE and 0.059 R^2. Random forest classifier model achieved 0.55 F1 score and 0.85 accuracy.





Insights for Business Application

In the setting of car insurance business, using the regressor models to predict the age of policy holder can aid targeted marketing strategy. Using the classifier model to predict whether a claim will be filed can aid budgeting and product pricing. The most important aspect is, however, the characteristic of random forest that it is tree based. Decision flow can be extracted to study its underlying logic. For age of policy holder regression, feature importance is extracted from the model, policy_tenure, age_of_car, car model M1, M8, M6 are the most important features. For claim classification, policy_tenure, age_of_car and their interaction features are the most important feature in determining whether a claim will be filed.