INSURANCE CLAIM PREDICTION

MANJU PENUMARTHI

Problem Statement

- The Vehicle Insurance company is trying to save money on material costs like promotional flyers, Phone call charges, and instead wants to accurately predict the number of claims and cost of the claims for future customers.

- The company wants to predict the number of claims and cost of the claims for future customers, so that they can direct their marketing to customers with fewer claims.

Objective

Segment Customer’s into multiple groups based on their claim history in order to:

- Save money on material costs
- Start targeted marketing campaigns

Glossary for the dataset

Name	Description
pol_number	policy number for the insurance policy
pol_eff_dt	auto insurance policy effective date
gender	gender of driver: F, M
agecat	driver's age category: 1 (youngest), 2, 3, 4, 5, 6
date_of_birth	driver's date of birth
credit_score	driver’s credit score
area	driver's area of residence: A, B, C, D, E, F
traffic_index	traffic index of driver’s area of residence
veh_age	age of vehicle(categorical): 1 (youngest), 2, 3, 4

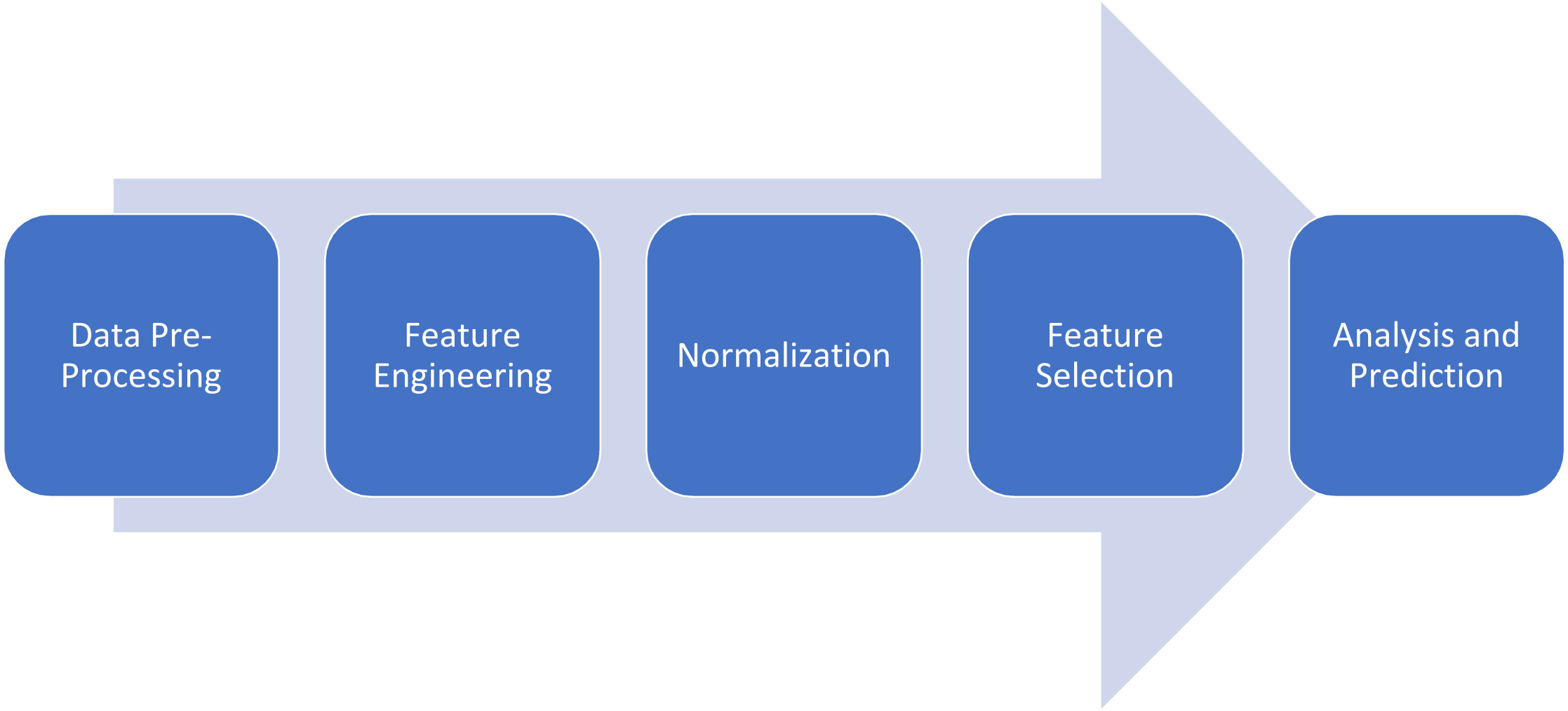
Glossary for the dataset

Name	Description
veh_body	vehicle body type
veh_value	vehicle value, in $10,000s
months_insured	number of months vehicle insurance is bought(integer)
claim_office	office location of claim handling agent: A, B, C, D
numclaims	number of claims(integer): 0 if no claim
claimcst0	claim amount: 0 if no claim
annual_premium	total charged premium i.e. the cost of insurance

TOOLS

1. Microsoft Excel
   - Initial evaluation of data
2. IDE
   - Jupyter Notebook
3. Python
   - Data Preprocessing
   - Feature Engineering
   - Predictive Modelling
   - Supervised Learning
   - Visualization

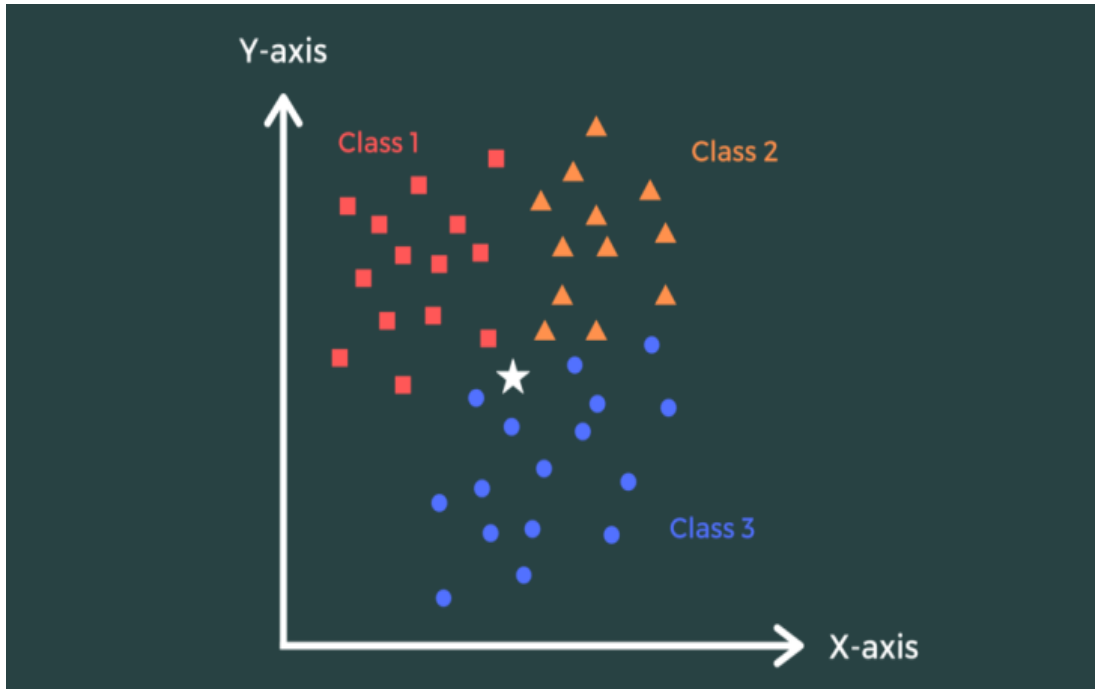
Methodology Behind the Recommendations



Algorithm

KNN Algorithm:

