ITCS/DSBA 6100
Final Project Presentation
Team 4

Predicting Loss Ratios for Auto Insurance Portfolios

Aditya Kamble

Urma Haldar

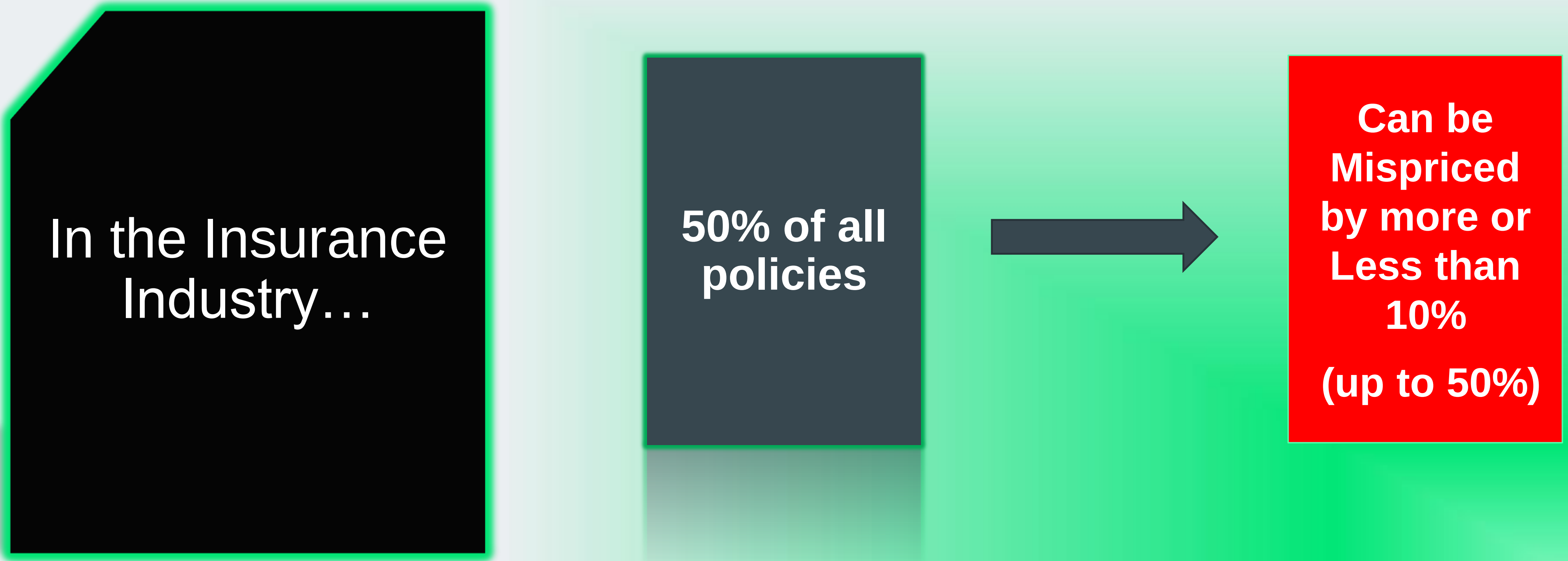
Niraj Bista

Mike Teague

Agenda

- Introduction
- Domain Knowledge
- EDA
- Preprocessing
- Modeling
- Evaluation
- Conclusions

Problem



In the Insurance Industry…

50% of all policies

Can be Mispriced by more or Less than 10%

(up to 50%)

Introduction

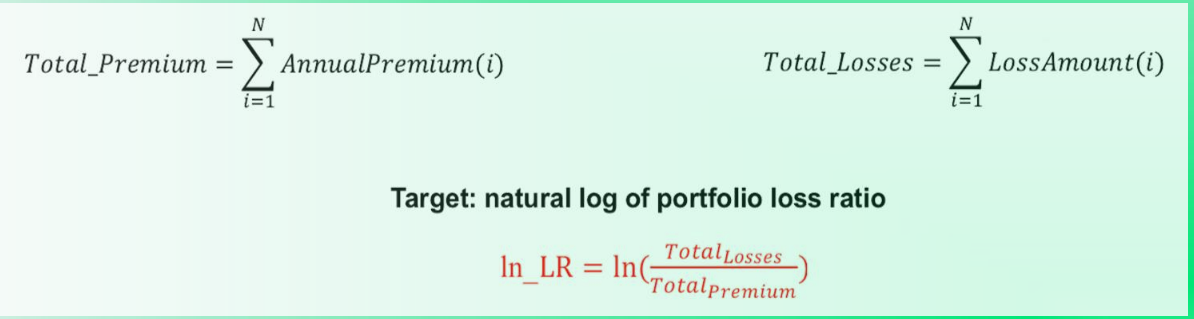
Datasets

- Historical dataset of around 400K input samples of Auto Insurance policies
  - 64 input features (ex. vehicle make year, vehicle performance, usage, miles to work etc.)
  - Most of these input features were categorical.
- Testing dataset of 330 policy portfolios
  - Each consisting of at least 1000 policies
  - Included almost all of features above EXCEPT those such as *loss_amount*.

Introduction

Our Objective

- Predict Missing Loss Amounts in Test Data
- Find *Natural Logarithm (ln_LR)* of the *loss ratio* for each portfolio in testing dataset



$$Total\_Premium = \sum_{i=1}^{N} AnnualPremium(i) \qquad Total\_Losses = \sum_{i=1}^{N} LossAmount(i)$$

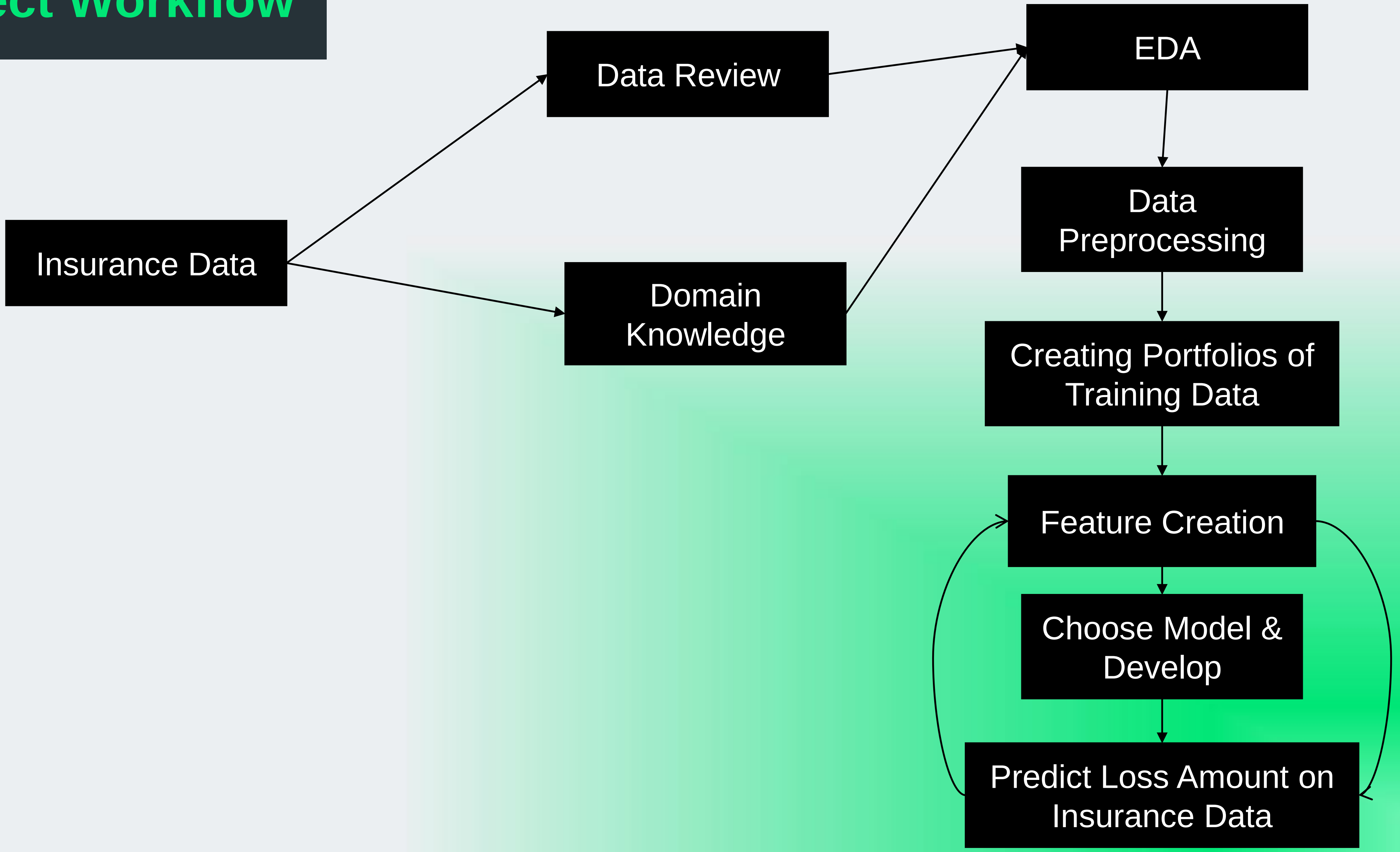
Target: natural log of portfolio loss ratio

