Aqsa

Aqsa.akhtar31@gmail.com

Car Insurance Policy Sales Prediction Based on Statistical/Machine Learning Model

Sales Prediction

# Section 1 (Outline and Exploration)

With recent advancements in technology, we are able to capture a huge pile of data. 90% of the world’s data is created in the previous year alone, and our current daily data output has reached 2.5 quintillion bytes. This data can be used in many optimal ways to implement versatile business solutions. Today, around 85% of the companies are trying to be data driven. The global data science market is expected to reach $128 billion by 2020 according to NewVantage Partner in their very recent survey [1]

Traditional data used to be small and structured, and thus, could be easily analyzed using simple BI tools. Today’s data, being humongous, is not well-structured and therefore, the raw data needs to be processed before exploring insightful trends. However, before processing and exploring the data, the first step is to have a clear understanding of the problem statement and the task.

## Defining Data Statement

The objective of this project is to consider various factors and dynamics of the available data to develop a Machine Learning based statistical model that can help us in identifying the people who are more likely to buy a car insurance policy from Hastings Direct.

## Programming Platform

For the completion of this project, I have used Python as the programming language. The libraries I have used from Python include: Pandas, Scikit-learn, Matplotlib, Seaborn.

## Getting Data

For this project, the data was provided by Hastings Direct. It consisted of 50,000 observations, where each had 12 features. Each observation defines the following attributes of the single policy holder:

* **Age**: The age of the customer.
* **Veh\_Value**: The market value of the vehicle in GBP.
* **Tax**: Insurance Premium Tax Rates which add an additional fee on top of the customer’s price.
* **Price**: The annual price a customer will pay for this policy.
* **Veh\_Mileage**: The mileage of the vehicle on this policy.
* **Credit Score**: The customer’s credit score.
* **Licence\_Length**: Amount of time a customer has held their full license in years.
* **Date**: The date when the policy will start.
* **Marital\_Status**: The customer’s marital status:
  + M: Married
  + S: Single
  + D: Divorced
* **Payment\_Type**: How the customer will pay their price:
  + Installments: Monthly payments where some interest rate is applicable.
  + Cash: Full amount for annual price up front.
* **Veh\_Reg\_Year**: The year when the vehicle was registered with the DVLA.
* **Sale**: Whether or not a policy is sold:
  + 0: Customer **did not** buy the policy.
  + 1: Customer **did** buy the policy.

## Data Analysis and Explorations

Data analysis and Exploration consists of looking for insightful trends and learning more about the data, e.g., its distribution, any anomalies and outliers. I used Matplotlib and Seaborn in Python to visualize the various trends accompanied by suitable plotting methodology, e.g., boxplot, histogram, bar plot for univariate and bivariate analysis.

## Data Cleaning

In any data science project, data cleaning along with data explorations takes the most of our time. Also, it is believed to be the most important part of a project, as all our predictions and analysis depend on how well we have transformed the raw data to the presentable form. In data exploration, I identified the missing values in each attribute as shown in *figure 1* along with outliers, and thus, replaced them based on the feature type as explained below.

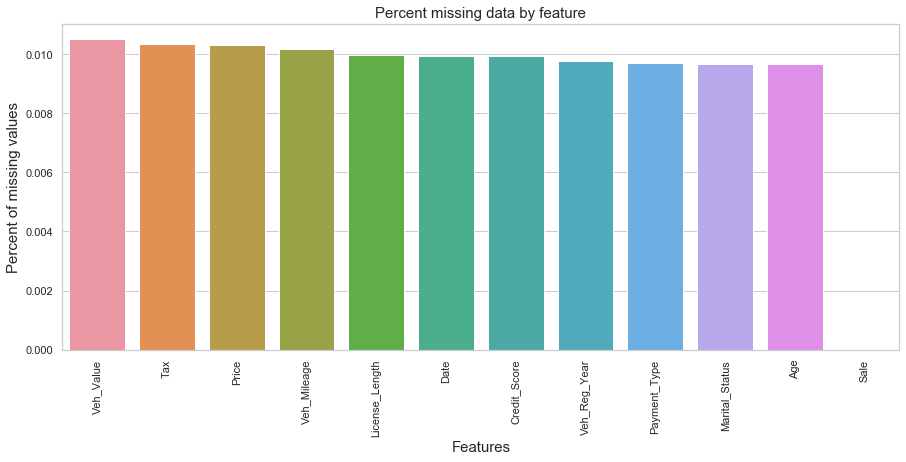


Figure 1: Percentage of missing values in dataset

**Categorical Columns**: Categorical values were imputed with the mode as it represents the value that is the most common for the given column.

