# vroom

Submitted by : Dileep Sathyan

Date: 17-September-2022

# Scope Of Analysis:

- **Identify the types of cars which cause most claims and refunds (EDA).**
- **Share the inferential insights to help reduce the operational budget.**
- **Build a Machine Learning model using the features to predict the cars that will be claimed post sales.**

# Skimming through the Datasets:

- **Car Sales**:
    - Some cars have been claimed by multiple Merchants.
    - More details about the merchants will help identify the correct record for them as well as in grouping the type of merchants (individual customer or business) from which most claims were raised.
- **Car Details**:
    - Type of cars (such as Brand_Name, Model_Name or Make_Type) if given, could help in classification.
- **Car Claims**:
    - Noticed multiple records of claim statuses for some cars.
    - A date field, if provided, could help in identifying the most recent status of such cars as well as understanding the actual flow of the refunding process.

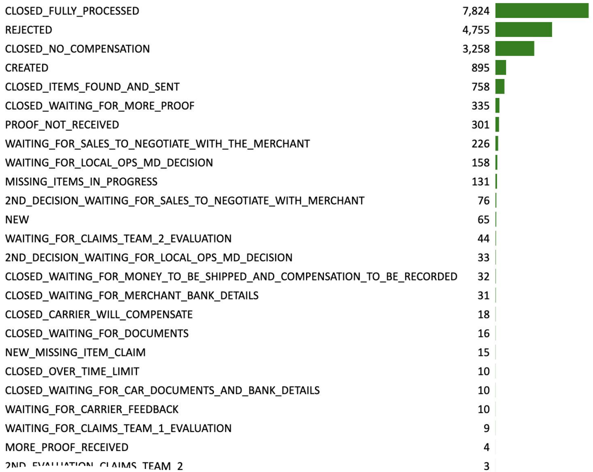
# Waterfall:

| | Count | | Amount in Euro | |
|---|---|---|---|---|
| **Unique Sales Records** | **69,551** | | **€ 518,596,726** | |
| **Total Claims Requested** | **19,021** | **27.35%** | **€ 40,567,601** | **7.82%** |

27.35% of the cars sold , are claimed by the customers.

It compounds to only 7.82% of the revenue made because the refunds are mostly done partially (and not always in full).

There are various statuses for the claims…



But, let's **group them into 4 as below**…

**Findings:**

1. 25% of the 19,021 claims are **REJECTED** and 22% are closed with **NO_COMPENSATION**.

2. 41% of the claims raised were finally **PAID** to customers (either fully or partially) which contributes to 53% of the total claimed amount. This excludes 11.9% claims which are still **UNDER_PROCESS**.

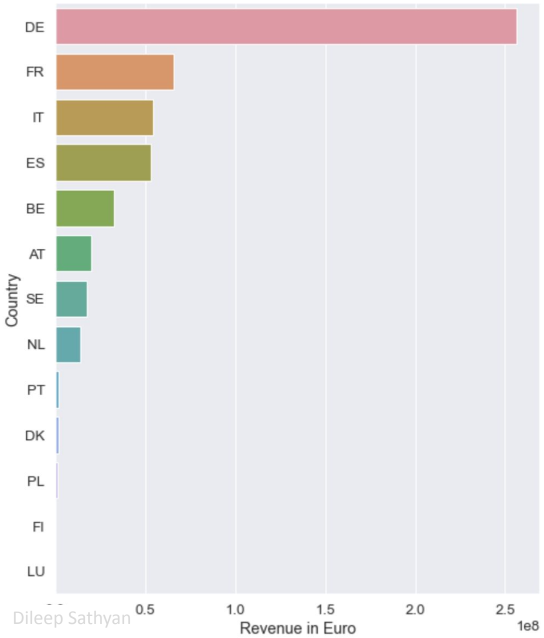
| | REJECTED | NO_COMPENSATION | UNDER_PROCESS | CLAIM_PAID | TOTAL |
|---|---|---|---|---|---|
| **# Cars** | 4,755 | 4,180 | 2,262 | 7,824 | **19,021** |
| **% Cars** | **25%** | **22%** | **11.9%** | **41%** | 100% |
| **Amount** | € 10,505,319 | € 4,940,276 | € 3,576,597 | € 21,545,408 | € 40,567,601 |
| **% Amount** | 25.9% | 12.2% | 8.8% | **53%** | 100% |

