哈尔滨工程大学专业选修课研究报告

面向二手车买卖的价格合理性智能判定系统

专　业　名　称：计算机科学与技术

学　生　姓　名：马骁

哈尔滨工程大学

2024年10月

摘　　要

随着经济的快速发展和消费者购车需求的增加，二手车市场的地位日益重要。相比于新车，二手车不仅能为消费者提供经济实惠的出行选择，还有助于资源的循环利用和环保。然而，由于二手车的复杂性，如车龄、行驶里程、保养记录和事故历史等，传统的人工评估方法在准确性和透明度上存在不足，常导致价格不合理和交易不公。 为了解决这一问题，本项目设计并开发了一套基于Python的二手车价格合理性智能判定系统。该系统利用数据分析与机器学习技术，通过对用户提供的车辆参数进行系统的预处理、特征提取和选择，以及机器学习模型的构建与优化，科学预测二手车的合理价格区间。系统的核心技术包括数据清洗、特征工程、可视化分析和模型优化，旨在提高预测的准确性和可靠性。 项目的主要目标包括帮助买家获取科学的价格参考，以避免因信息不对称而受到不合理定价的影响；同时，帮助卖家合理定价，提升交易效率。通过引入机器学习技术，系统不仅能提高市场透明度，还能为行业提供技术创新与应用推广的示范。最终，本项目希望为二手车市场注入更多科技手段，使价格评估不再依赖于主观经验，而是基于数据驱动的科学分析，为买卖双方提供公平、透明的交易基础，推动二手车市场的健康发展。

关键词：二手车价格评估；数据分析；机器学习

目　　录

[第1章 应用背景 3](#_Toc357160353)

[1.1 项目背景 3](#_Toc357160354)

[1.2 当前无人售货机现状 3](#_Toc357160355)

[1.3 项目目标 4](#_Toc357160356)

[第2章 核心技术 6](#_Toc357160357)

[2.1 数据预处理 6](#_Toc357160358)

[2.2 特征工程与选择 7](#_Toc357160362)

[2.3 可视化分析 7](#_Toc357160366)

[2.4 机器学习建模 8](#_Toc357160370)

[2.5 模型优化 9](#_Toc357160370)

[2.6 系统实现与部署 9](#_Toc357160370)

[第3章 项目结构与流程可视化 1](#_Toc357160371)1

[3.1 项目架构图 1](#_Toc357160372)1

[3.2 项目流程图 1](#_Toc357160373)4

[3.3 功能结构图 1](#_Toc357160379)6

[第4章 功能模块分析和代码介绍 1](#_Toc357160371)9

[4.1 功能模块 2](#_Toc357160372)0

[4.2 功能模块的交互与协同 2](#_Toc357160373)2

[4.3 代码分析 2](#_Toc357160373)3

[第5章 软件功能演示 3](#_Toc357160371)3

# 第1章 应用背景

## 1.1 项目背景

在经济发展日新月异、消费者购车需求逐渐增加的今天，二手车市场的地位日益重要。二手车的购买不仅能够为消费者提供更加经济实惠的出行选择，还为资源的循环利用和环保事业贡献了一份力量。然而，由于二手车本身的复杂性，如车龄、行驶里程、保养记录、事故历史等，价格的评估也因此变得十分复杂。传统的估价方法大多依赖于人工经验或简单的市场均值评估，这样的方式往往缺乏数据支持，在面对个体车辆多样化的情况时，往往显得力不从心。因此，为了能够帮助消费者更加科学地理解二手车的价格合理性，提供一套基于数据分析和机器学习的二手车价格评估系统显得尤为必要。 此外，随着信息技术的快速发展，数据分析与机器学习等技术在各个行业中的应用不断深化，二手车价格评估也成为了这些新兴技术的重要应用场景。通过挖掘大量的市场数据，建立有效的模型，结合用户提供的车辆信息，科学地预测二手车的合理价格区间，可以帮助用户更加透明、全面地了解市场行情，降低交易过程中的不确定性，提高买卖双方的满意度。