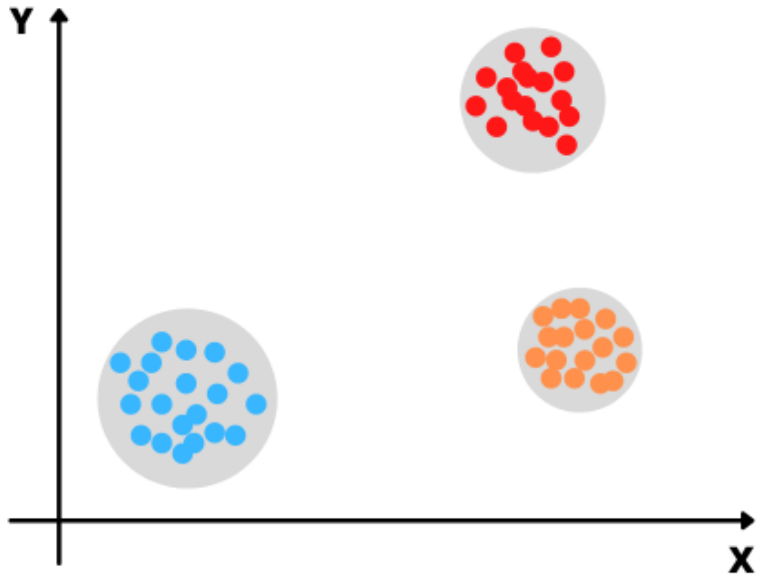
- The KNN algorithm uses 'feature similarity' to predict the values of any new data points. This means that the new point is assigned a value based on how closely it resembles the points in the training set.



KNeighborsClassifier

To Predict Number of Claims per Customer:

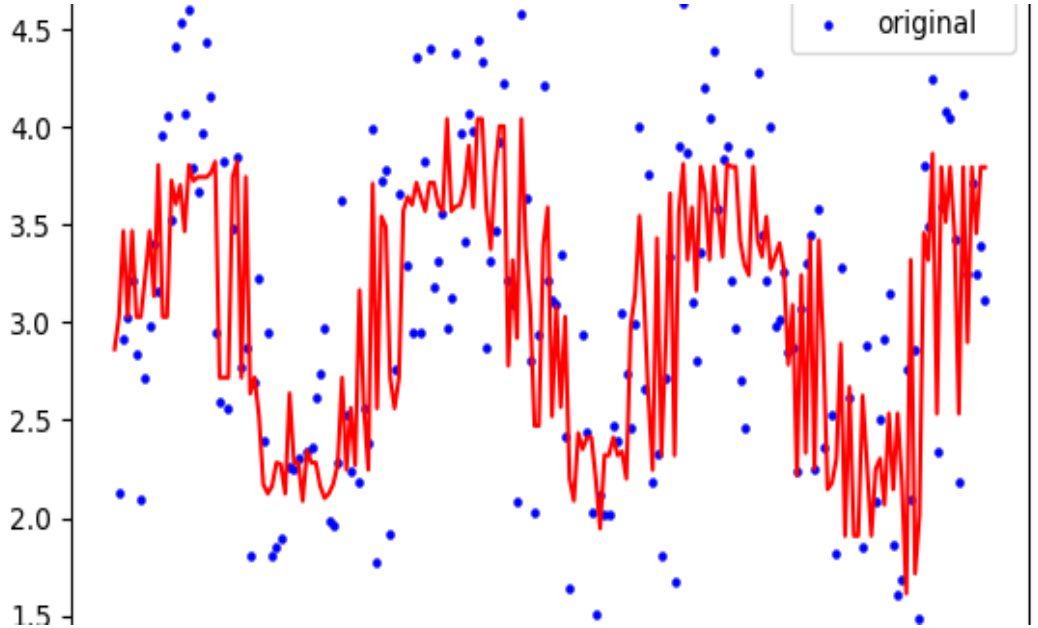
- KNeighborsClassifier implements classification based on voting by nearest k-neighbors of target point.
- Using KNeighborsClassifier and then the argument inside determines how many nearest neighbors you want your datapoint to look at. There is no rule of thumb for how many neighbors you should look at.



KNeighborsRegressor

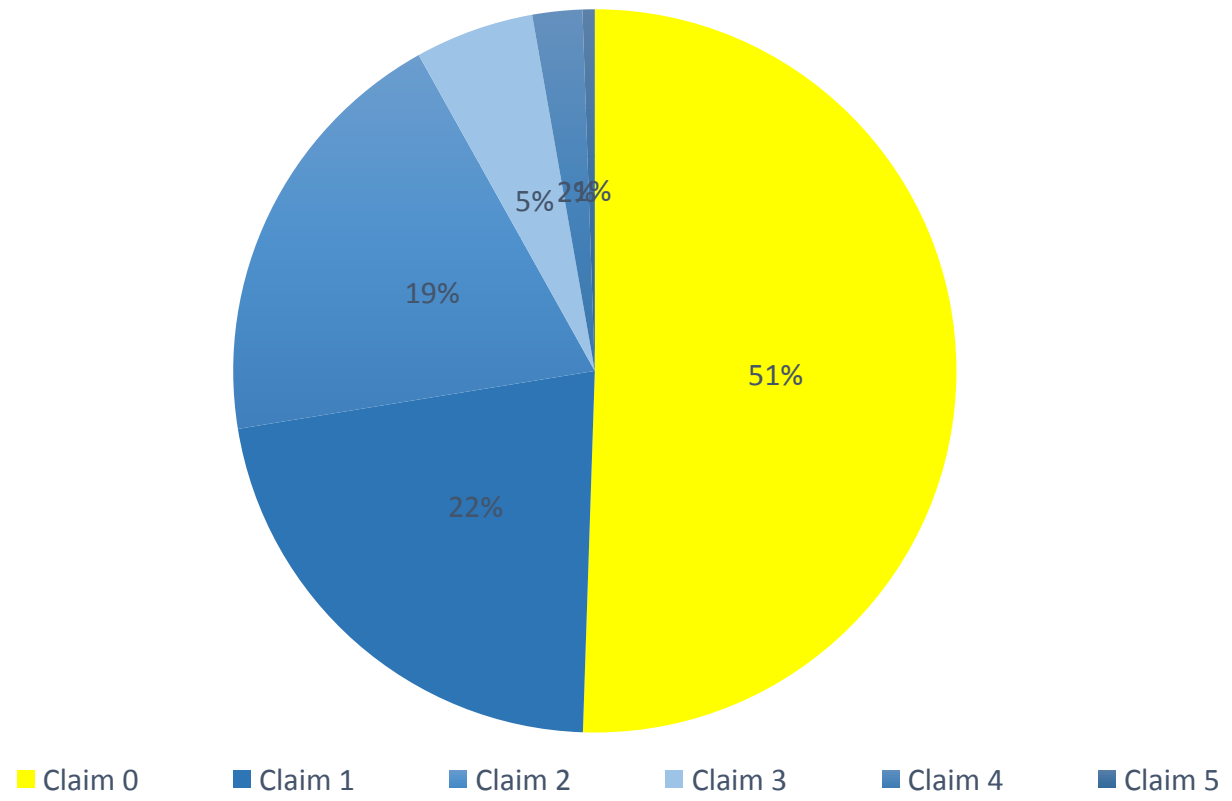
To Predict Cost of Claims per Customer:

- In a regression task, which predicts continuous values (not labels), KNN takes the mean of the nearest *k* neighbors.

