PLAGIARISM SCAN REPORT

Date	December 05, 2021
Exclude URL:	NO

	Unique Content	92%	Word Count	1,124
	Plagiarized Content	8%	Records Found	5
	Paraphrased Plagiarism	0		

CONTENT CHECKED FOR PLAGIARISM:

Analyzing the accelerating sales of used cars in a digitally disrupted market Mayuravarsha P Computer Science & Engineering PES UniversityBengaluru, Indiamayurvp72@gmail. com Vedant Mantri Computer Science & Engineering PES University Bangalore, Indiavedantmantri3@gmail.com Pranav Mekal Mahesh Computer Science & Engineering PES University Bangalore, India pranavmm25@gmail.com Abstract— Used-car selling is being disrupted by the internet revolution, and it's for the better. This new era of digital commerce is more than just about technology; it emphasises the importance of the consumer experience in the used-car purchase process. Introduction Online suppliers are beginning to erode traditional used-car dealers' standing and promote growth by empowering digitally savvy customers via three primary capabilities, as indicated by our unique customer research: complete end-to-end purchasing capabilities extensive vehicle data and photos, along with effective search tools unique delivery options With the development of digital players and the possibility of incumbent-dealer consolidation on the horizon, the developing market will present new dangers and opportunities for firms looking to gain a competitive advantage in an already crowded Furthermore, while customer purchasing habits are evolving, it is true that the needs of used-car buyers differ significantly from those of new-car buyers. As a result, all used-car merchants must identify their target client categories and quickly design the best ways among a growing number of accessible options in order to provide a uniformly distinctive and distinguishing customer experience. Handling The Dataset Dataset right632460 00 Our dataset is a used cars dataset chosen from Kaggle. It contains most all relevant information that Craigslist provides on car sales including columns like price, condition,

manufacturer, latitude/longitude, and 18 other categories. The link for the same can be found below in references. Preprocessing or Data Cleaning: The dataset that we chose was of the size of over 1GB and that means that we have too many redundant columns and data in the dataset and these were removed and dealt with accordingly. The attributes with NULL values we either simply dropped or dealt with using various basic techniques. Exploratory Data Analysis: The visual summary of the data makes it easier to identify patterns and trends than looking through thousands of rows and columns on a spreadsheet. Thus, we plotting several data plots and visualization for the better understanding of the data. The different plots that we have used are Histograms, Box plots, Count plots and Bar Graphs. Before we plot any graphs, we found that the dataset is huge and might we may face some hitch while doing visualization, so we thought of splitting the dataset and did it randomly. 22034596520 00 245745354965 00 left2319020 00 left185420 00 90805185420 00 781052661920 00 Existing Appoaches And Their Shortcoming Linear Regression This model does not seem to work well at all as we found results where the RMSE (Root Mean Squared Error) was in the power of 19 and such models are extremely unreliable as they overfit to a huge extent and provide no value in predicting or analyzing. This model assumes that there are only linear relations in the dataset which is a huge gamble and mostly isn't necessarily true and can lead to being extremely costly as mentioned before. Gradient Boosting Does relatively well compared to Linear Regression. 99695114300 00 Data Modelling We have built multiple models to predict the likeliness of a used car getting sold based on its various attributes. We followed the standard procedure of cleaning and preprocessing the data before proceeding with model testing. We have used correlation coefficient and root mean square error as the defining basis of which models have performed the best and the worst with the test data. StratifiedKFold from sklearn module has been used to index the training and testing data. The models used are as follows: Decision Tree Baseline Performs far better than Linear Regression but still doesn't belong anywhere near other machine learning models which have been used by us for better efficiency. The drawbacks of decision trees are pruning and overfitting. There's a high chance of pruning as in most cases we are going to come across outliers and Decision Trees tend to classify each of the data in the training dataset perfectly and hence not only overfitting but also making the length of the tree unnecessarily longer and costlier. XGBoost Baseline XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable. It implements machine learning algorithms under the Gradient Descent framework. This performs significantly better than Gradient boosting. XGBoost with Parameters The overall parameters have been divided into 3 categories by XGBoost :General Parameters: Guide the

overall functioning Booster Parameters: Guide the individual booster (tree/regression) at each step Learning Task Parameters: Guide the optimization performed This performs significantly better than XGBoostRandom forest baseline Random Forest is an ensemble learning algorithm that constructs many decision trees during the It predicts the mode of the classes for classification tasks and mean prediction of trees for regression tasks. It is using random subspace method and bagging during tree construction. It has built-in feature importance. Since it uses ensemble models, it is highly accurate and less error prone compared to previous models. Decision Tree HyperParam Tuning Hyperparameter tuning is searching the hyperparameter space for a set of values that will optimize your model architecture. Due to this property, it performs way better than a Decision Tree and predicts more accurately. Other Efficient Models LightGBM with parameters MLP Regressor with parameter tuning LightGBM with categories & parameters All of the above-mentioned models are the most efficient and helpful when it comes to analyzing and the proof for the same can be seen below with their r2 and RMSE values. Comparing all the algorithms: 2749551803400 400000 center182880 4000020000 Conclusion, Prediction and Analysis By using an MLP model we saved this model while choosing RELU as the activation function as it performs extremely well in real world examples. right384810 00 The MLP model is saved for future insights on the same The predictions made for the same can be seen below 95253249930 4000020000 Acknowledgment We would first like to thank our professor Dr. Gowri Srinivasa and her Teaching Assistants for guiding us throughout the course of this project. This work was supported by grants from the Computer Science and Engineering department in PES University, Bengaluru. Everyone from our team have made contributions in all aspects including Data segmentation, Data visualization, Segmentation and Data modelling and analyzing. References https://www.mckinsey.com/~/media/McKinsey/Industries/Automotive and Assembly/Our Insights/Used cars new platforms Accelerating sales in a digitally disrupted market/Used-cars-new-platforms-Accelerating-sales-in-a-digitally-disrupted-market-vF.pdf U. Dinesh Kumar - Business Analytics_ The Science of Data-driven Decision Making Wilkinson, Christopher- Python data science: an ultimate guide for beginners to learn fundamentals of data science using Python Smith, James, V- Data Analytics: What Every Business Must Know About Big Data And Data Science https://www.kaggle.com/austinreese/craigslist-carstrucks-data

MATCHED SOURCES:

towardsdatascience.com - *2% Similar*Compare

https://towardsdatascience.com/lightgbm-vs-xgboost-which-alg....

---

www.op.salesperformance.com - *2% Similar*Compare

https://www.op.salesperformance.com/goqplrs/team-gymnastics-....

---

medium.com - *1% Similar*Compare

https://medium.com/coinmonks/predicting-a-startups-profit-su....

---

mljar.com - *<1>Compare*

https://mljar.com/machine-learning/catboost-vs-random-forest....