$$\mathrm{ln\_LR} = \ln\left(\frac{Total_{Losses}}{Total_{Premium}}\right)$$

Goal

Create a competitive advantage for an Insurance Provider by developing a model(s) capable of predicting the loss ratios of policy portfolios, which will enable them to more accurately access the overall risk of a given portfolio as well as make better informed decisions on premium rates.

Project Workflow



Insurance Data

Data Review

Domain Knowledge

EDA

Data Preprocessing

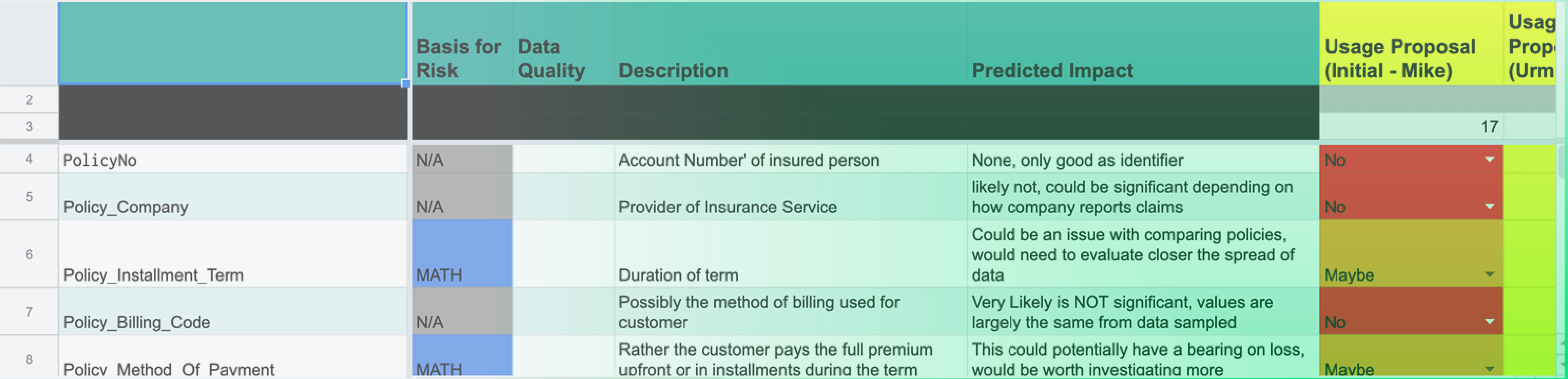
Creating Portfolios of Training Data

Feature Creation

Choose Model & Develop

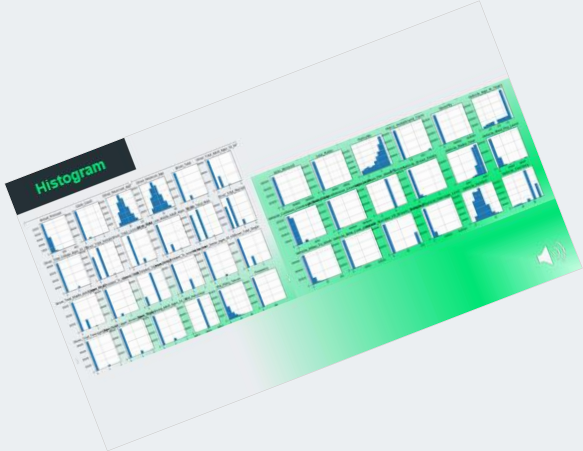
Predict Loss Amount on Insurance Data

Domain Knowledge



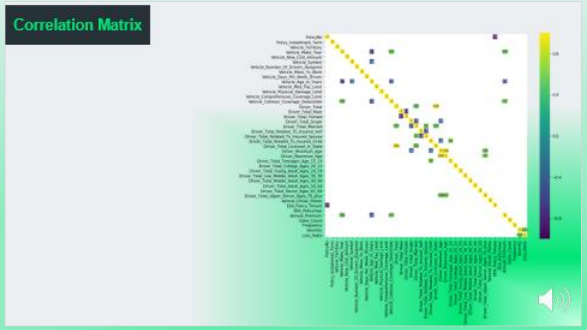
		Basis for Risk	Data Quality	Description	Predicted Impact	Usage Proposal (Initial - Mike)	Usag Prop (Urm
2							
3						17	
4	PolicyNo	N/A		Account Number' of insured person	None, only good as identifier	No	
5	Policy_Company	N/A		Provider of Insurance Service	likely not, could be significant depending on how company reports claims	No	
6	Policy_Installment_Term	MATH		Duration of term	Could be an issue with comparing policies, would need to evaluate closer the spread of data	Maybe	
7	Policy_Billing_Code	N/A		Possibly the method of billing used for customer	Very Likely is NOT significant, values are largely the same from data sampled	No	
8	Policy_Method_Of_Payment	MATH		Rather the customer pays the full premium upfront or in installments during the term	This could potentially have a bearing on loss, would be worth investigating more	Maybe	