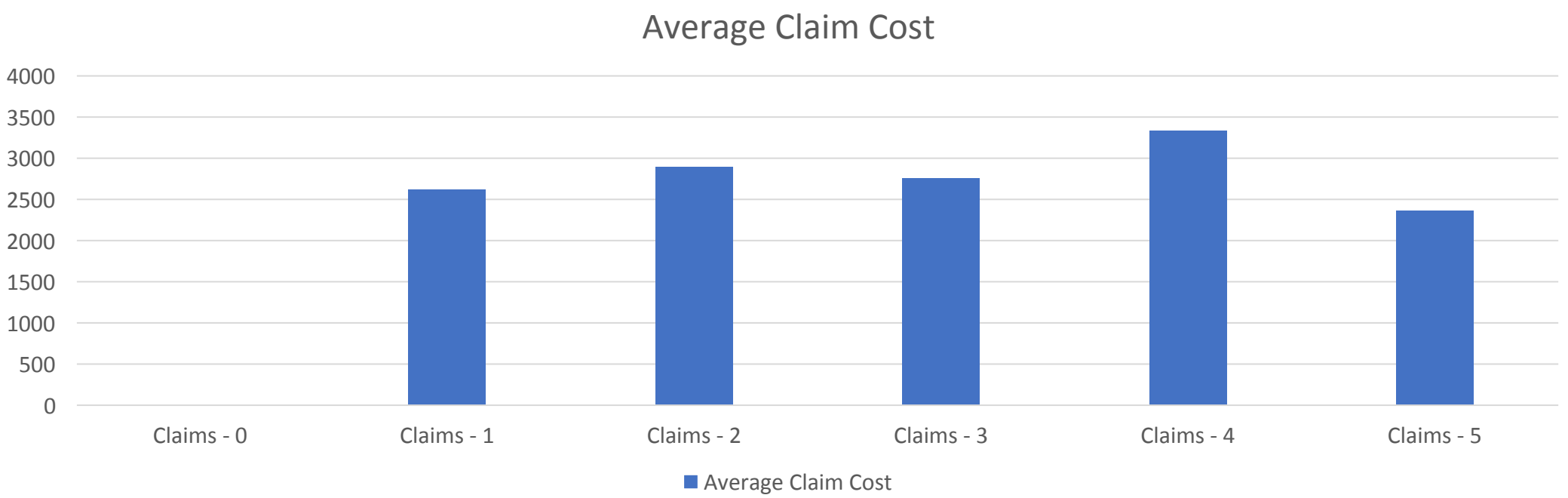


RESULTS

Numclaims Category Percentage



Number of Claims Category vs Average Claim Cost Prediction



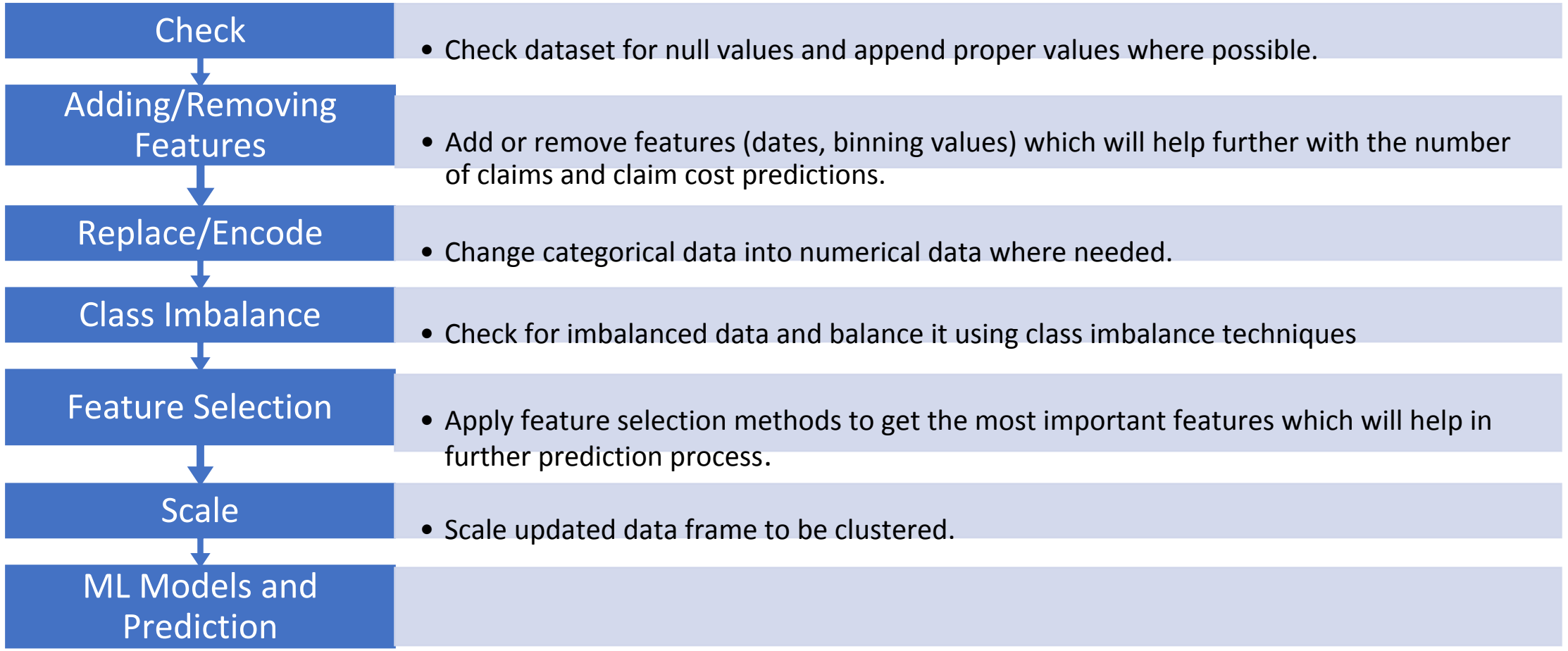
Recommendations

Number of Claims	Average Claim Cost	Total No. of Customers	Send Mails and Give a Call
0	$0.00	3888	Yes
1	$2619.57	1599	Yes
2	$2889.14	1378	May or May not
3	$2758.47	377	May or May not
4	$3334.10	173	No
5	$2363.69	49	No

Questions?