vroom

# Country Wise Performance:

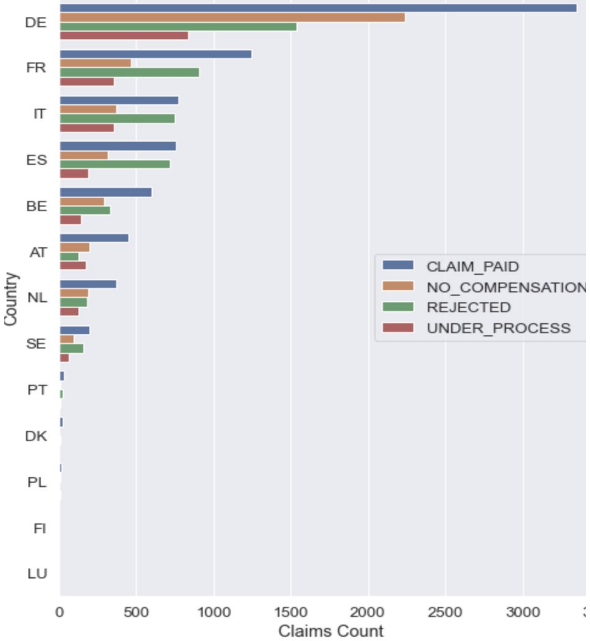
- ❖ Denmark followed by France, Italy, Spain & Belgium are the top 5 countries in terms of the revenue for vroom.
- ❖ The claims from those countries are also aligning in the same order.

## Total Revenue

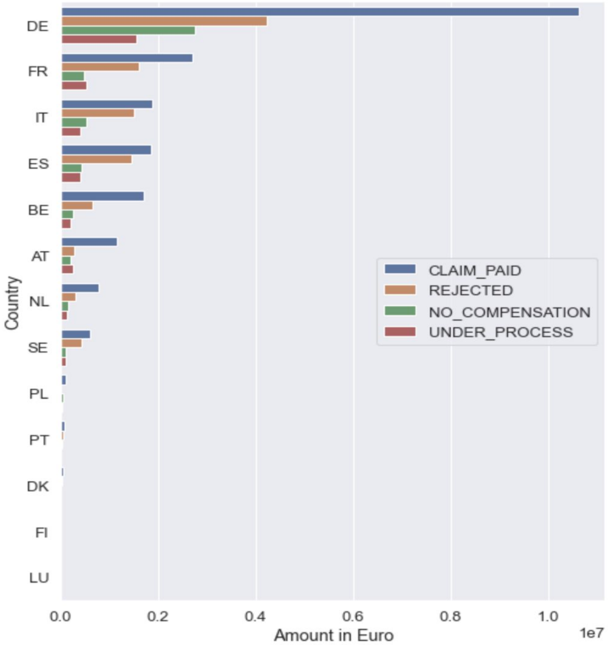


Dileep Sathyan

## Claims Count #



## Claim Amount €



# Merchants:

★ Merchant_Id: 8819318 has **claimed 100% cars** he purchased…!!!.

★ Out of his total 823 purchases, 677 claims ie., >**82% were REJECTED** and 47 claims, which is around **6% were closed with NO_COMPENSATION**.

★ The above 88% of his claim requests could have been avoided and saved considerable overhead expenses of vroom.

★ **Vroom might need to closely monitor this merchant, if not they wish to close the contract with them permanently!**

### Claims Count #

| merchant_id | REJECTED | NO_COMPENSATION | UNDER_PROCESS | CLAIM_PAID |
|---|---|---|---|---|
| 2593085 | 0 | 9 | 10 | 73 |
| **8819318** | **677** | 47 | 30 | 69 |
| 5482442 | 16 | 14 | 14 | 53 |
| 6717821 | 1 | 32 | 0 | 31 |
| 10667222 | 0 | 7 | 0 | 28 |
| 7604076 | 4 | 0 | 3 | 26 |
| 2491964 | 8 | 18 | 11 | 25 |
| 6142462 | 0 | 4 | 4 | 24 |
| 12044106 | 1 | 0 | 2 | 21 |
| 13683897 | 8 | 7 | 0 | 19 |