**Continuous Columns**: I used the median to replace the missing values present in the numeric columns.

**Outliers:** Looking at the nature of the data, outliers were replaced with the 5th and 95th quantile score of each column.

## Feature Engineering

Feature engineering can boost the performance of our Machine Learning model as it can interpret data to make it easier for us to understand. It involves cleaning the data as well as making new variables from the already present variables along with converting categorical data to numeric values.

One hot encoding technique was used to convert categorical data to numeric values and was applied on two attributes of our dataset, i.e., *Marital Status* and *Payment Type*. Each category was converted into a binary column, where only one can have ‘True’ value.

*Day*, *month* and *year* were separated from the ‘*Date*’ category column. Each of these attributes was stored in a new column with their respective names.

## Apply Predictive Model

Logistic Regression was applied on the binary class classification, which is explained in detail in *section 2* of this document.

## Evaluation Metrics

Building Machine Learning is not just about getting predictions, rather it goes way beyond. We need to check the model, evaluate it based on the feedback we obtain from our evaluation metrics used and improve it. These metrics are critical to our machine learning model because it is possible that our model works well with one evaluation metric and not with another metric. Hence, the need to check whether our model is working correctly and optimally is essential. The evaluation metrics I used in this project are:

### Confusion Matrix

A confusion matrix is typically a 2x2 table, which is often used to describe the performance of our model. It gives us the measure of accuracy, sensitivity, specificity, as well as recall, and precision. Test data set is used to create this matrix using the predictions obtained from the model. Classification matrix produces four outcomes

1. True Positive: Correct Positive Predictions
2. False Positive: False Positive Predictions
3. True Negative: Correct Negative Predictions
4. False Negative: False Negative Predictions

Using these four outcomes we can calculate the values of accuracy, error rate, precision, recall, F1 score, sensitivity and specificity.

### AUC ROC Curve

Receiver Operating Characteristics (ROC) are plotted using ‘True Positive Rate’ with ‘False Positive Rate’ at various thresholds. AUC represents the measure of how well our model is capable of separating the binary class, whereas ROC is the probability curve. Higher the AUC, better is the result, as it distinguishes well between 1 and 0 class. This results in higher chances of predicting whether a person bought the policy or not. Here, True Positive rate is defined as

And false positive rate is defined as,

The value of 0.5 AUC shows that our model is not able to distinguish between positive class and negative class, whereas the value of 0 shows that it’s showing negative class as positive class and vice versa.

### Precision, Recall and F1 score

Another solution to our problem is using precision, recall and F1 score. Precision is the amount of retrieved instances that are actually relevant, whereas, recall expresses the ability to get the list of all relative instances. If we classify all the people who have not bought our policy as the one, it means we will have 0 precision and recall, as there are no true positive in it. However, if we tweak our model a bit and classify one person who has bought our policy right, we will still have a very low recall as there are a lot of false negatives in our model. Also, if we classify all the people like the one who has bought the policy, we will be adding those people in the list as well who have not bought the policy and thus, our precision will be low due to the presence of false positives. Increasing recall, decreases precision and vice versa. In an ideal model, our target should have both recall and precision high.

To get the optimal results from the precision and recall, we can blend the two together with the help of the F1 score, which is given as the harmonic mean of precision and recall.

Harmonic mean is used to cater with the extreme values. Recall 1 and precision 0 gives an average 0.5 but with the F1, it gives 0 and tries to balance both the recall and precision to get the optimal results.

## Data Analysis

Data visualization is the graphical representation of the information and data. By using visual elements like charts, graphs and maps, data visualization tools provide an accessible way to see and understand trends, outliers and patterns in data.

Let’s start with the ‘describe’ command to gain some insight of data distribution:

Figure 2: Numeric data descriptive analysis

‘Describe’ function of pandas allows us to view simple statistical description of our numeric data. As shown in the figure above, all the variables have some missing values along with outliers in columns *Credit\_Score* where the maximum *Credit\_Score 9999* lies far away from other a values alongside *License\_Length* where it is not likely to have the negative value for the amount of time a customer has held their full licence in years. A summary of each columns mean, min, max is given accompanied by standard deviation value which is rather high for some columns.