## 1.2 当前二手车市场现状

二手车市场在全球范围内，尤其是中国等新兴经济体，近年来呈现出了迅猛的发展势头。根据相关数据显示，中国二手车交易量近年来以每年两位数的速度增长，成为了汽车产业链中不可忽视的一部分。与新车相比，二手车市场的价格波动大，主要原因在于车辆的使用状况、市场需求变化和车主心理预期等多种因素的共同影响。

车辆的状况直接影响其价值。包括车辆的使用年限、行驶里程、保养和维修记录等信息，这些都将直接决定车辆在二手市场上的价格。其次，市场的供需关系也是影响价格的关键因素。不同车型的市场需求量不同，例如一些热门品牌和型号的二手车可能会由于市场热度而维持较高的价格，而冷门车型的价格则可能较低。此外，环保政策的变化、油价波动、季节性需求变化等外部因素也会影响市场对二手车的需求，从而引起价格波动。

信息不对称是二手车市场面临的另一大挑战。买家和卖家之间对于车辆状况的了解程度差异显著，卖家可能因为更了解车辆的真实状况而开出不合理的高价，而买家则由于对车辆细节的不完全了解，面临价格不透明的困扰。这种信息不对称导致市场中常常出现交易不公正的情况，影响市场的活力和买卖双方的信任。因此，如何通过技术手段解决二手车市场中的信息不对称问题，提高交易的公平性和透明度，成为了行业亟待解决的问题。

## 1.3 项目目标

基于当前二手车市场面临的挑战，本项目的目标是设计并开发一套基于Python的二手车价格评估系统，利用数据分析与机器学习技术，对用户提供的车辆参数进行合理的价格区间估算。具体来说，本项目有几个主要目标：

**1）帮助买家评估合理价格**：

通过科学的价格评估模型，买家可以获得车辆的合理价格区间，避免因信息不对称而被动接受高于市场的售价，从而在购车过程中有更好的谈判能力和决策基础。

1. **帮助卖家设定合理售价**：

卖家可以根据系统的建议，合理定价，既避免了过高定价导致车辆滞销的情况，也防止因低估价格而造成经济损失。通过对车辆状况和市场数据的综合分析，卖家可以快速了解车辆在当前市场中的最佳售价区间，从而提高交易效率。

**3）提高市场透明度和交易效率**：

通过将数据分析和机器学习引入二手车价格评估，本项目致力于提高二手车市场的透明度，减少交易中的不公平现象。这不仅能够提高交易效率，还能增加买卖双方的信任度，推动整个市场的健康发展。

**4）技术创新与应用推广**：

本项目还旨在通过Python编程技术的应用，探索数据科学在二手车价格预测领域的更多可能性，积累相关的经验和方法，为未来的数据分析应用奠定基础。这套价格评估系统不仅可以用于买卖双方，也可以推广到车商、保险公司等需要对二手车进行估价的场景中，具有广泛的应用潜力。

通过以上目标的实现，本项目希望为二手车交易市场注入更多的科技手段，让价格评估不再依赖于经验和主观判断，而是基于科学的分析和市场的真实数据，为买卖双方提供公平、透明的价格参考，推动整个二手车市场的良性循环和发展。

# 第2章 核心技术

本项目的核心技术涵盖了数据预处理、特征工程、可视化分析、机器学习建模以及模型优化等方面。具体技术实现如下：

## 2.1 数据预处理

数据预处理是确保模型质量的关键步骤，涉及对原始数据的清洗、缺失值处理和标准化。本项目选用的二手车数据集包含40万条交易记录，数据维度复杂，包括车辆品牌、型号、年份、行驶里程等26个字段。在预处理中，我们进行了以下处理：