### Claim Amount €

| merchant_id | REJECTED | NO_COMPENSATION | UNDER_PROCESS | CLAIM_PAID |
|---|---|---|---|---|
| 6142462 | € 0 | € 28,398 | € 31,508 | € 281,739 |
| 2593085 | € 0 | € 0 | € 14,436 | € 185,465 |
| 10667222 | € 0 | € 10,129 | € 0 | € 179,093 |
| 11360764 | € 0 | € 0 | € 11,907 | € 134,155 |
| 5482442 | € 38,218 | € 12,550 | € 35,888 | € 130,516 |
| 703705 | € 6,762 | € 0 | € 0 | € 123,567 |
| **8819318** | **€ 1,437,082** | € 53,992 | € 33,926 | € 118,640 |
| 9575737 | € 0 | € 0 | € 5,649 | € 117,461 |
| 3260198 | € 0 | € 17,630 | € 0 | € 108,248 |
| 5509281 | € 4,817 | € 0 | € 18,800 | € 100,915 |

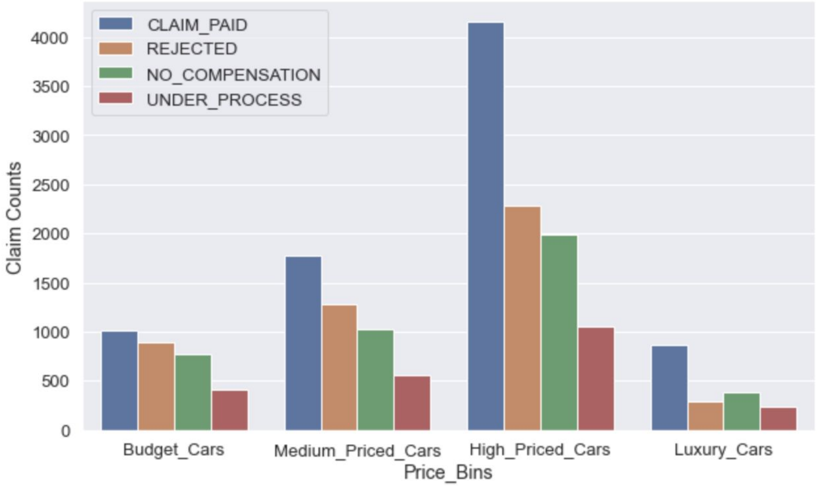
→ Merchant_ids of all car purchases are not available in the dataset for further deep dive.

# Selling Prices:

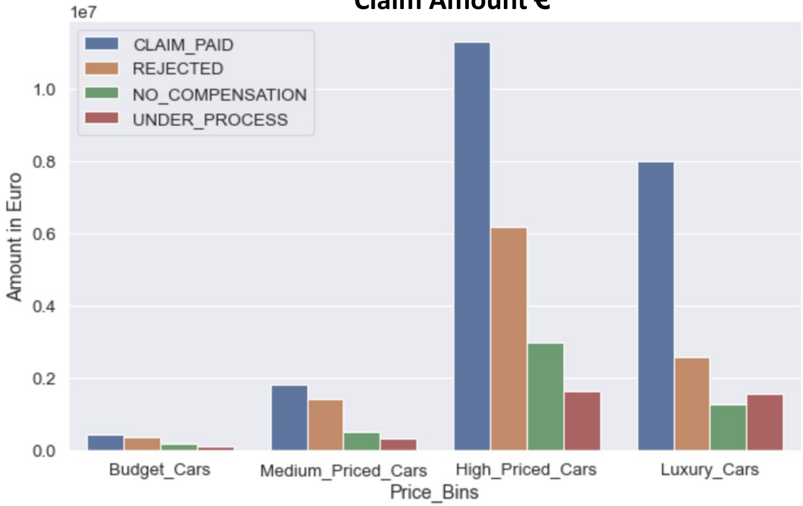
- **High Priced Cars** followed by **Medium Priced Cars** are claimed **more frequently** than the **Budget Cars** or **Luxury Cars**.
- However Amount of Money Claimed by Luxury Cars buyers overtakes Medium Ranges due to their high sell price.