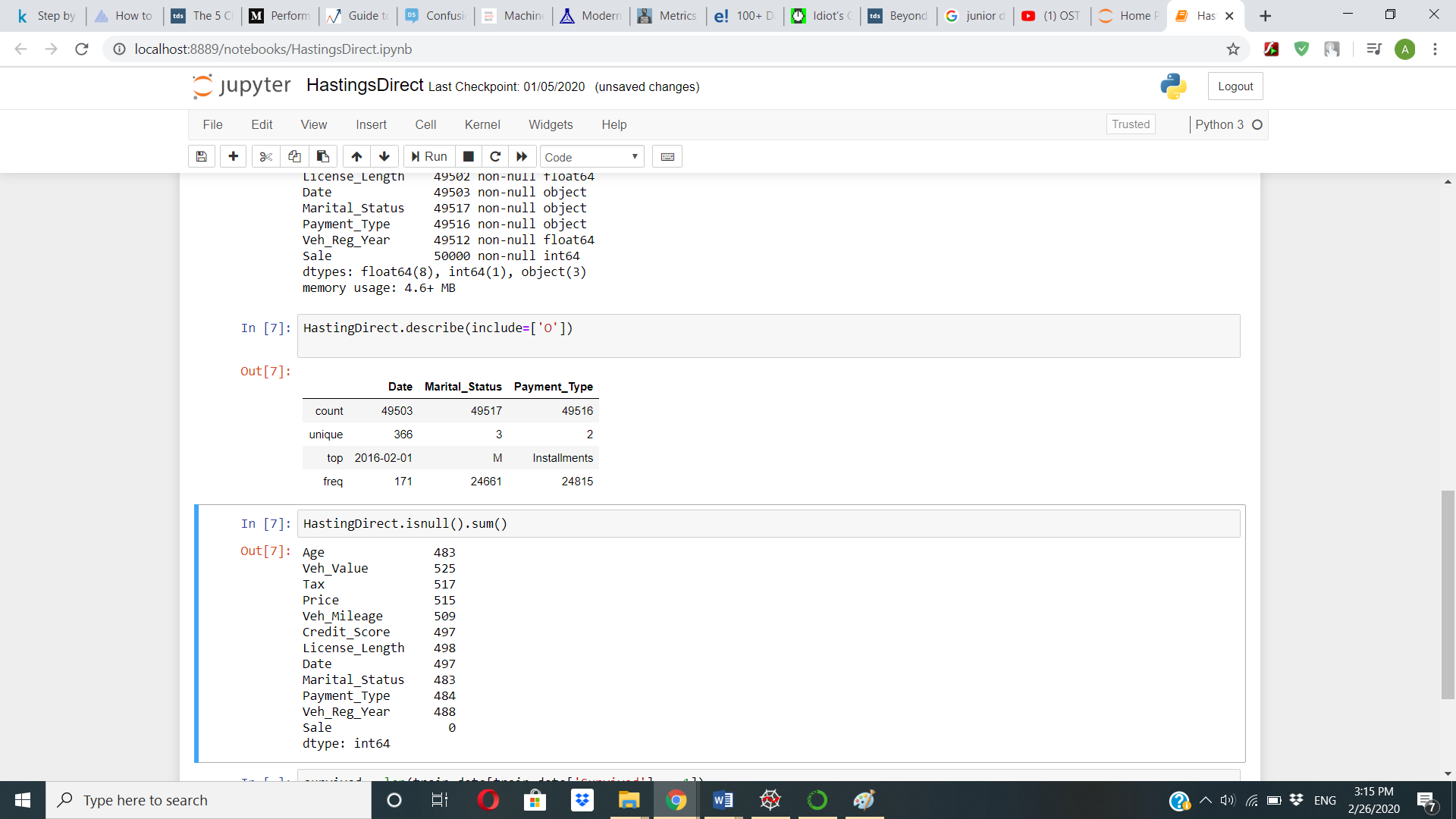


Figure 3: Categorical data descriptive analysis

As we can clearly see from the *figure 3*, the most people in our dataset are *married* and the majority of people paid policy price in *installments*.

A quick bar chart of dependent variable ‘*Sale*’ showed us the class that dominated our dataset.