Exploratory Data Analysis



Histogram

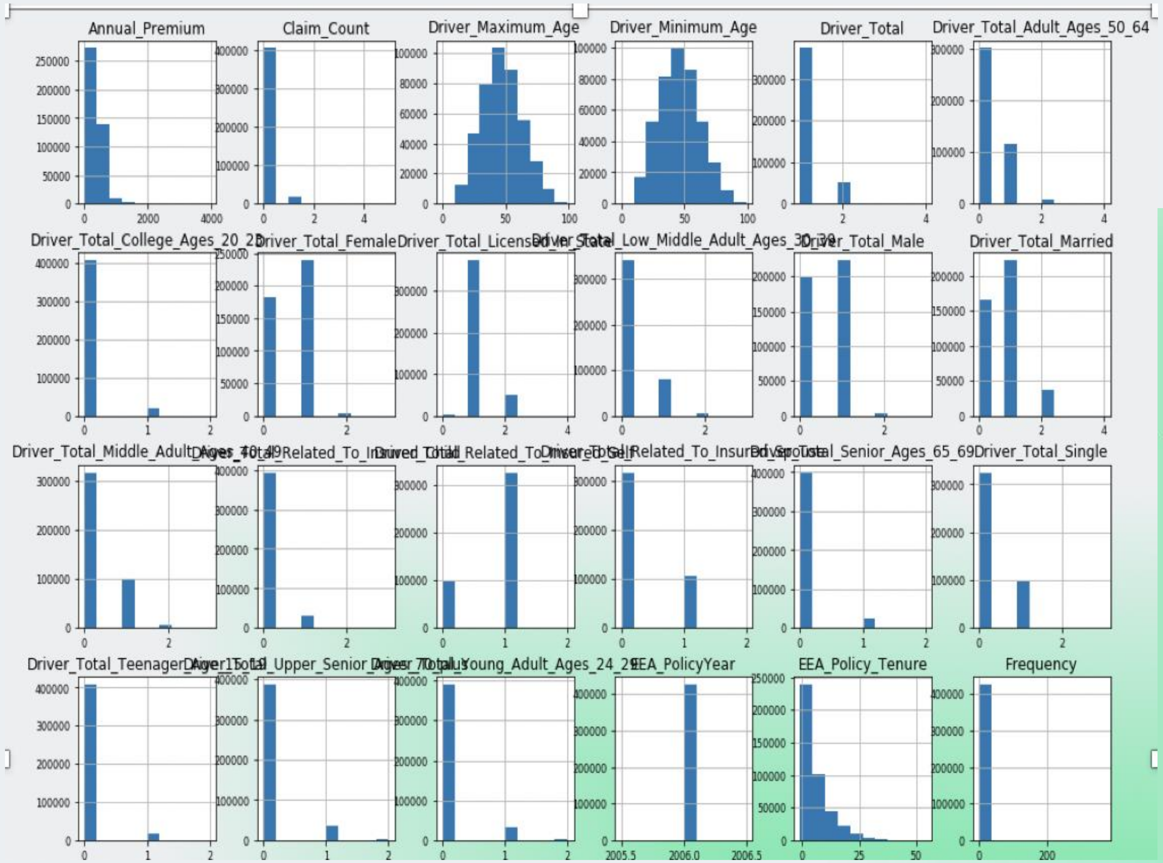
Pair Plot

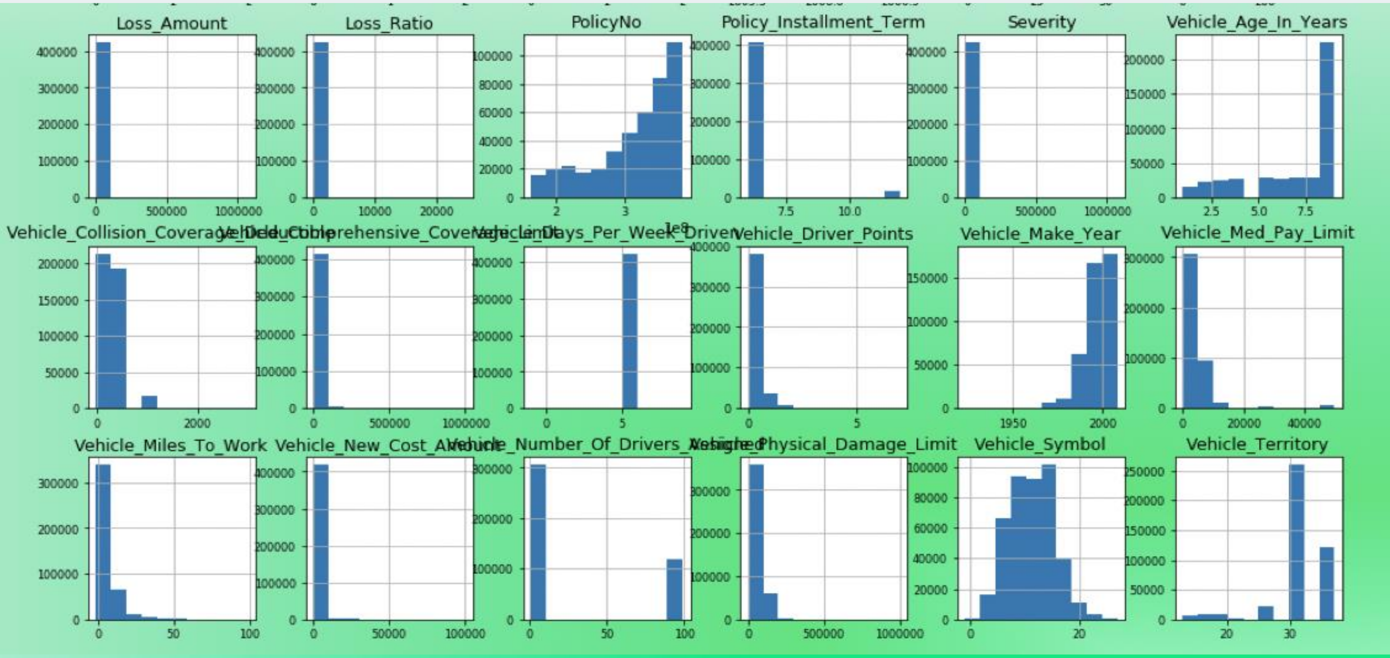


Correlation Matrix

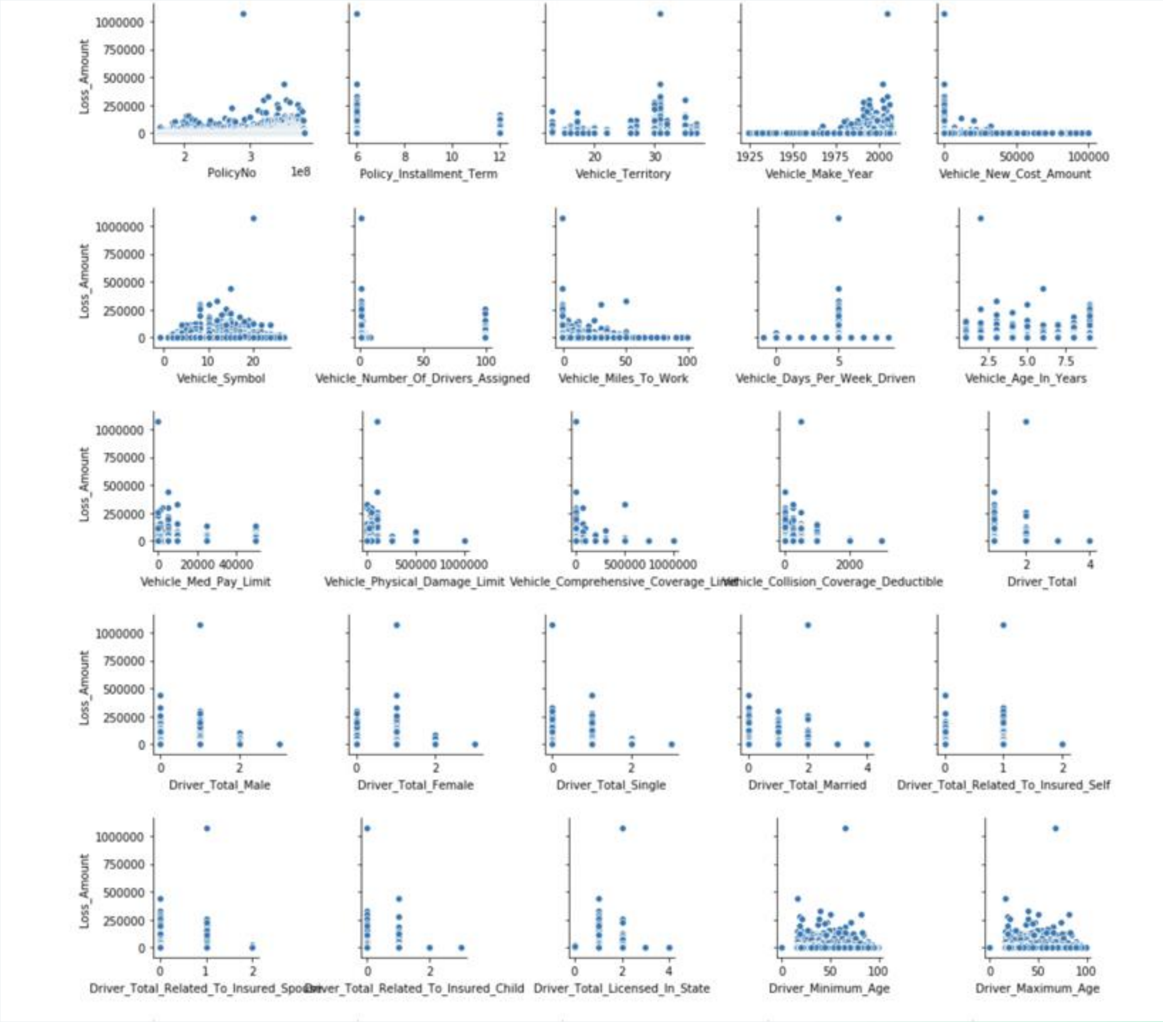
Bar Chart Comparison

Histogram

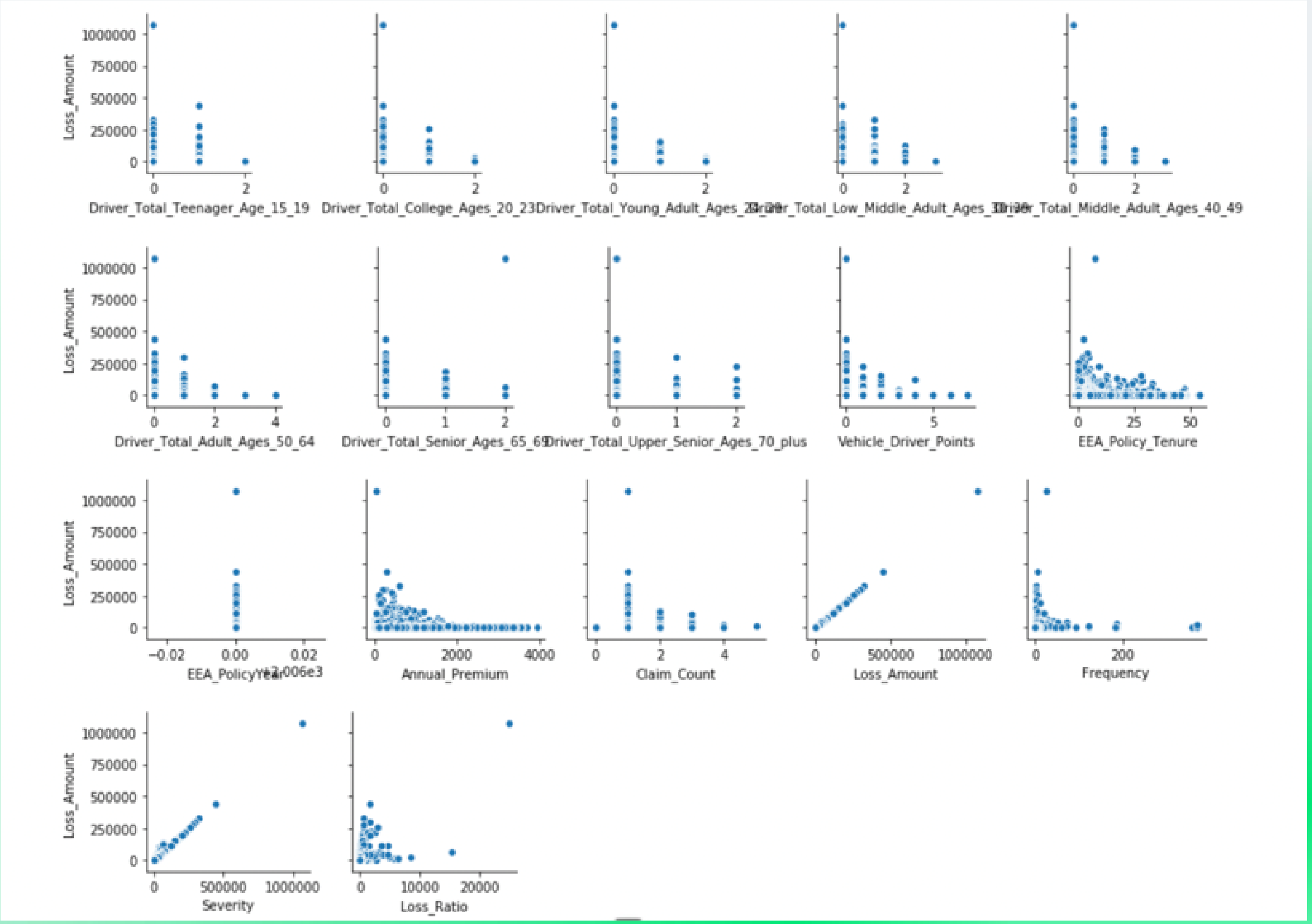

Annual_Premium
Claim_Count
Driver_Maximum_Age
Driver_Minimum_Age
Driver_Total
Driver_Total_Adult_Ages_50_64
Driver_Total_College_Ages_20_23
Driver_Total_Female
Driver_Total_Licensed_In_State
Driver_Total_Low_Middle_Adult_Ages_30_39
Driver_Total_Male
Driver_Total_Married
Driver_Total_Middle_Adult_Ages_40_49
Driver_Total_Related_To_Insured_Child
Driver_Total_Related_To_Insured_Self
Driver_Total_Related_To_Insured_Spouse
Driver_Total_Senior_Ages_65_69
Driver_Total_Single
Driver_Total_Teenager_Age_15_19
Driver_Total_Upper_Senior_Ages_70_plus
Driver_Total_Young_Adult_Ages_24_29
EEA_PolicyYear
EEA_Policy_Tenure
Frequency


Loss_Amount
Loss_Ratio
PolicyNo
1e8
Policy_Installment_Term
Severity
Vehicle_Age_In_Years
Vehicle_Collision_Coverage_Deductible
Vehicle_Comprehensive_Coverage_Limit
Vehicle_Days_Per_Week_Driven
Vehicle_Driver_Points
Vehicle_Make_Year
Vehicle_Med_Pay_Limit
Vehicle_Miles_To_Work
Vehicle_New_Cost_Amount
Vehicle_Number_Of_Drivers_Assigned
Vehicle_Physical_Damage_Limit
Vehicle_Symbol
Vehicle_Territory