| Price_Bins | Price Range | Total Revenue | Percentage% |
|---|---|---|---|
| Budget_Cars | 0 - 2500 € | € 26,204,010 | 5% |
| Medium_Priced_Cars | 2501 - 5000 € | € 60,109,897 | 12% |
| High_Priced_Cars | 5001 - 20000 € | € 282,488,826 | 54% |
| Luxury_Cars | 20001 & Above € | € 149,793,993 | 29% |
| | | **€ 518,596,726** | **100%** |

## Claims Count #



## Claim Amount €

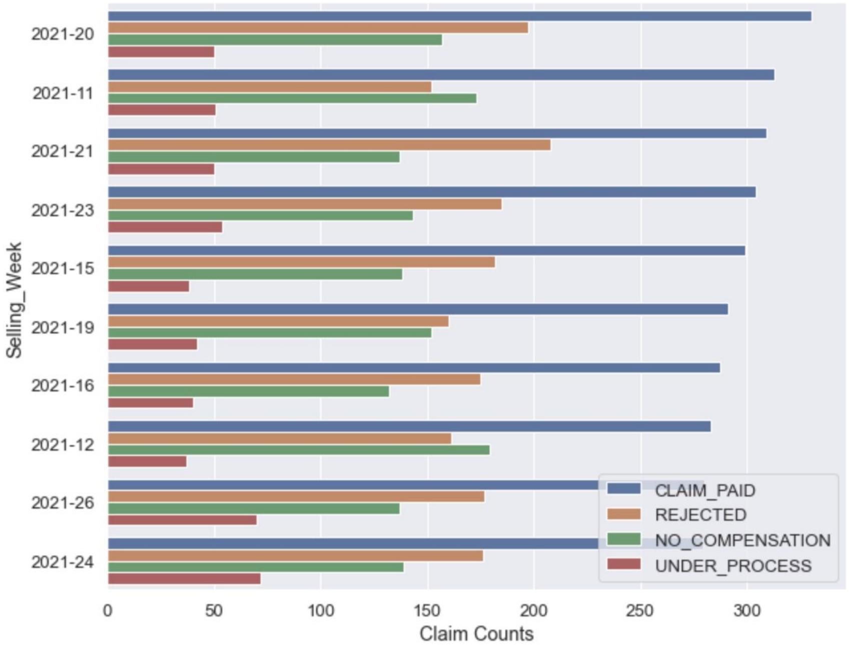


# Selling Week:

❖ Claims are randomly distributed across the Selling Weeks & hence doesn't have much correlation.