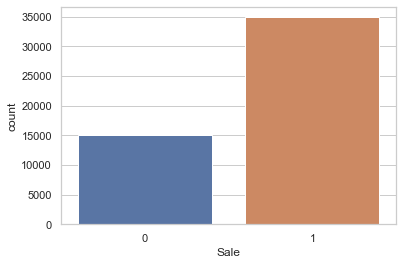


Figure 4: People who bought policies vs people who didn't buy policy

The number of policies that were bought clearly dominate the number of policies that were rejected by the customers, as 69.8% of the policies were bought as compared to 30.2% of the policies that were rejected.

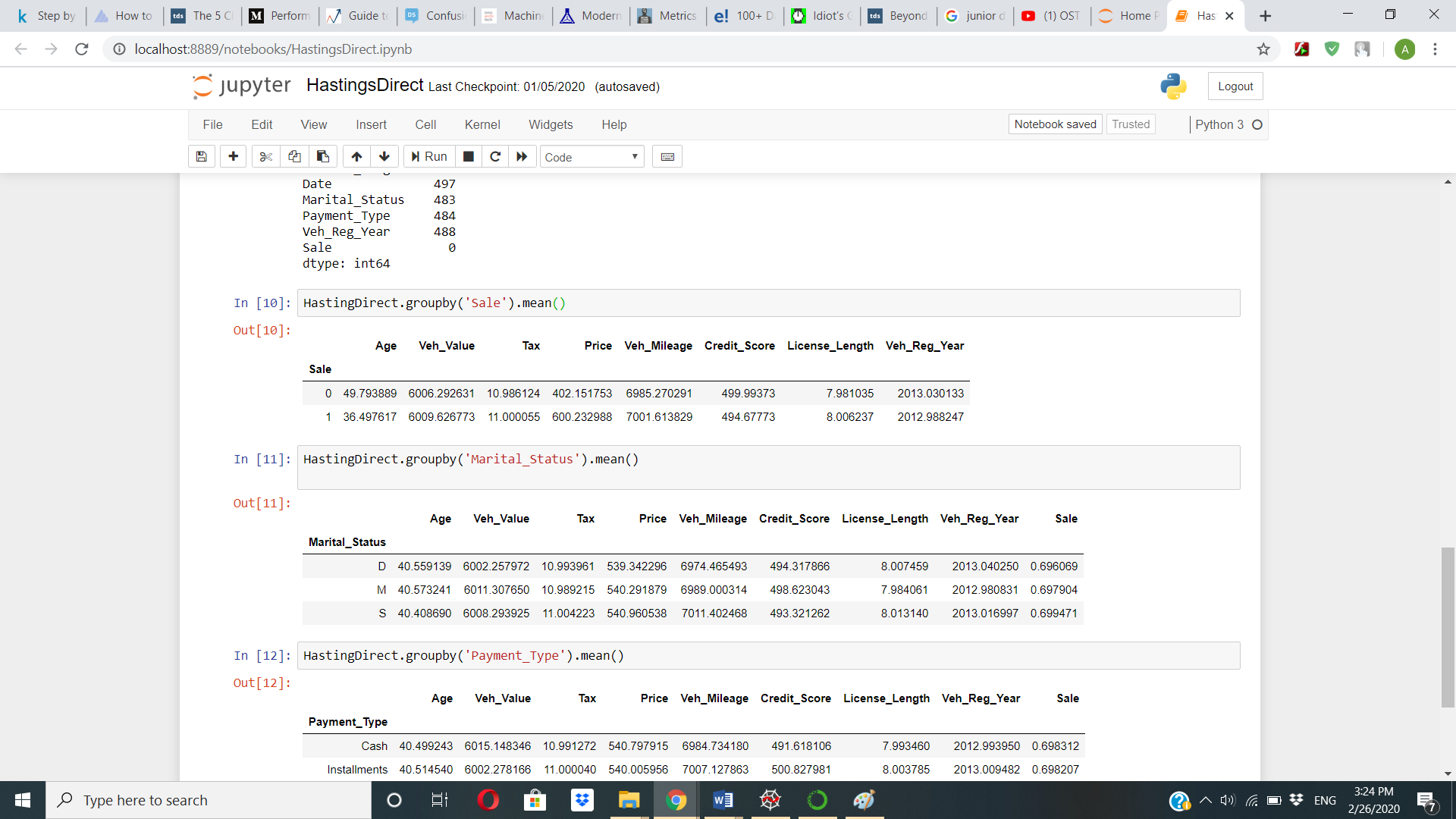
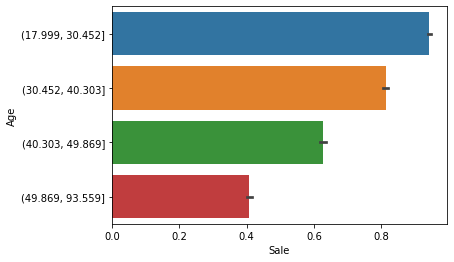


