Authors: Imangaliyev Daulet (IT-2104), Biakhmet Daulet (IT-2101)

Idea proposal: car prediction for Kazakhstan’s market

Problem: There is no tool to evaluate car’s price in Kazakhstan

Background Information: Kolesa.kz is the biggest auto marketplace in Kazakhstan.

The high prices of cars in our hometown make it difficult for consumers to make informed purchasing decisions. The goal of this project is to develop a machine learning model that can accurately predict the prices of cars based on a variety of factors, such as make, model, and year. By providing consumers with this information, they will be able to make more informed purchasing decisions and potentially save money. In the car industry, determining the right price for a car is crucial. Many factors such as make, model, year, and condition affect the price of a car. With the rise of technology, it has become important for car dealerships to use advanced methods for pricing their cars. This is where machine learning comes in. It can be used to analyze large amounts of data and make accurate predictions about car prices. There are several existing solutions for predicting car prices, such as Kelley Blue Book and Edmunds. However, these solutions rely on a limited set of data and may not take into account all of the relevant factors that affect car prices. Additionally, these solutions are not always accessible to consumers. Our proposed solution aims to improve upon existing solutions by using a wider range of data and machine learning techniques to make more accurate predictions.

The data for this project will be collected from multiple sources such as Kolesa.kz. The data will include information such as make, model, year, and condition of the cars. The data will be cleaned and preprocessed before being used to train the machine learning model. The solution for this project will be a machine learning model that can predict the prices of cars based on a variety of factors. The model will be trained on the collected data using techniques such as linear regression and decision trees. The model will be tested and evaluated using metrics such as mean absolute error and R-squared.

The technologies that will be used in this project include Python, NumPy, Pandas, Scikit-learn, and TensorFlow. These tools will be used for data cleaning and preprocessing, training and evaluating the machine learning model. In conclusion, this project aims to develop a machine learning model that can accurately predict the prices of cars based on a variety of factors. By providing consumers with this information, they will be able to make more informed purchasing decisions and potentially save money. The proposed solution is expected to improve upon existing solutions by using a wider range of data and machine learning techniques

Structure:

○ Introduction:

■ Problem

■ Literature review with links (another solutions)

■ Current work (description of the work)

○ Data and Methods

■ Information about the data (probably analysis of the data with some visualisations)

■ Description of the ML models you used with some theory

○ Results

■ Results with tables, pictures and interesting numbers

○ Discussion

■ Critical review of results

■ Next steps

Problem: Kolesa.kz has no transparent tool for evaluate your car price, while there are so much different parameters, which user should understand, but without experience, seller could evaluate in a wrong side, which will harm both sellers and market



We know difference from mean value, but why is it so, is not clear.

**Literature review with links**

* [**https://www.autotrader.co.uk/cars/valuation**](https://www.autotrader.co.uk/cars/valuation)
* **https://www.carwale.com/used/carvaluation/**

**Изображение выглядит как текст

Автоматически созданное описание**

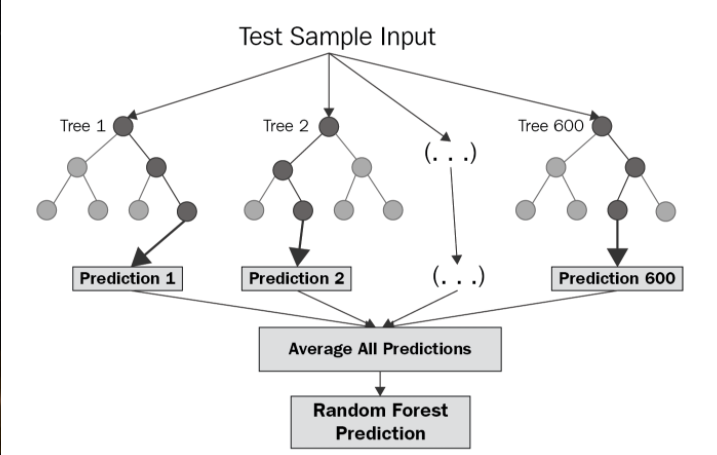
**Изображение выглядит как текст, внутренний, снимок экрана

Автоматически созданное описание**

These services also work not transparent and not adapted to Kazakhstan’s market.