INSURANCE CLAIM PREDICTION
Detailed Approach

Steps


Check
Check dataset for null values and append proper values where possible.
Adding/Removing Features
Add or remove features (dates, binning values) which will help further with the number of claims and claim cost predictions.
Replace/Encode
Change categorical data into numerical data where needed.
Class Imbalance
Check for imbalanced data and balance it using class imbalance techniques
Feature Selection
Apply feature selection methods to get the most important features which will help in further prediction process.
Scale
Scale updated data frame to be clustered.
ML Models and Prediction

Feature Engineering and Supervised Learning

1. Feature Engineering:

- Feature engineering is the process of selecting, manipulating, and transforming raw data into features that can be used in supervised learning. In order to make machine learning work well on new tasks, it might be necessary to design and train better features.

2. Supervised Learning:

- Supervised learning, as the name indicates, has the presence of a supervisor as a teacher. Basically supervised learning is when we teach or train the machine using data that is well labeled. Which means some data is already tagged with the correct answer. After that, the machine is provided with a new set of examples(data) so that the supervised learning algorithm analyses the training data(set of training examples) and produces a correct outcome from labeled data.

Data Preprocessing/Data Cleaning

1. **Initial Inspection/Check:**

➢ Checked the Null value count in each column

➢ Missing Data Imputation: Filled the null values using Median and Mean methods.

- Filled the null values in the “agecat” column using Median() method.
- Filled the null values in the “credit_score” column using Mean() method.
- Filled the null values in the “traffic_index” column using Mean() method.

Data Preprocessing/Data Cleaning

2. Adding or Removing Features:

- Added a column named “person_age” using the “date_of_birth” column values. The “person_age” column is the age of the policy holder at the time of policy purchase.

- Removed few columns from the data frame because they did not contribute to the driver’s risk:

- Dropped "claim_office" column as it consists of more than 80 percent of null values and has no effect on customer claim prediction.
- Dropped "pol_eff_dt" column as it don't have any effect on customer claims
- The annual premium column was removed from the data frame because the it was the same for all customers.

- Binning Values:

- Created bins for “credit_score” and “traffic_index” columns to reduce the effects of minor observation errors.

Data Preprocessing/Data Cleaning

3. Replace/Encode: Using numeric values, helps to more easily determine a probability for our values.

➢ **Replacing Strings with Numerics:**

- Replaced the characters in “area” column [A, B, C, D, E, F] with Numbers [1, 2, 3, 4, 5, 6] respectively.

➢ **One-Hot Encoding:**

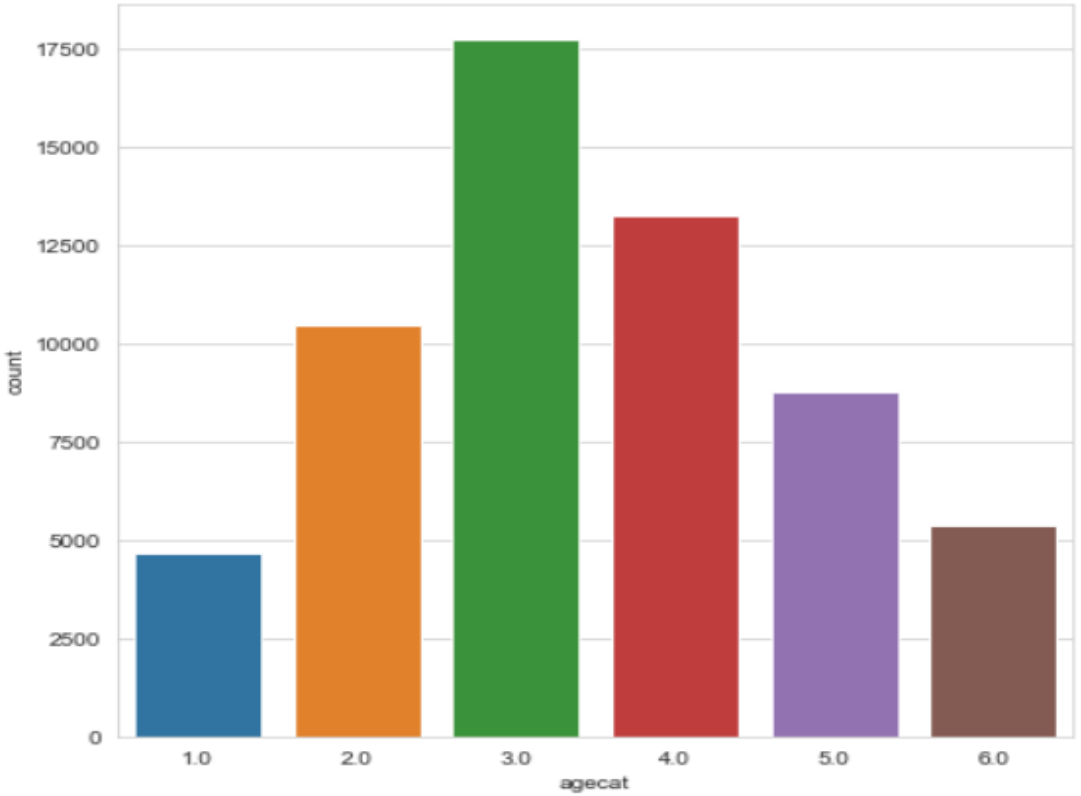
- Applied one-hot encoding to “gender” and “veh_body” columns to convert them to numerics (0,1).

