

**FLIGHT PRICE PREDICTION**

Submitted by:

Albin Reji

# ACKNOWLEDGMENT

I would like to thank Sapna Verma, who’s the guide for the internship phase at FLIP ROBO, and the Dr. Deepika Sharma from DataTrained who conducted the training in the institute.

References:

<https://stackoverflow.com/>

<https://medium.com/analytics-vidhya/>

<https://www.youtube.com/user/krishnaik06/>

# INTRODUCTION

## Description

Anyone who has booked a flight ticket knows how unexpectedly the prices vary. The cheapest available ticket on a given flight gets more and less expensive over time. This usually happens as an attempt to maximize revenue based on - 1. Time of purchase patterns (making sure last-minute purchases are expensive) 2. Keeping the flight as full as they want it (raising prices on a flight which is filling up in order to reduce sales and hold back inventory for those expensive last-minute expensive purchases)

## Conceptual Background

Knowledge of real-life real-estate companies would help to better understand the different variables given in the dataset. It will in turn help us to leverage the correlation of the variables.

## Review of literature

I did research <https://www.magicbricks.com/> to find the prospect factors that would influence the prices of houses in an area. Even though the filters available in the commercial website it limited, we can definitely get an insight on how different factors affect the pricing. The important ones I noticed was the square ft area, and the neighbourhood. There might also be differences that must be taken into account when checking houses in the US and in India, but the differences are evident in the filters.

From the business perspective, the info is limited apart from the filters the website provides. There are no other detailed insights given by any competitor platforms.

## Motive

The motive behind this project is to create a ML model that predicts with the given conditions the price of used cars.

# ANALYTICAL PROBLEM FRAMING

## Web scraping

All the the data for the analysis was scraped from the cars24 website which I found to have an extensive collection of used cars details

* Most of the cars are previously used by just a single owner, there is a small number which was used by 4 owners
* All the cars in the dataset has no accident history
* Majority cars have manual transmission
* Most cars use petrol or diesel engines
* Most of the cars have 3rd party insurance, there care cars with zero depreciation and comprehensive insurance coverage



## Hardware and Software Requirements and Tools Used

* Intel i7 – 10705H
* 8 GB DDR4
* 1TB SSD
* Anaconda – Jupyter Notebook
* Chrome

# MODEL(S) DEVELOPMENT AND EVALUATION

## Possible Problem-Solving Approach

The output variable was continuous in nature, so the problem had to be modelled using the Regression method.

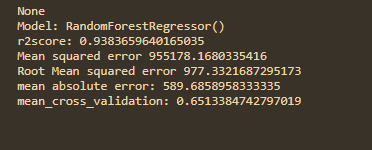
## Testing of Identified Approaches (Algorithms)

Models:

* Linear Regression
* K-Neighbors Regressor
* Decision Tree Regressor
* SVR
* Random Forest Regressor
* Lasso
* Ridge
* Elastic Net

## Run and Evaluate selected models

After running all the models, Ridge turned out to be the best model.



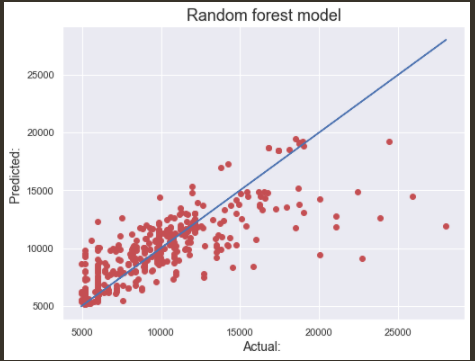
Running the Cross validation with the parameters set as shown below, and re-running the model with the best parameters, we get a 89.2% accuracy.

# Key Metrics for success in solving problem under consideration

* alpha = 0.1
* fit\_intercept = False

# Visualizations

The plot below shows the code and the fitting of the model. We can conclude that the model is not over/under fit, and works as it says it does.



## Interpretation of the Results

We’re getting a model with about 91% accuracy post Cross Validation and Grid Search CV with all listed parameters.

# CONCLUSION

## Key findings and Conclusion

Many models gave a good R2 Score in the preliminary run, but on further inspection, random forest turned out to be the best working model. On further running with the Cross validation, we were able to find that the difference between R2 Score and CV score were less for all the models, however, I went with ridge taking into consideration the low difference in values for all the models and the comparatively higher score for the Ridge Model.

## Learning Outcomes of the Study in respect of Data Science

I have created a model which gives me a good accuracy in terms of predicting the prospect price of a house with the parameters listed.

## Limitations of this work and Scope for Future Work

I feel that the libraries which we run are limited in a way that they don’t have an option for parallel processing.

Some libraries like TensorFlow-GPU, H2O4GPU (Scikit replacement) etc utilizes the GPU which has more cores and help run the models a lot faster while not compromising on the accuracy.

If the common Classification models that we use come up with a replacement that utilizes the GPU, we’ll be more comfortable running those models on personal laptops.