Figure 5: Sale summary using descriptive data mean

Figure 5 shows some interesting insights of the available data. People who did not buy the policy has an average age of 50 while younger people tend to buy the insurance policy even with high price which is opposite of what we analyzed in the older age group. *Price* shows positive correlation with *Sale* proving that as the policy price increase *Sales* goes up as well.



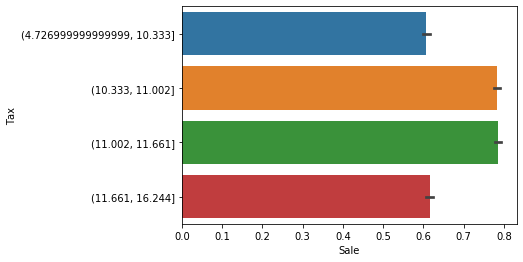


Figure 6: Numeric features vs Sale

*Age* and *Price* have the most effect on our target variable as evident from the figures above, followed by *Vehicle Mileage* and *Tax*.

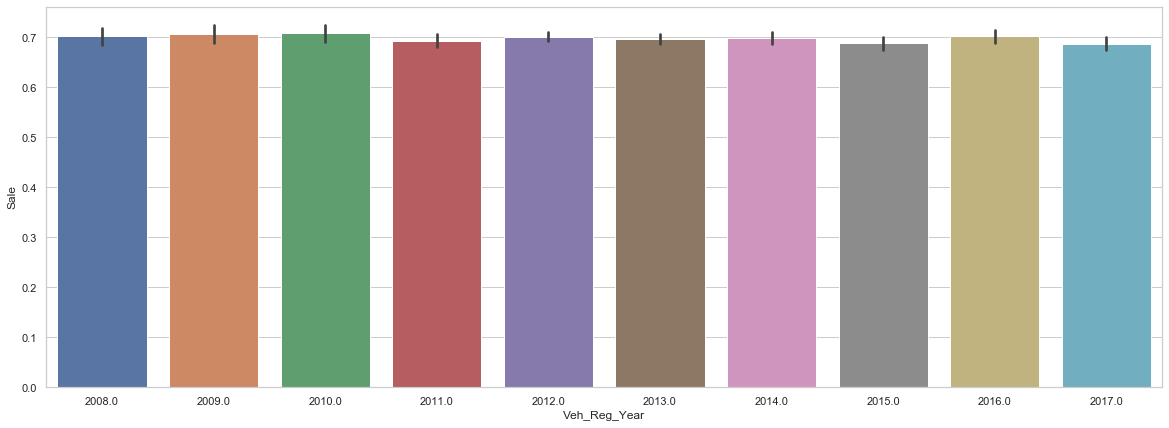


Figure 7: Vehicle Registration Year vs Sale

*Sale* vs *Vehicle Registration Year* shows that people with cars from 2011, 2015, and 2017 seem least interested in getting insurance, however, the difference is quite small.

Bar plot technique was used to show the relation between categorical data and target variable to identify if any *Marital Status* or *Payment Type* has any type of effect on *Sales*

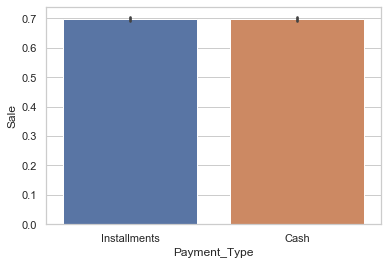
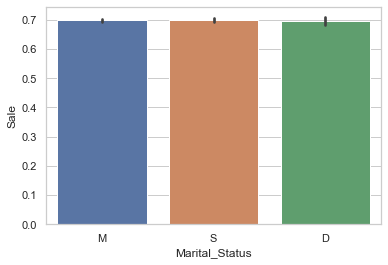


Figure 8: Marital Staus and Payment Type impact on Sale

Neither the *Marital Status* not the *Payment Type* had any effect on the *Sale* of the policy as the probability of a person who is buying the policy is the same as the probability of a person who is single or divorced. Also, 50% of the people bought the policy on *Installments* and the rest bought the policy on *Cash*.