**Current work (description of the work)**

Car price predictor parses data from kolesa.kz, it const of random forest



All predictions are made after making trees from previous data and each prediction

Random Forest Regression: What is it?

A type of machine learning methods known as "Random Forest Regression" combines many random decision trees, each of which has been trained on a subset of data. The method is more stable and less erratic when many trees are used. The random forest regression approach, which performs well for big and most types of data, is a frequently used model.

Each tree is produced by the algorithm using a unique sample of input data. A separate sample of characteristics is chosen for splitting at each node, and the trees proceed independently of one another. A single outcome, the forecast of the Random Forest, is produced by averaging the predictions from each of the trees.

After set of experiments we can out group explored efficiency of different algorithms, where Random Forest and XGBoost performed the best, but we can see that Random Forest has higher accuracy.

**Information about the data (probably analysis of the data with some visualisations)**

Изображение выглядит как текст

Автоматически созданное описание

We made special parser, in out training set we have 371 documents, while all of them are different cars, we could not physically store all data from kolesa.kz, which implemented restrictions on several spheres of our solution.

As we can see, we use MongoDB, database to store all data for training. We use special data schema for data validation.

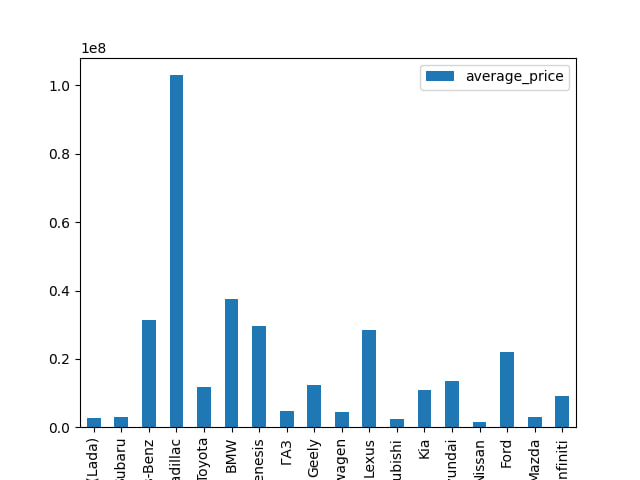
Изображение выглядит как текст

Автоматически созданное описание Изображение выглядит как текст

Автоматически созданное описание

We can see defined data types for all values

Here you can see the vizualizations of data



**Description of the ML models you used with some theory.**

As we said, we had limited storage space for train data, so we made experiments, and we discovered, that we can noticeably reduce loss and error, by fitting around 20 new examples, that are look-a-like of the example that we see.

Many decision trees are built during the training phase of the random forests or random decision forests ensemble learning approach, which is used for classification, regression, and other tasks. The class that most of the trees choose is the output of the random forest for classification problems. The mean or average forecast of each individual tree is returned for regression tasks. The tendency of decision trees to overfit their training set is corrected by random decision forests. Although they frequently outperform decision trees, gradient enhanced trees are more accurate than random forests. Yet, their effectiveness may be impacted by data peculiarities.

In 1995, Tin Kam Ho developed the first algorithm for random decision forests using the random subspace technique. Ho's formulation offers a way to put Eugene Kleinberg's "stochastic discrimination" approach to classification into practice.

Leo Breiman and Adele Cutler created an adaptation of the method, and in 2006 they filed "Random Forests" as a trademark (as of 2019, owned by Minitab, Inc.). The extension builds a set of decision trees with controlled variance by combining Breiman's "bagging" concept with randomly choosing features, which were independently presented by Ho, Amit, and Geman after Ho.

**Results with tables, pictures and interesting numbers**

Interesting fact, if you do around 10 requests on Kolesa.kz using foreign IP address, you will get restricted access to site this is why we use sleep(1) before requesting

Изображение выглядит как текст

Автоматически созданное описание

**Discussion**

Next Steps

We want to implement all modern software engineering paradigms, and want to make this service accessible to all.

Our current r2 rate ~91%, we want to reach level of 96% at least

Also in future we want to make web-service from it using Django.