1. 数据清洗：

原始数据集中包含一些无效字段（如county）和无用数据（如lat和long），这些字段对价格预测没有实际意义，因此在数据清洗过程中将其去除。此外，还针对数据集中可能存在的脏数据进行了筛除，例如价格字段中出现的负值记录，确保所有数据符合现实逻辑。

1. 缺失值处理：

数据集中部分字段存在缺失值，如size、condition、cylinders等字段。对于这些缺失值较多的字段，我们首先进行了分析，决定保留重要性较高的字段，而删除其他影响较小的字段。对于缺失值较少的字段，我们采用了均值填充或中位数填充的方法来补全数据，确保数据的完整性。

1. 数据脱敏：

为了保护用户的隐私，本项目对敏感信息（如VIN码）进行了脱敏处理，将其替换为随机值或直接覆盖，这样可以避免用户隐私泄露，同时确保数据在模型训练中的安全性。

1. 数据标准化：

为避免特征之间的量纲差异导致的模型偏差，对数据进行了标准化处理，以提升模型的鲁棒性和准确性。

2.2 特征工程与选择

为了提高模型的预测能力，本项目对原始数据进行了系统的特征工程，包括特征提取、特征选择和特征编码：

1. 特征提取：从数据集中提取了对车辆价格具有代表性的信息，例如车辆的年份、行驶里程、燃料类型、车辆状况、变速器类型等。这些特征经过分析被认为对二手车价格有显著影响，因此被选为模型的输入变量。
2. 特征选择：通过相关性分析和特征重要性分析，我们评估了各个特征与目标变量（价格）之间的相关性，筛选出最具影响力的特征，例如year、cylinders、condition等，剔除那些对预测无关紧要或噪声较大的特征。这一过程不仅简化了模型结构，还减少了数据维度，提高了计算效率。
3. 特征编码：由于类别型特征对价格预测有重要意义，为了使模型能够处理这些特征，我们使用了独热编码（One-Hot Encoding）和标签编码（Label Encoding）等方法将类别型特征转换为数值型特征。这些编码方式确保了模型在处理类别型数据时具有较高的准确性和泛化能力。

2.3可视化分析

在数据预处理之后，采用了可视化分析手段对数据集进行探索性数据分析（EDA），以直观地展示各个特征对价格的影响，帮助理解数据的内在关系：

使用缺失值矩阵图展示各个特征的缺失情况，帮助确定需要处理的字段。通过可视化分析，我们能够直观地看到哪些字段存在较多的缺失值，从而决定处理策略。对于价格分布分析，则利用直方图、箱型图对价格数据进行可视化，分析发现价格数据存在右偏，并且存在较多的极端离群值。为此，我们剔除了价格大于60000的离群值，并对价格数据进行了对数转换，以使数据分布更符合正态分布，这有助于提高模型的稳定性和预测能力。并通过绘制相关性热力图，我们能够直观地看到不同特征之间的相关性大小，特别是价格与其他特征之间的关系。通过相关性热力图，可以有效地选择对价格预测有贡献的特征，避免特征共线性对模型性能的负面影响。对于不同类别特征（如年份、车辆状况、燃料类型等）与价格的关系，我们使用箱型图和条形图进行了可视化。结果显示，车辆的年份、状况、汽缸数等因素对价格有显著影响，为后续的模型构建提供了有力支持。

2.4 机器学习建模

本项目采用了Scikit-Learn作为核心人工智能库，构建了多种机器学习模型对二手车价格进行建模，并对比了不同模型的效果。

1）Lasso回归：为了应对多重共线性问题并提高模型的解释性，我们采用了Lasso回归。Lasso回归通过引入L1正则化约束来选择最具代表性的特征，降低了过拟合的风险，同时保留了对价格影响较大的变量。

2）其他模型对比：除了Lasso回归，我们还尝试了其他回归算法（如线性回归、随机森林回归和支持向量机回归）来对比模型的性能。随机森林回归由于其处理复杂非线性关系的能力，表现出较好的预测效果。通过交叉验证对不同模型进行评估，最终选取了表现最优的模型用于价格预测。