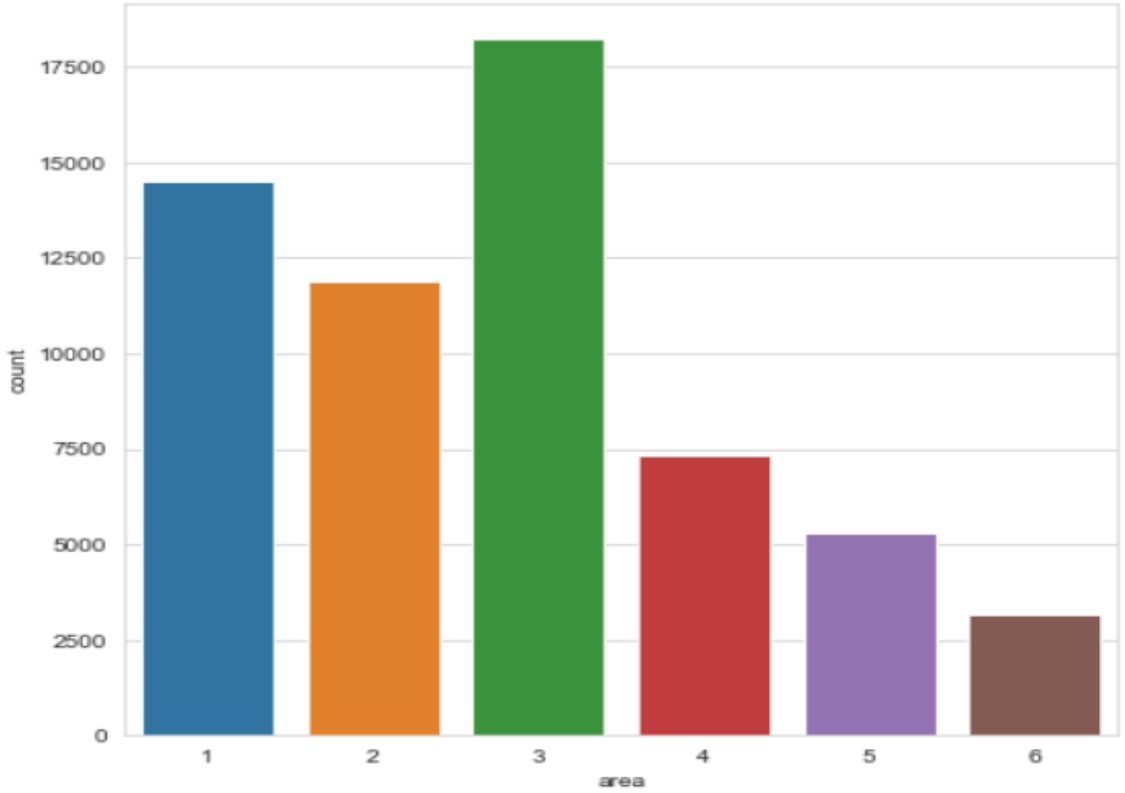
Data Preprocessing/Data Cleaning

4. Handling Class Imbalance:

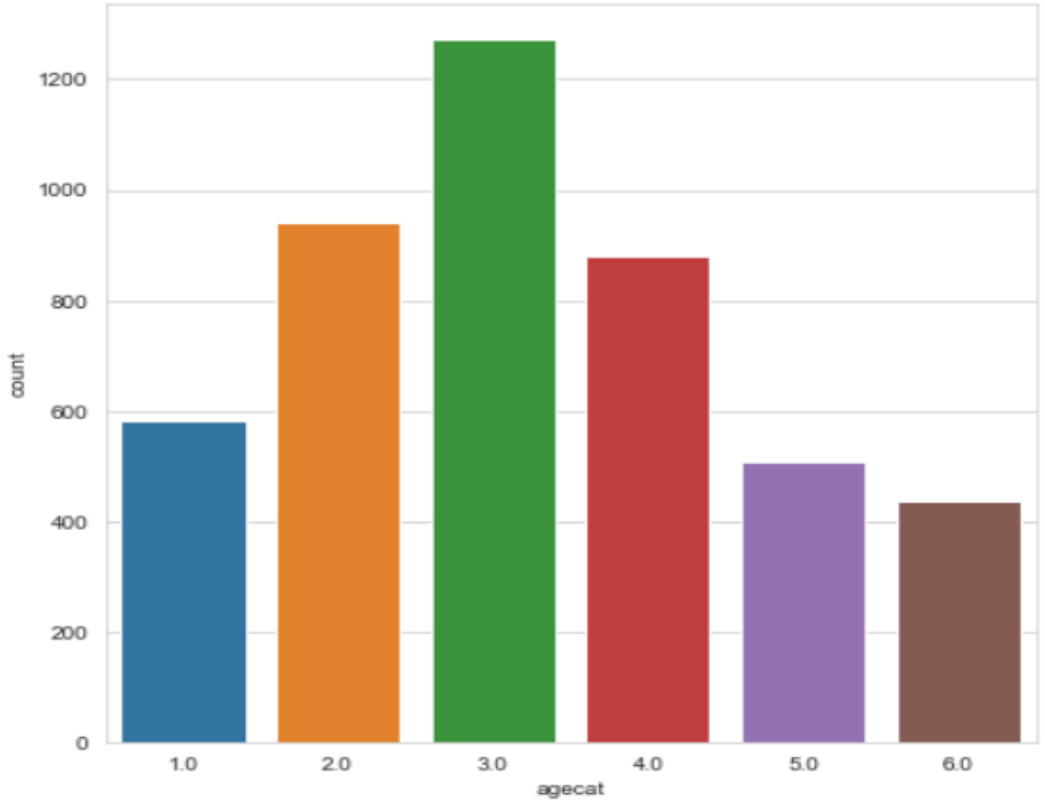
- It happens when the data categories are not equally represented.
- To handle class imbalance we have multiple imbalance techniques such as Undersampling, Oversampling and SMOTE.
- Here in our case, the numclaims column as an unbalanced data like below:
- Numclaim value: Count = 0:37779, 1:6021, 2:1004, 3:343, 4: 106, 5:41
- I used undersampling technique to reduce the 0 numbered claims to match with the remaining minority class values.
- The plots in next slides depicts the importance of balancing data.

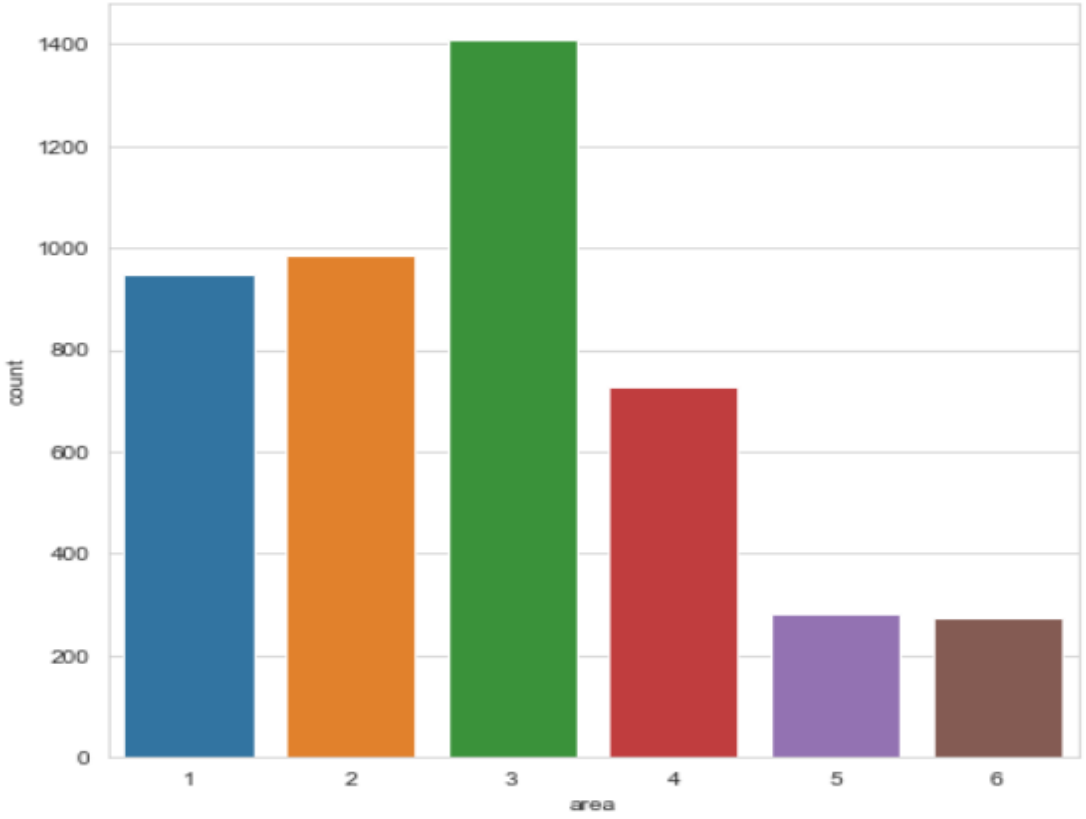
Plots For “agecat” and “area” Features Before Sampling