Pair Plot

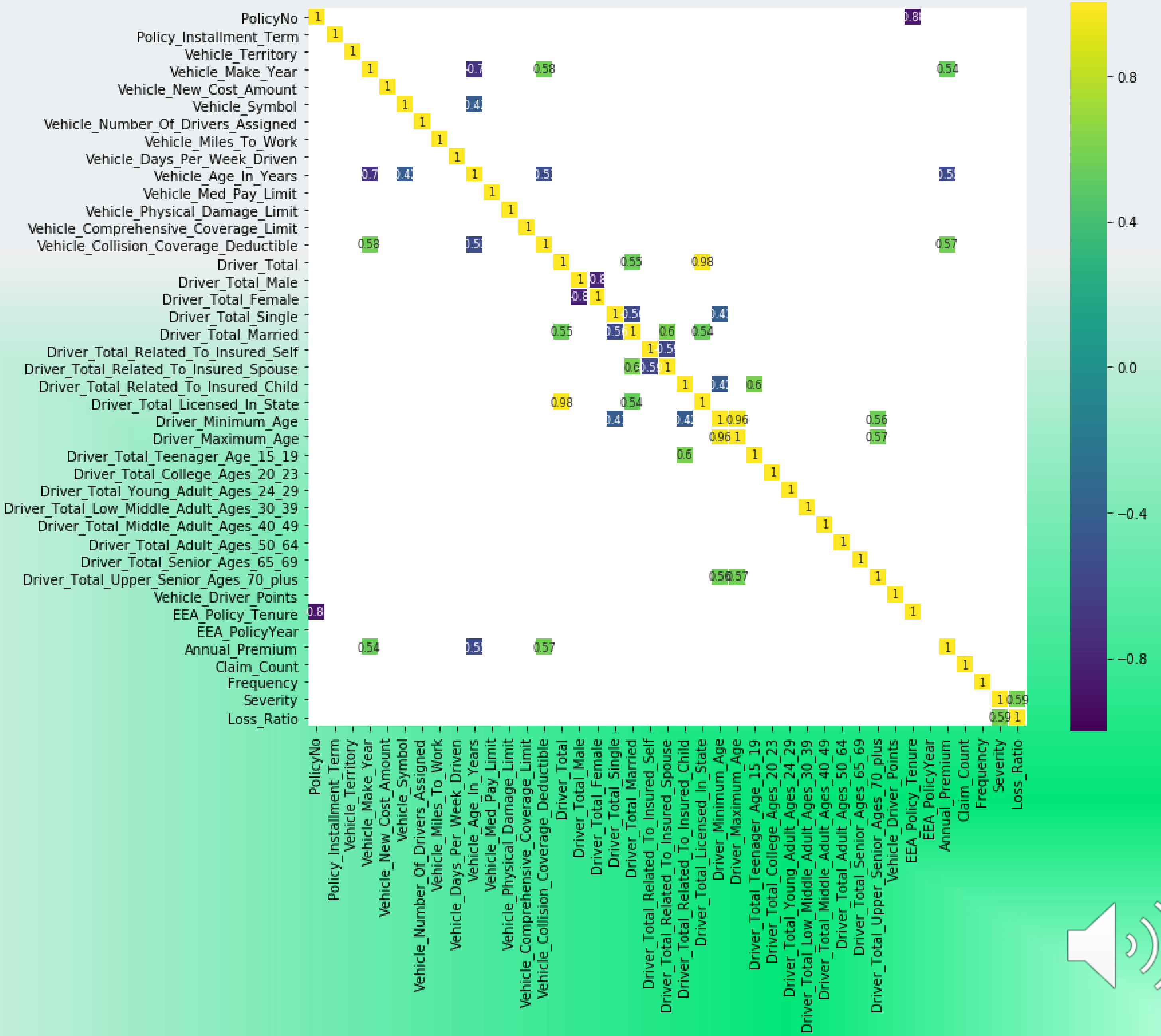


Loss_Amount
1000000
750000
500000
250000
0
PolicyNo
1e8
Policy_Installment_Term
Vehicle_Territory
Vehicle_Make_Year
Vehicle_New_Cost_Amount
Vehicle_Symbol
Vehicle_Number_Of_Drivers_Assigned
Vehicle_Miles_To_Work
Vehicle_Days_Per_Week_Driven
Vehicle_Age_In_Years
Vehicle_Med_Pay_Limit
Vehicle_Physical_Damage_Limit
Vehicle_Comprehensive_Coverage_Limit
Vehicle_Collision_Coverage_Deductible
Driver_Total
Driver_Total_Male
Driver_Total_Female
Driver_Total_Single
Driver_Total_Married
Driver_Total_Related_To_Insured_Self
Driver_Total_Related_To_Insured_Spouse
Driver_Total_Related_To_Insured_Child
Driver_Total_Licensed_In_State
Driver_Minimum_Age
Driver_Maximum_Age