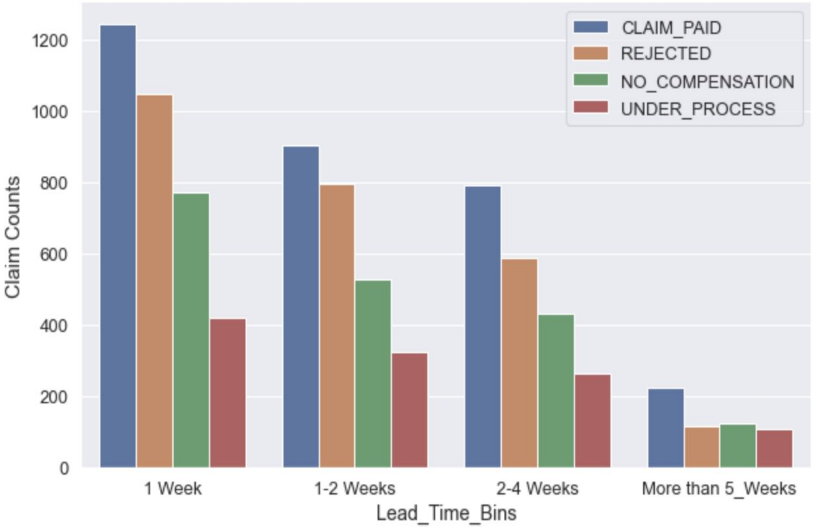
**Claims Count #**

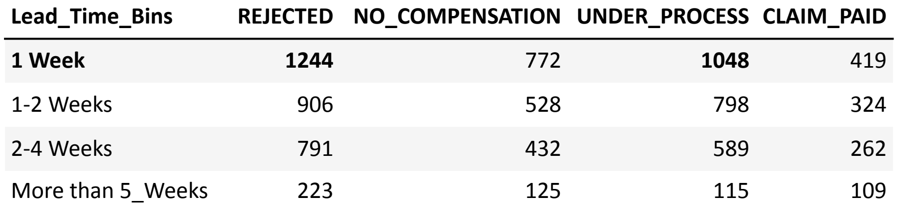
| SELLING WEEK | REJECTED | NO_COMPENSATION | UNDER_PROCESS | CLAIM_PAID |
|---|---|---|---|---|
| 2021-20 | 197 | 157 | 50 | 330 |
| 2021-11 | 152 | 173 | 51 | 313 |
| 2021-21 | 208 | 137 | 50 | 309 |
| 2021-23 | 185 | 143 | 54 | 304 |
| 2021-15 | 182 | 138 | 38 | 299 |
| 2021-19 | 160 | 152 | 42 | 291 |
| 2021-16 | 175 | 132 | 40 | 287 |
| 2021-12 | 161 | 179 | 37 | 283 |
| 2021-26 | 177 | 137 | 70 | 280 |
| 2021-24 | 176 | 139 | 72 | 279 |



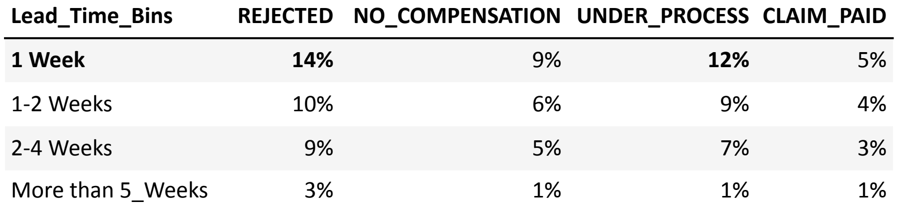
# Days to Transport:

- **The sooner the cars were transported after payment date, the frequent the claims are…!!**
- The cars which were transported within 1 week of payment, seem to have claimed more frequently than the cars having more lead time to ship.
- We could further drill down this perspective if the claim_dates were provided in the dataset, in order to identify the types of claims requested in the first week, 2nd week etc..





| Lead_Time_Bins | REJECTED | NO_COMPENSATION | UNDER_PROCESS | CLAIM_PAID |
|---|---|---|---|---|
| **1 Week** | **1244** | 772 | **1048** | 419 |
| 1-2 Weeks | 906 | 528 | 798 | 324 |
| 2-4 Weeks | 791 | 432 | 589 | 262 |
| More than 5_Weeks | 223 | 125 | 115 | 109 |



| Lead_Time_Bins | REJECTED | NO_COMPENSATION | UNDER_PROCESS | CLAIM_PAID |
|---|---|---|---|---|
| **1 Week** | **14%** | 9% | **12%** | 5% |
| 1-2 Weeks | 10% | 6% | 9% | 4% |
| 2-4 Weeks | 9% | 5% | 7% | 3% |
| More than 5_Weeks | 3% | 1% | 1% | 1% |

# Other Car Features:

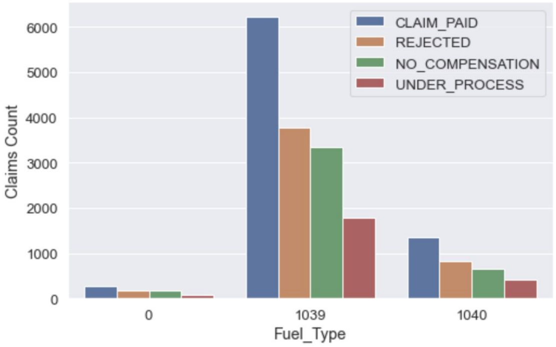
❖ Fuel Type: ‘1039’ seems to have more claims than ‘1040’.