3）模型训练与验证：将数据集划分为训练集和测试集，按照3:1的比例进行划分，以保证模型在训练和测试阶段都能得到充分的样本。通过交叉验证的方法评估模型的性能，确保模型具有较好的泛化能力和稳定性。

2.5 模型优化

在初步建模之后，对模型进行了进一步的优化，以提高预测精度和模型的泛化能力：

1）超参数调优：通过网格搜索和随机搜索对Lasso回归中的alpha参数进行调优，以找到最佳的模型参数组合。超参数调优的过程可以有效地提高模型的预测精度，避免因参数设置不当而导致的模型性能下降。

2）标准化与正则化：在模型训练中加入了数据标准化步骤，确保特征值的均一性，避免某些特征值范围过大导致的模型偏差。同时，通过正则化技术有效防止模型的过拟合问题，增强了模型的泛化能力。

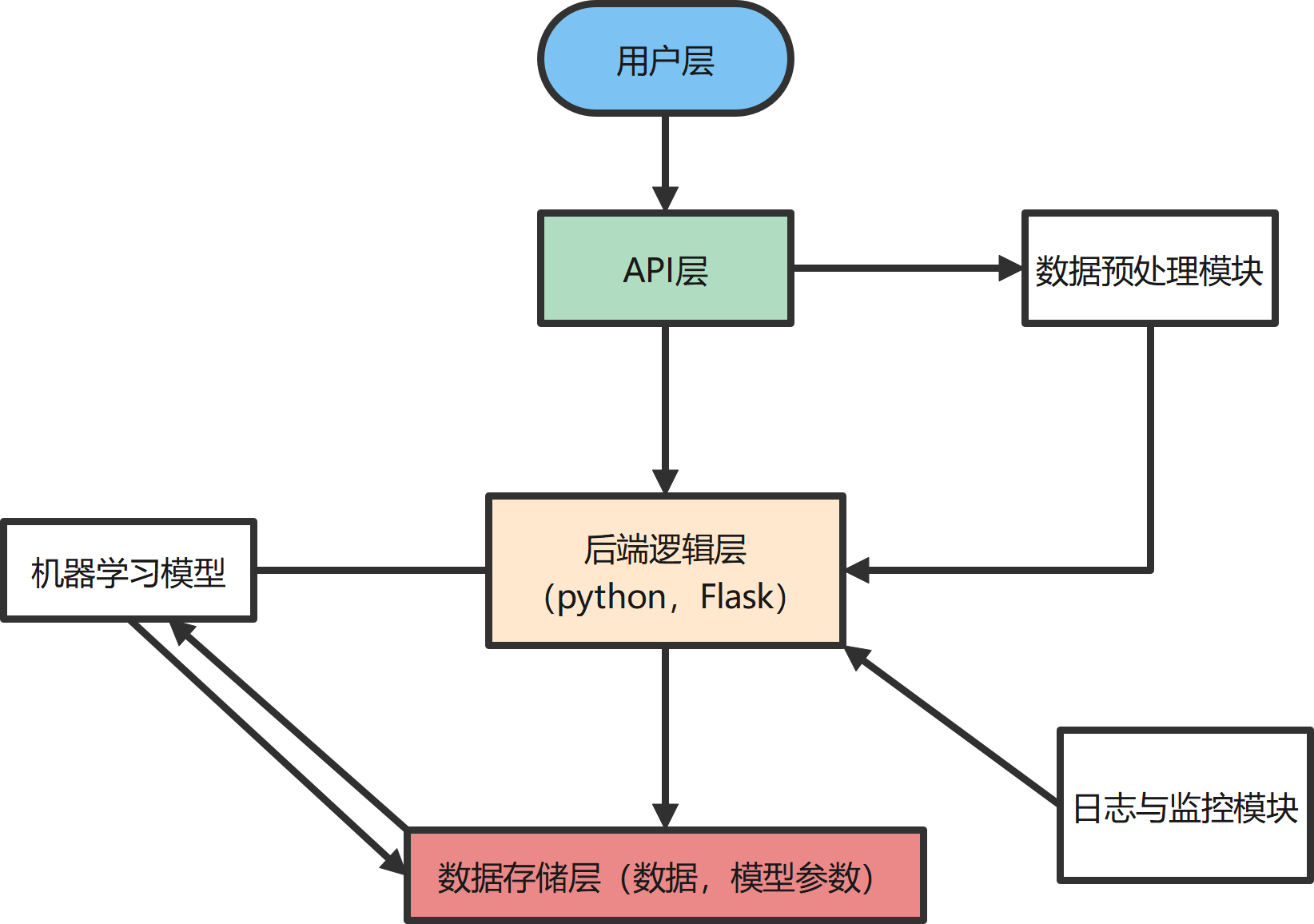
3）误差评估与残差分析：使用均方误差（MSE）、均方根误差（RMSE）和决定系数（R²）等指标来评估模型的表现，确保模型的高精度和稳健性。此外，通过残差图分析模型的拟合效果，以识别模型中可能存在的系统误差。