However, an interesting trend was observed in the data as given below:

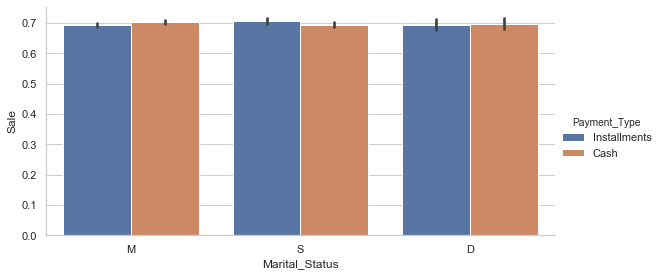


Figure 9: Martial Status, Payment Type vs Sale

*Married* people preferred to buy the policy on *cash* which is quite different from the people who were *single* and preferred *Installment* method. Whereas, most of the *divorced* people went for *Cash* option as well. Though the difference is not huge it can still be a crucial information for business.

Another comparison of *Marital Status* and *Sale* showed no major observation.

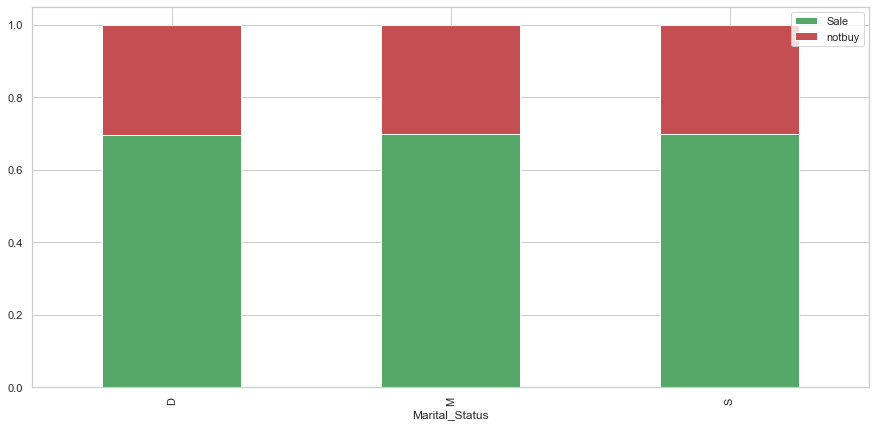


Figure 10: Marital Status vs Sale

Our initial analysis showed that *Payment Type* and *Marital Status* have little to no correlation with *Sale*, however *Age* and *Price* can affect the dependent variable.

A heatmap *Figure 9* was used to show the correlation of each variable with the dependent class as well as the independent classes.

*Tax* showed some correlation with *Vehicle* *Mileage* and *Vehicle* *Value*. *Age* showed negative correlation with *Sale*, indicating that older people rarely buys the policy. *Price* showed the highest correlation with *Sale*.