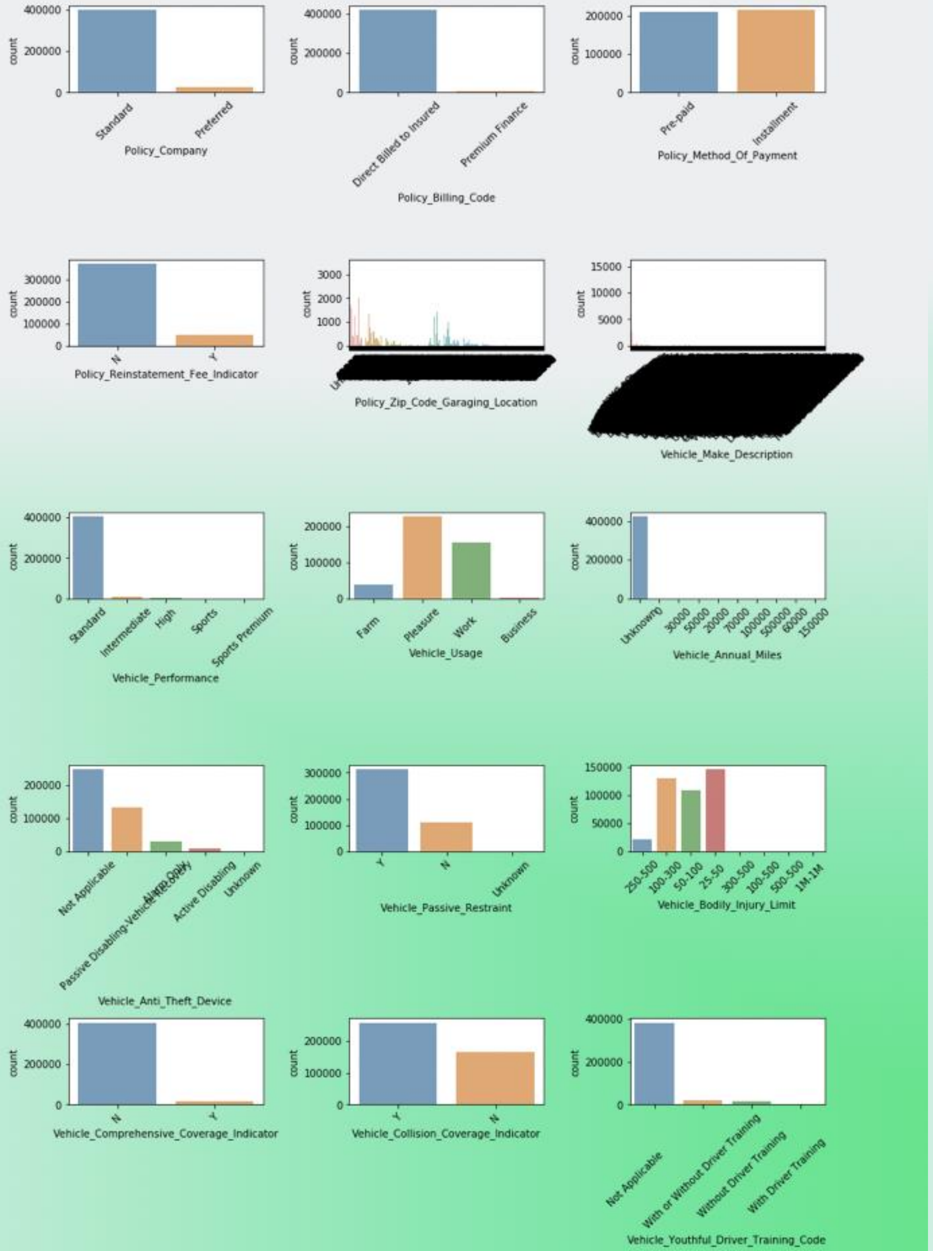


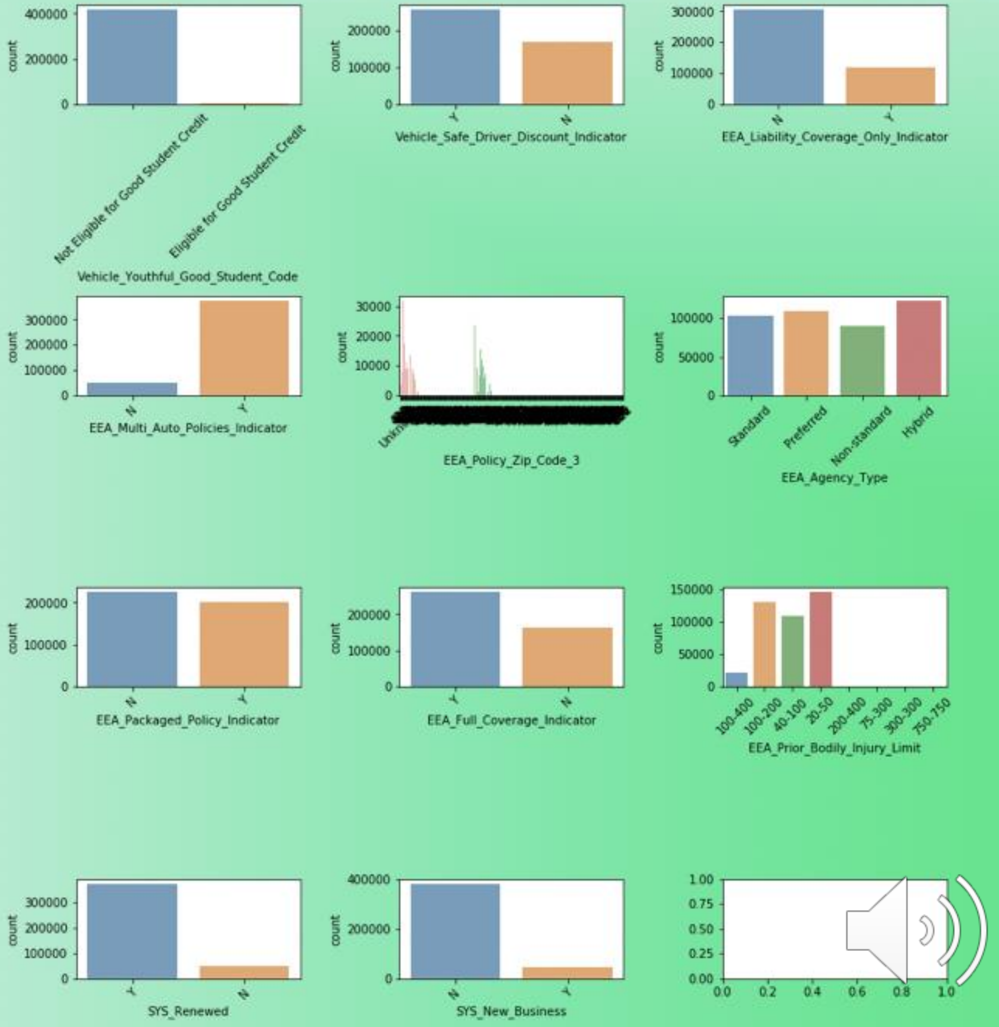
Loss_Amount
1000000
750000
500000
250000
0
Driver_Total_Teenager_Age_15_19
Driver_Total_College_Ages_20_23
Driver_Total_Young_Adult_Ages_24_29
Driver_Total_Low_Middle_Adult_Ages_30_39
Driver_Total_Middle_Adult_Ages_40_49
Driver_Total_Adult_Ages_50_64
Driver_Total_Senior_Ages_65_69
Driver_Total_Upper_Senior_Ages_70_plus
Vehicle_Driver_Points
EEA_Policy_Tenure
EEA_PolicyYear
+2.006e3
Annual_Premium
Claim_Count
Loss_Amount
Frequency
Severity
Loss_Ratio