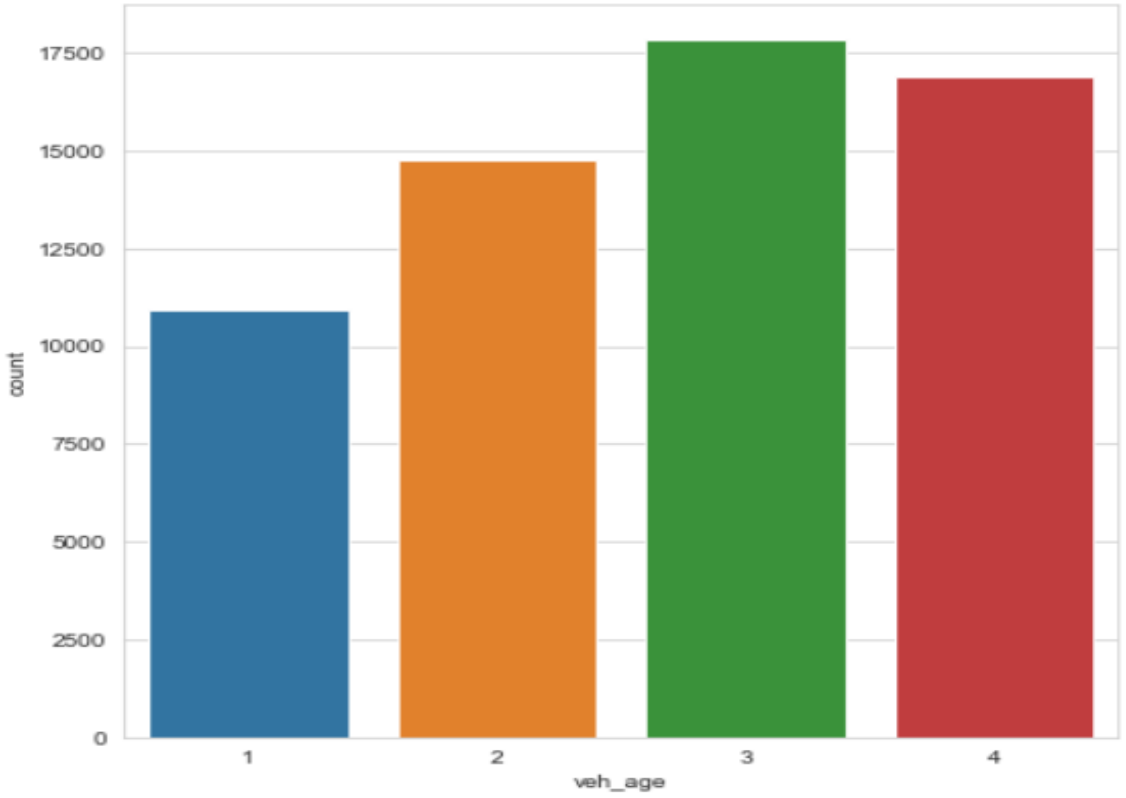


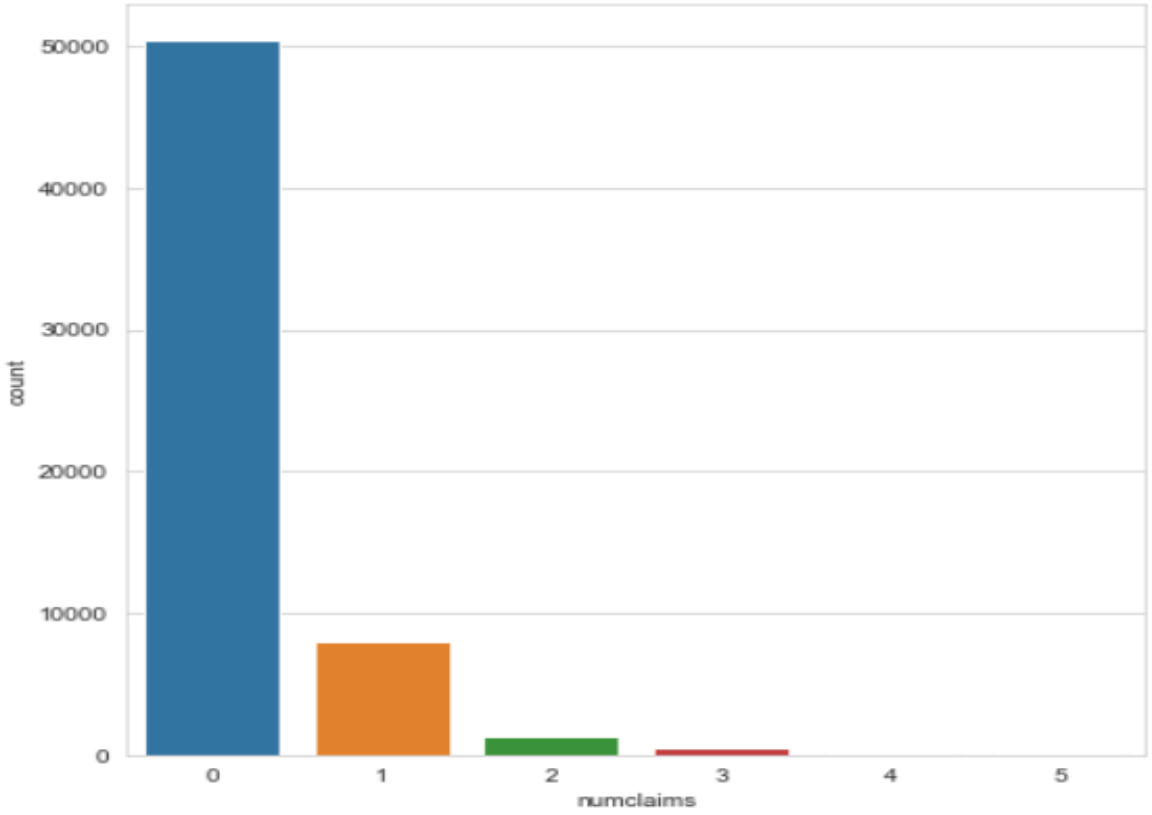
Plots For “agecat” and “area” Features After Sampling