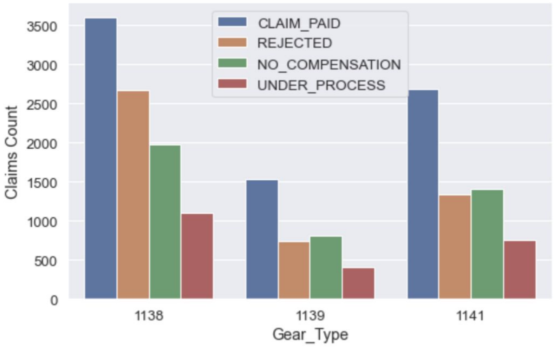
❖ Gear Type: ‘1138’ followed by ‘1141’ have more claims than ‘1139’

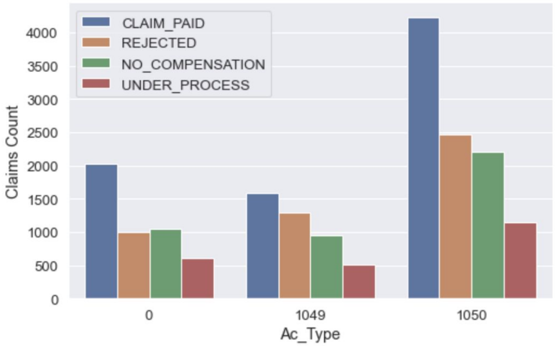
❖ AC Type: ‘1050’ has considerably higher claims than other AC types.

➔ More info about these codes will help in further analysis.

❑ Other features such as has_tuning, has_airbags, has_alarm_system, navigation_system, xenon_lights and radio_system have considerable NULL values and 0 values which makes the meaningful summarisation difficult.







# Focusing on the 4755 REJECTED Claims:

- Out of the 4755 Rejected claim requests, 32% are from Denmark.
- Almost half of the Rejected claims are for High Priced Cars which range from € 5,000 to € 20,000.
- Around 41% of the Rejected claims for the cars which got transported to the customer within 1 week of payment.

| Country | Amount in Euro | Claims Count | Percentage |
|---|---:|---:|---:|
| **DE** | **€ 4,218,028** | **1536** | **32%** |
| FR | € 1,594,313 | 907 | 19% |
| IT | € 1,510,409 | 749 | 16% |
| ES | € 1,441,395 | 717 | 15% |
| BE | € 647,437 | 335 | 7% |
| NL | € 311,180 | 185 | 4% |
| SE | € 425,781 | 161 | 3% |
| AT | € 283,197 | 127 | 3% |
| PT | € 41,635 | 24 | 1% |
| DK | € 25,928 | 12 | 0% |
| PL | € 6,016 | 2 | 0% |
| **TOTAL** | **€ 10,505,319** | **4755** | **100%** |

| Price_Bins | Amount in Euro | Claims Count | Percentage |
|---|---:|---:|---:|
| **High_Priced_Cars** | € 6,165,567 | **2289** | **48%** |
| Medium_Priced_Cars | € 1,406,976 | 1282 | 27% |
| Budget_Cars | € 359,601 | 897 | 19% |
| Luxury_Cars | € 2,573,175 | 287 | 6% |
| **TOTAL** | **€ 10,505,319** | **4755** | **100%** |

| Lead_Time | Amount in Euro | Claims Count | Percentage |
|---|---:|---:|---:|
| **1 Week** | € 1,859,061 | **1048** | **41%** |
| 1-2 Weeks | € 1,854,277 | 798 | 31% |
| 2-4 Weeks | € 1,204,716 | 589 | 23% |
| More than 5_Weeks | € 270,807 | 115 | 5% |
| **TOTAL** | **€ 5,188,861** | **2550** | **100%** |

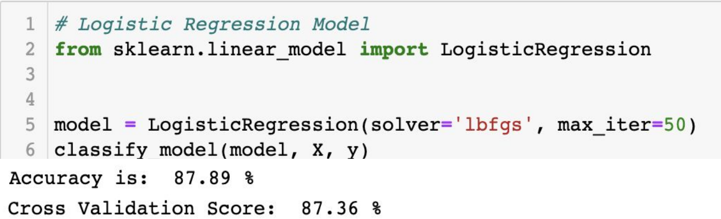
➔ Only 2550 cars out of 4755, were requested to be transported to customers. The rest were collected from vroom's premises.

# Building Predictive Model:

**Steps taken to build the model.**

- Data Cleaning and handling missing data points
- Feature Engineering: ‘days_to_ship’, ‘is_claim_successful’ (the dependant variable has been calculated by grouping both already paid claims & the ones under process)
- Data Normalization: Amount fields were normalised by taking their logs for better prediction results.
- Built Correlation Matrix to identify the prominent numerical variables
- Label Encoding: to code the categorical variables suitable for modeling.
- Train Test Split: Separated 80% of data for model training and 20% of testing.
- Model building using Logistic Regression, Decision Tree Classifier and Random Forest Classifiers.
- Identifying the best model using the performance and scores after Cross Validation (cv=10).
- Visualizing the prediction using a Confusion Matrix.