Correlation Matrix

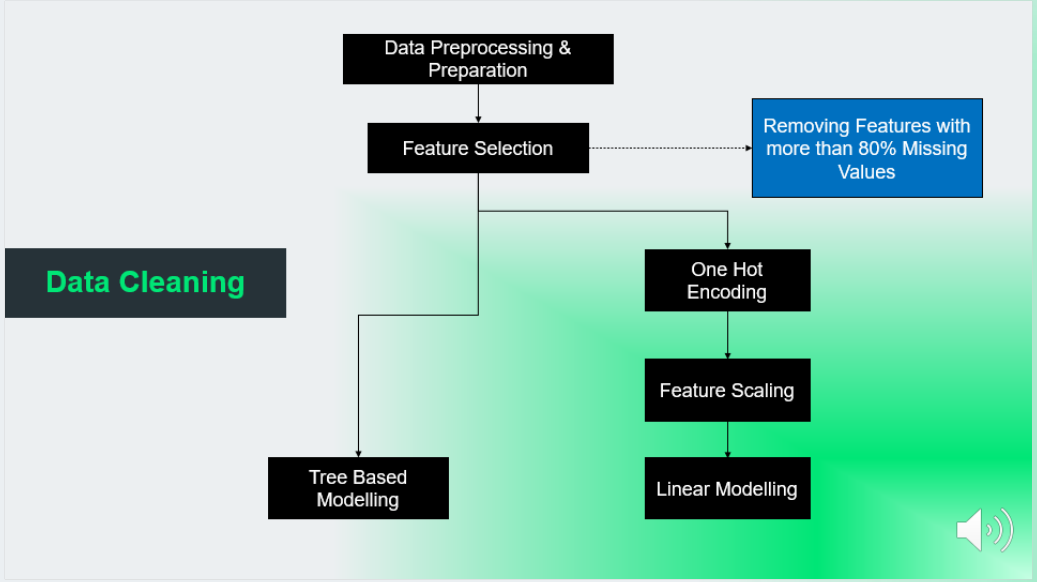


Bar Chart Comparison





Data Preprocessing



Data Cleaning

Data Preprocessing & Preparation

Feature Selection

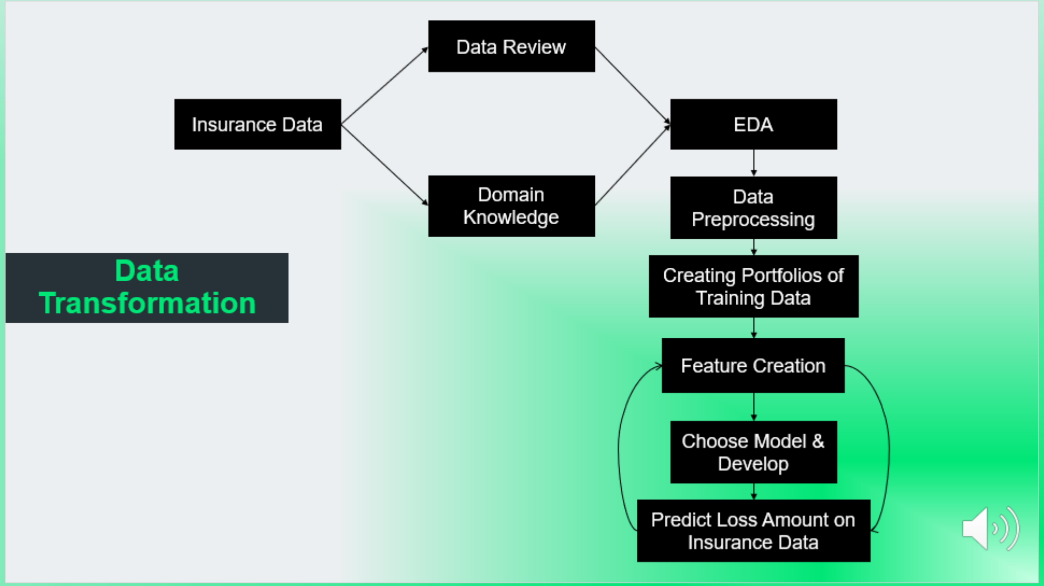
Removing Features with more than 80% Missing Values