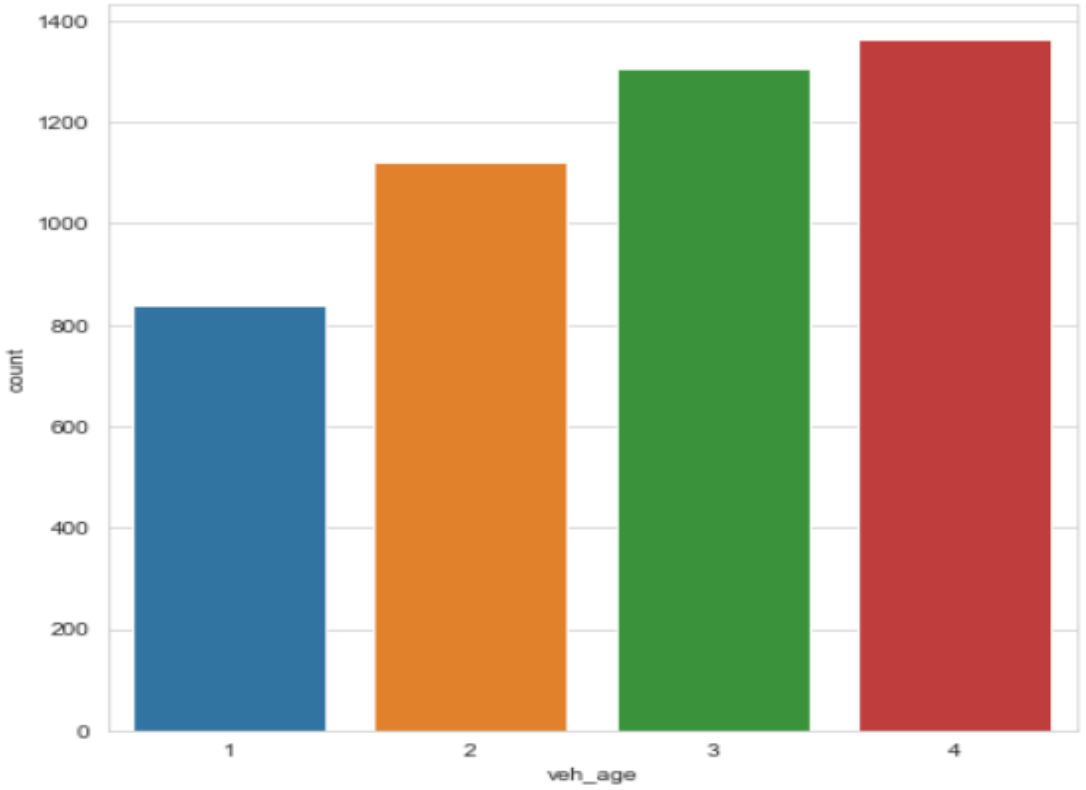


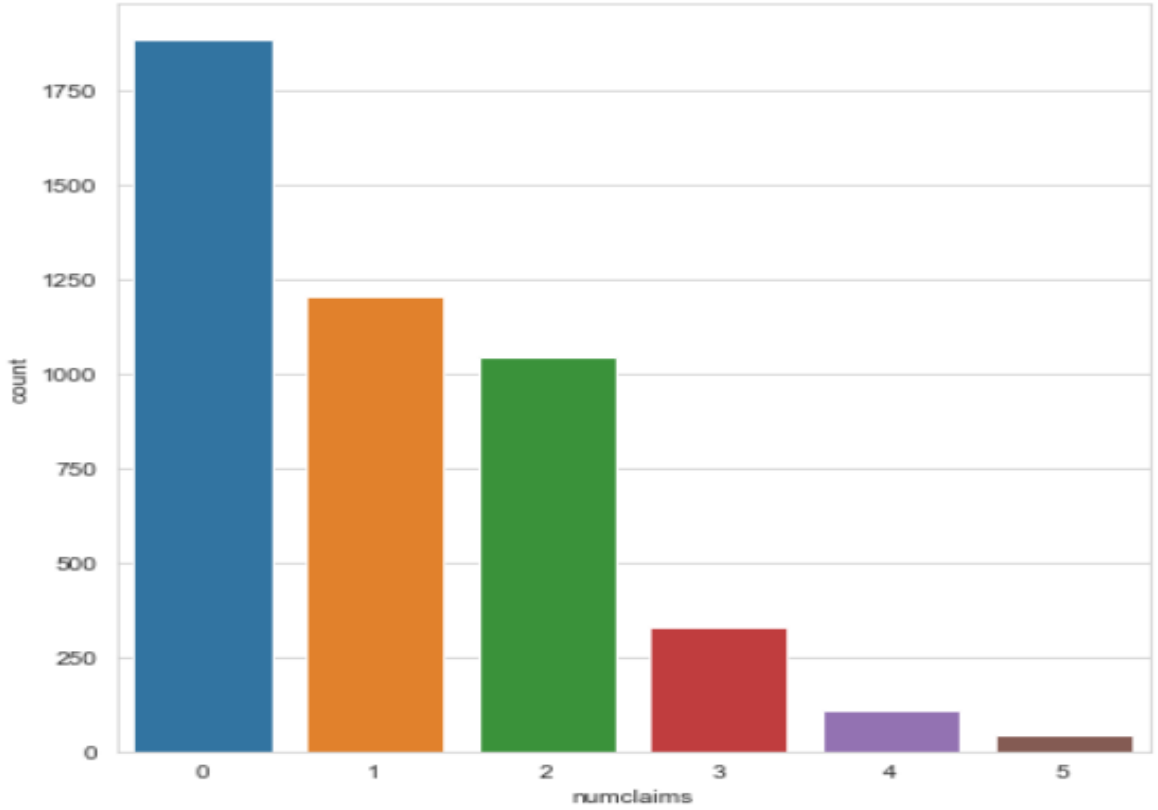
Plots For “veh_age” and “numclaims” Features Before Sampling





Plots For “veh_age” and “numclaims” Features After Sampling





Data Preprocessing/Data Cleaning

5. Feature Selection:

- Feature Selection is the method of reducing the input variable to your model by using only relevant data and getting rid of noise in data. It is the process of automatically choosing relevant features for your machine learning model based on the type of problem you are trying to solve.
- It is an important problem in machine learning, where we will be having several features in line and have to select the best features to build the model.
- We have different feature selection methods namely Filter methods, Wrapper methods, Principal Component Analysis, etc.
- In order to choose the relevant features for prediction of number of claims and claim cost in our dataset, I used Chi-squared test.
- The chi-square test helps you to solve the problem in feature selection by testing the relationship between the features.

Data Preprocessing/Data Cleaning

6. Scaling data:

- Feature Scaling is a method used to normalize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data preprocessing step.

- In simpler words, we can say that the scaling is used for making data points generalized so that the distance between them will be lower.

Data Preprocessing/Data Cleaning

6. Appling Machine Learning Models:

(a)

- In order to predict the numclaims column values for each potential customer, at first I started finding the performance metric value (Precision score or f1 score) on the balanced train dataset with numcliams as a target variable.

- As numclaims column consists of five categories 0, 1, 2, 3, 4, 5, I used **KNN Classifier** to predict the performance. It resulted in nearly 70% of both precision and f1 score.

- It indicates that the model performed well and we can reply on the prediction of number of claims of the future customers for targeted marketing in order to reduce the material costs and improve the profits.