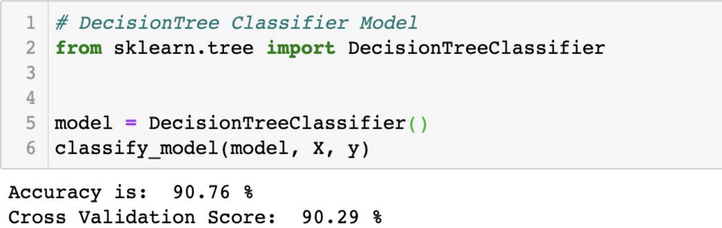
# Model Results & Evaluation:



```python
# Logistic Regression Model
from sklearn.linear_model import LogisticRegression


model = LogisticRegression(solver='lbfgs', max_iter=50)
classify_model(model, X, y)
```

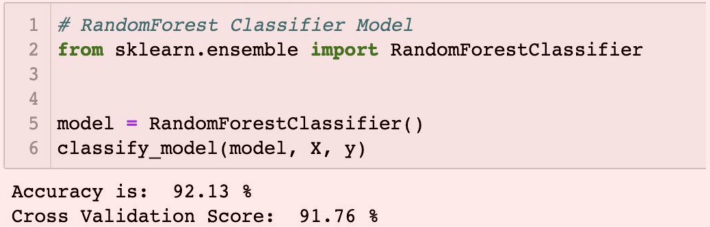
```
Accuracy is:  87.89 %
Cross Validation Score:  87.36 %
```



```python
# DecisionTree Classifier Model
from sklearn.tree import DecisionTreeClassifier


model = DecisionTreeClassifier()
classify_model(model, X, y)
```

```
Accuracy is:  90.76 %
Cross Validation Score:  90.29 %
```



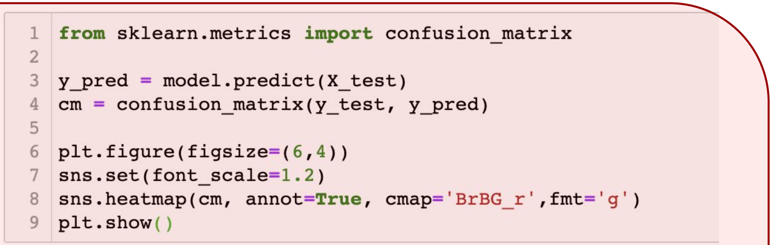
```python
# RandomForest Classifier Model
from sklearn.ensemble import RandomForestClassifier


model = RandomForestClassifier()
classify_model(model, X, y)
```

```
Accuracy is:  92.13 %
Cross Validation Score:  91.76 %
```

- ★ **Random Forest Classifier Model predicts the cars which could get claimed, @ 92.13% Accuracy.**

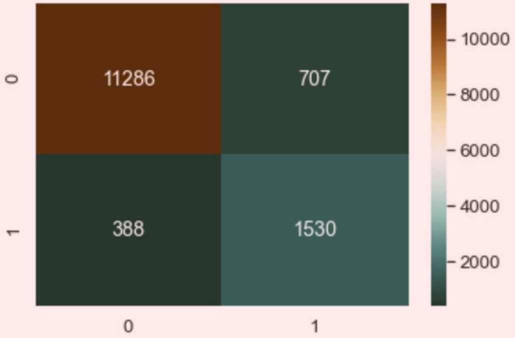
- ★ This is achieved with the features available in the dataset and **can be improved considerably if given more features** (such as Merchant details of ALL cars, Brand_Name, Model_Name, Make_Type, Claim_Date etc…)



```python
from sklearn.metrics import confusion_matrix

y_pred = model.predict(X_test)
cm = confusion_matrix(y_test, y_pred)

plt.figure(figsize=(6,4))
sns.set(font_scale=1.2)
sns.heatmap(cm, annot=True, cmap='BrBG_r',fmt='g')
plt.show()
```



vroom

# Thank you..!

Dileep Sathyan