One Hot Encoding

Feature Scaling

Linear Modelling

Tree Based Modelling



Data Transformation

Insurance Data

Data Review

Domain Knowledge

EDA

Data Preprocessing

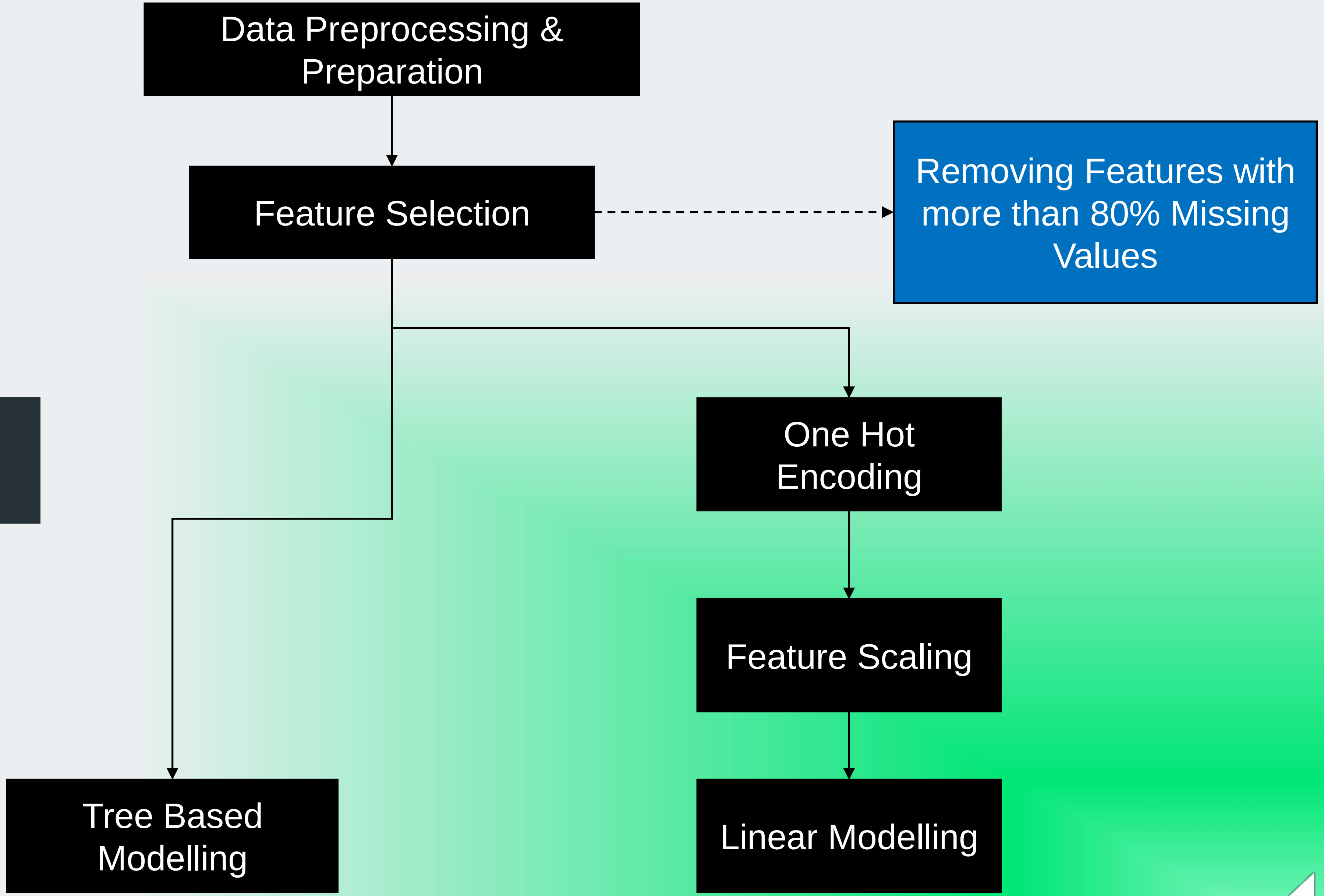
Creating Portfolios of Training Data

Feature Creation

Choose Model & Develop

Predict Loss Amount on Insurance Data

Data Cleaning



Data Preprocessing & Preparation

Feature Selection

Removing Features with more than 80% Missing Values

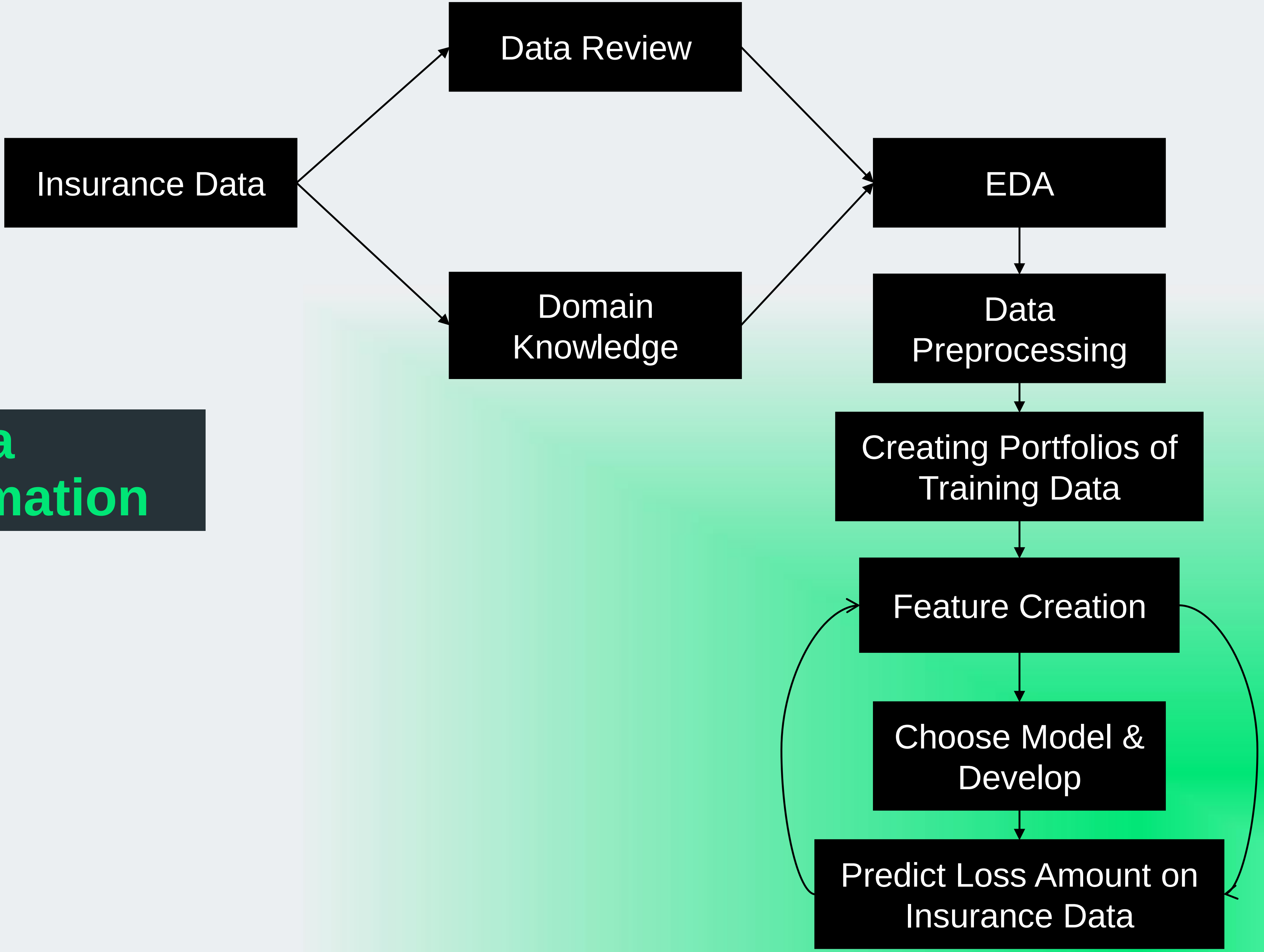
One Hot Encoding

Feature Scaling

Linear Modelling

Tree Based Modelling

Data Transformation


Insurance Data

Data Review

Domain Knowledge

EDA

Data Preprocessing

Creating Portfolios of Training Data

Feature Creation

Choose Model & Develop

Predict Loss Amount on Insurance Data

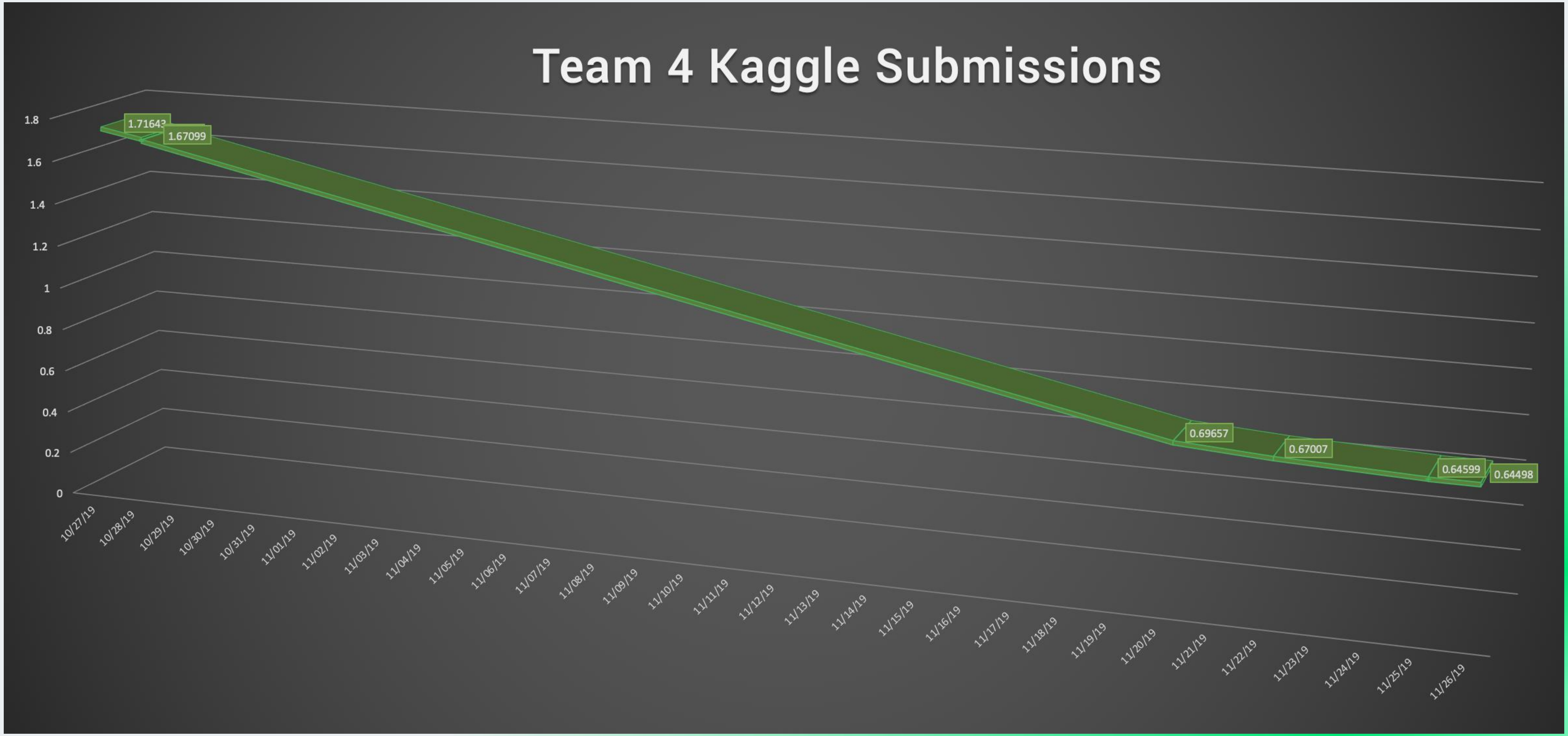
Modeling Techniques

- Decision trees
- Random Forest
- Neural Networks
- Ada Boost
- XGBoost

Evaluation Methods

- Splitting training portfolios into training and holdout sets
- Calculated the Mean Absolute Error (MAE) on these new splits
- Predicted loss amounts on our test portfolios using our model
- Calculated the loss ratio and corresponding Natural Logarithm on the testing portfolios
- Submitted results on Kaggle which provided a rank based on the MAE of our submission

Evaluation Methods



Conclusions

Statistical modeling offers a unique competitive advantage in many industries…

It offers a notable advantage for the Insurance Industry.

Thank You!