Data Preprocessing/Data Cleaning

6. Appling Machine Learning Models:

(b)

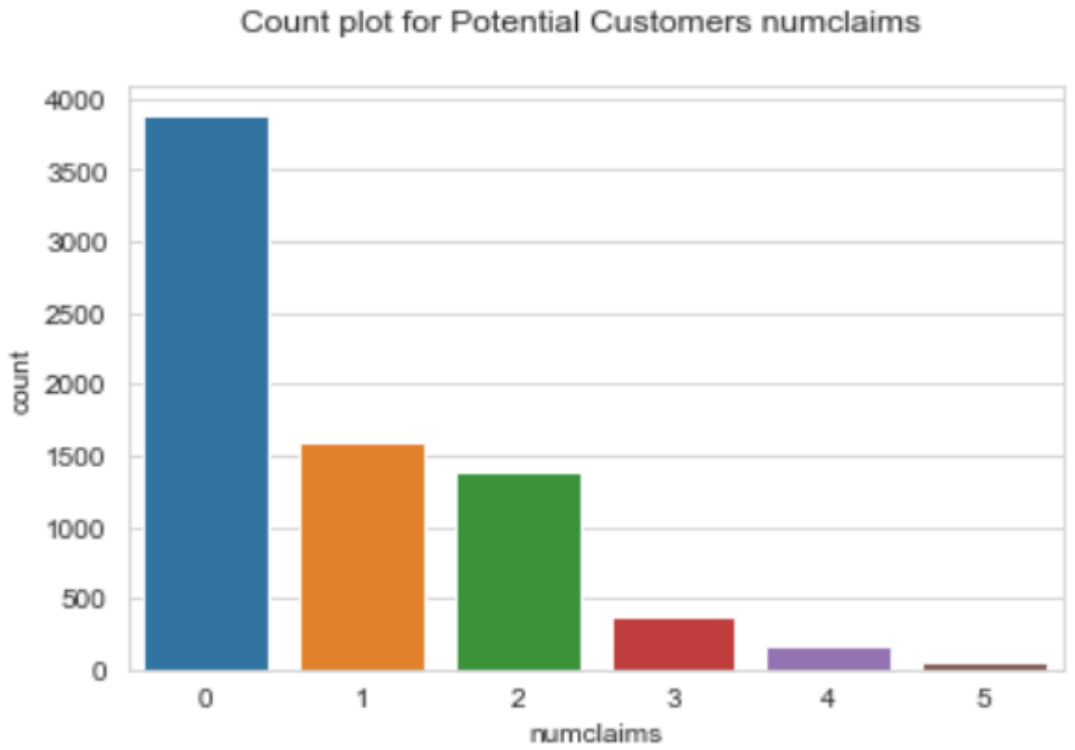
- In order to predict the claimcost column values (continuous variable) for each potential customer, at first I started finding the error metric value (Mean Absolute Error) on the balanced train dataset with claimcost as a target variable.

- As claimcost column are continuous variables I used **KNN Regressor** to predict the performance. It resulted in 2494.045 MAE.

- It indicates that the model performed well and we can reply on the prediction of claimcost of the future customers for targeted marketing in order to reduce the material costs and improve the profits.

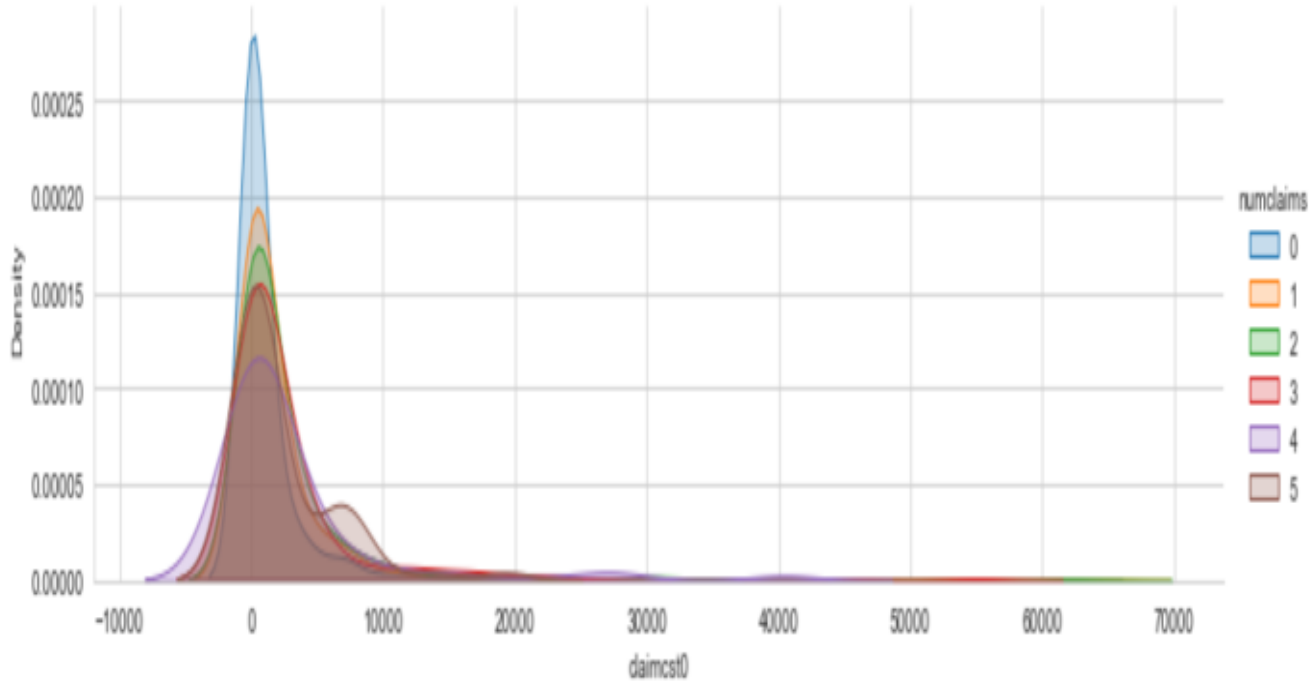
KNN Classifier Algorithm and EDA

- Using Chi-squared test with k = 20 for Feature selection and KNN Classifier algorithm with k = 7 to predict the target variable, I predicted the numclaims values for the future customers.



KNN Regressor Algorithm and EDA

- Using Chi-squared test with k = 20 for Feature selection and KNN Regressor algorithm with k = 7 to predict the target variable, I predicted the claimcst values for the future customers.



KNN Prediction Summary

- I interpreted the results from KNN using the best K. Appended the predictions to the original test dataset provided and performed summary statistics/visualizations on it.

- Out of 7464 total new customers:

- 2403 customers are predicted to have no claims and zero claimcost – 32%.
- 1042 customers are predicted to have 1 claim and non-zero claimcost – 14%.
- 926 customers are predicted to have 2 claims and non-zero claimcost – 12%.
- 253 customers are predicted to have 3 claims and non-zero claimcost – 3.4%.
- 106 customers are predicted to have 4 claims and non-zero claimcost – 1.5%.
- 26 customers are predicted to have 4 claims and non-zero claimcost – 0.4%.

Conclusion

- KNN gave predictions that were more precise for potential customer segmentation in determining their number of claims and claim cost for targeted marketing campaigns.

- There are many different prediction algorithms that could be used to find the potential customers that are insured, but from the knowledge and tools available KNN was found to be the best algorithm for this application.

Questions?