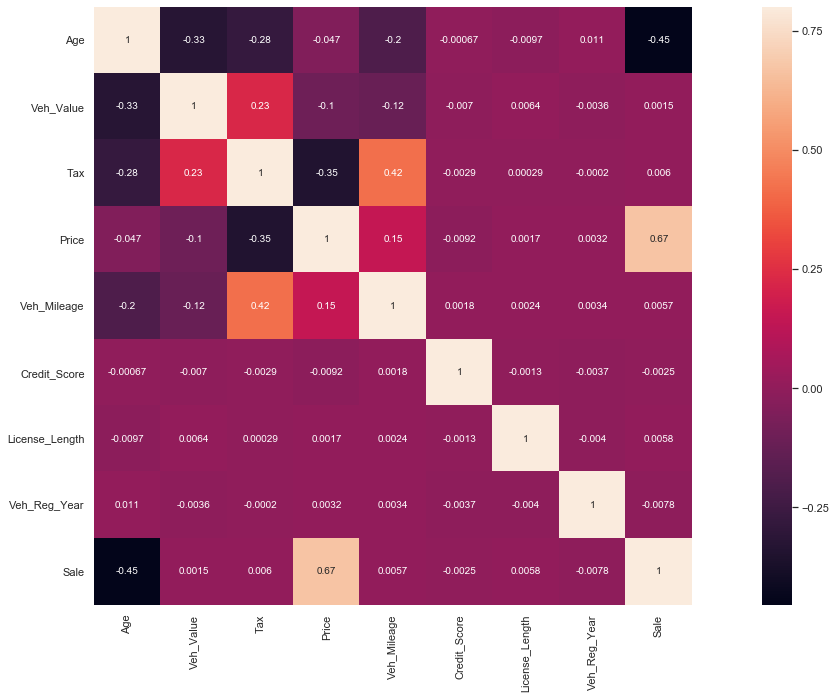


Figure 11: Features correlation map

# **Model Building and Results**

First, we need to identify the type of problem we are dealing with. For instance, does it need a supervised model or unsupervised model? Also, if it is a supervised problem whether it is a classification problem or regression problem.

Supervised problem, is the type of system in which we are given both input and output of the data. And using machine learning algorithms we map this input data to the output for future data processing and predictions. Unsupervised Problems is a technique which does not need input or output data, it helps you find patterns in data through Clustering and Association.

Our dataset has input data of 12 features which we need to map on our target class to yield results that can bring in high sale and income. As our class type is binary in type we need some Classification technique to train our mode.

Another significantly important step is choosing the right Machine Learning model to apply. There are several models available, each with its own pros and cons. The machine learning problem chosen for this problem is ‘**Logistic Regression’**. Even though it starts with logistic and ends with ‘Regression’, it is not really a model for regression problem, but deals with binary classification.

Logistic Regression is similar to Linear Regression in terms that both uses the equation of line in their model. But to recall the major difference, Logistic Regression uses a sigmoid function and thus, produces an S shape curve that takes care of outliers.



Figure 12: Logistic Regression

Logistic Regression was chosen among other models because of the underlying data structure. The available data is quite simple with symmetric distribution and we do not really need more complex models like Random Forest and SVM.

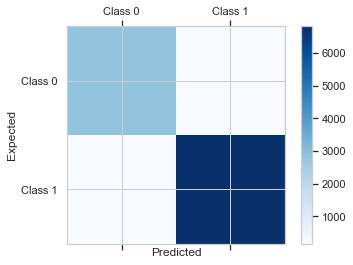
## Results

After standardizing our variables, it was ready to be tested by our model. For training data model showed the accuracy of **96.77%** and for test data model showed the accuracy of **96.42%.**

## Evaluation Metrics:

As described in detail in Section 1, we used three Evaluation metrics to test the performance of our model.

## Confusion Matrix:



6813

189

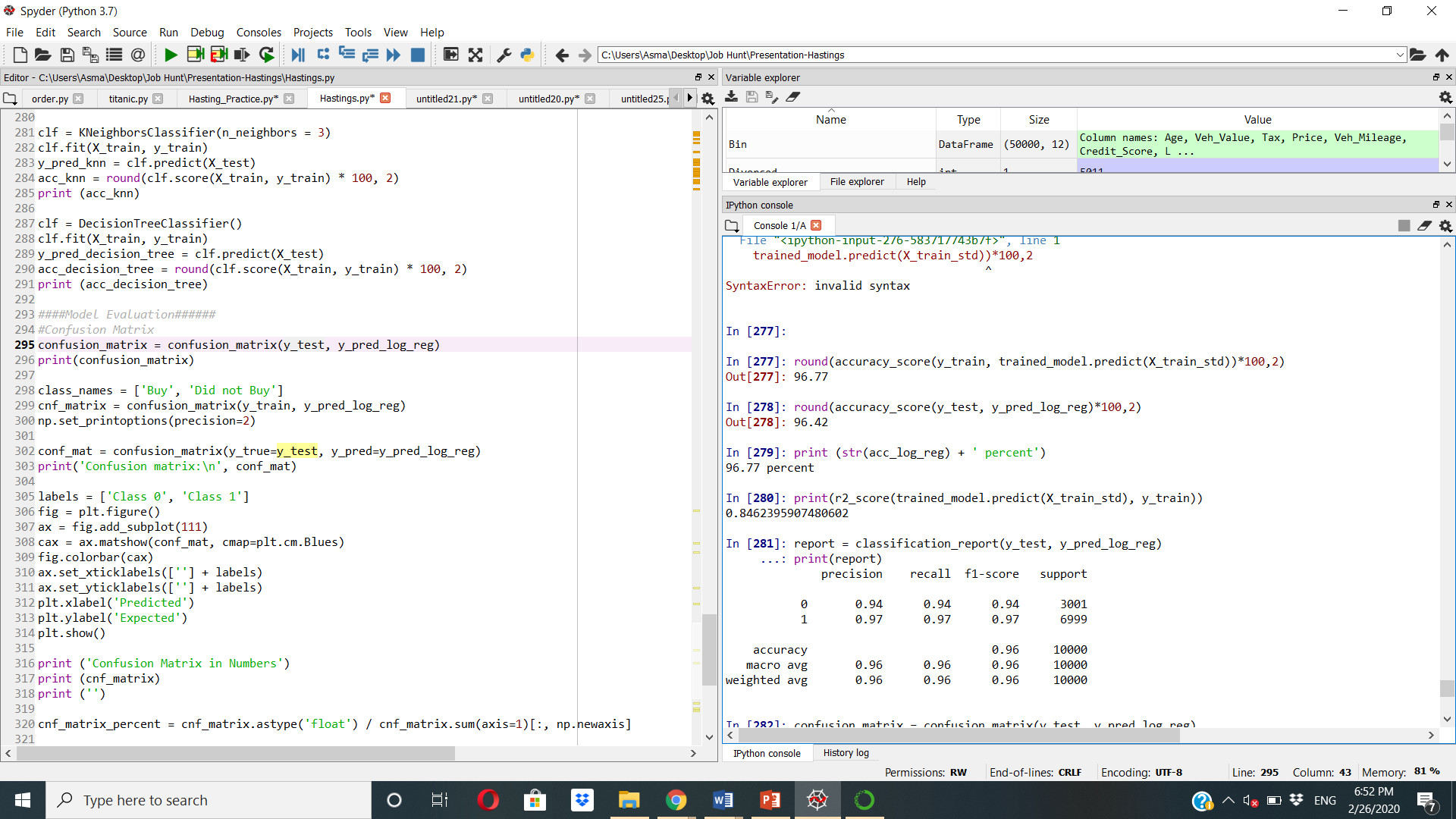
161

2837

Figure 13: Confusion Matrix

94% of the values were classified as True Negative, our model predicted customer will not buy the policy and the prediction was correct. 98% of the values that were classified as True Positive where our prediction that customer will buy the policy was true.

## Precision Recall F1 score



As described in detail in Section 1, our focus is to find the optimal F1 score along with high precision and high accuracy. Our models show the most optimal results with high precision along with high recall and F1 Score.

## AUC ROC

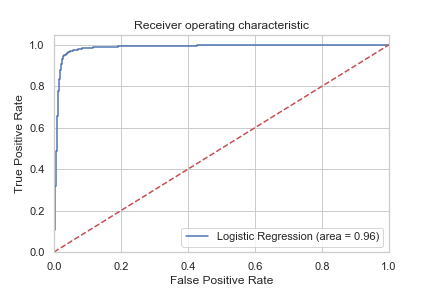


Figure 14: Area under curve ROC

ROC shows a high area under the curve of 0.96. This shows our model can easily describe the difference between our Target variable.

# **Insights**

Undoubtedly, *Age* and *Price*are the main features of the dataset, which can facilitate Hastings Direct in making the right company decisions and policies. For instance, a marketing manager can decide to steer his investments to target a certain group of people. As depicted by data, he may decide to target the customers with age band of 20-50, as they are the one more likely to buy the insurance even if the price is high.  This may result in better return on investment (ROI), focusing less on people, whose age lies between 50 and 90.

Similarly, an offering manager can look onto the analysis results and decide to offer premium services (e.g., include more insurance covers) to people, even if that means higher prices, because as evident from our analysis, lower price policy does not bring in much sales.  
Another interesting fact in the analysis was the payment type which people of different marital status opted for. Although the difference wasn’t huge, however, it can still impact the business plans. As married people prefer to pay in cash and single people prefer installments, it can be valuable to give better installment plans to single people and better cash plans to both married and divorced customers.

It would be beneficial for the company if further variables could be added to the already existing ones, such as, driving history, points on license, license type, no-claim years, etc., as it can directly impact the sales and revenue. Apart from this, further information regarding the car make and car price can be helpful for in-depth analysis.