2.6 系统实现与部署

在完成模型训练和优化后，项目还开发了一个用户界面，以方便用户使用模型来进行二手车价格评估：系统前端采用Web技术进行开发，基于Python的Flask框架来搭建后端。通过用户友好的界面，用户可以输入车辆的相关信息，如年份、行驶里程、车辆状况等，系统会实时返回该车辆的合理价格区间。通过将模型封装为RESTful API，前端系统可以调用模型接口，实现价格预测功能的实时性和便利性。API的集成使得系统具有良好的扩展性，未来可以方便地进行功能扩展或与其他系统集成。为了保证用户数据的安全性，系统对用户输入的数据进行了严格的数据保护措施，包括数据加密和隐私信息的脱敏处理，确保用户的隐私不会在系统使用过程中泄露。

第3章 项目结构与流程可视化

## 3.1 系项目架构图



该项目架构图展示了整个二手车价格评估系统中各个部分之间的关系，以及它们如何协同工作来实现最终的目标。架构图包含以下几个主要部分：

1)用户层:

用户通过浏览器或移动设备与系统进行交互，输入车辆的参数信息（例如品牌、型号、年份、行驶里程等）。这些信息将通过前端界面发送到系统的后端进行处理。

1. API层:

API层用于管理用户请求与后端逻辑之间的通信。在用户输入车辆信息之后，这些数据通过API层传递到后端逻辑模块。API层确保数据在前端与后端之间的安全传输，并负责处理请求的路由。

3）后端逻辑层：

后端逻辑是整个系统的核心，由Python和Flask框架构建。后端逻辑层负责处理从API层接收到的用户输入数据，执行必要的操作，如数据预处理、价格预测等。该模块还会与其他模块（例如数据预处理模块、机器学习模型模块）进行交互，协调整个系统的运行。

4）数据预处理模块：

数据预处理模块是系统处理用户输入数据的重要环节。该模块对输入数据进行清洗、编码、标准化等操作，确保数据能够被机器学习模型正确理解和处理。该模块与后端逻辑层紧密配合，确保输入数据的质量和一致性。

5）机器学习模型模块：

机器学习模型模块用于预测二手车的合理价格。经过预处理的数据被输入到机器学习模型中进行计算。模型基于之前训练的数据对输入信息进行预测，最终输出价格估算结果。该模块和数据存储层也有交互，用于读取已训练的模型参数和进行更新。

6）数据存储层：

数据存储层用于存储系统需要的数据和模型，包括用户输入的历史数据、用于模型训练的训练数据集以及训练后的模型文件。数据存储层支持后端逻辑和机器学习模块的调用，确保模型可以持续进行改进和更新。

7）日志与监控模块：

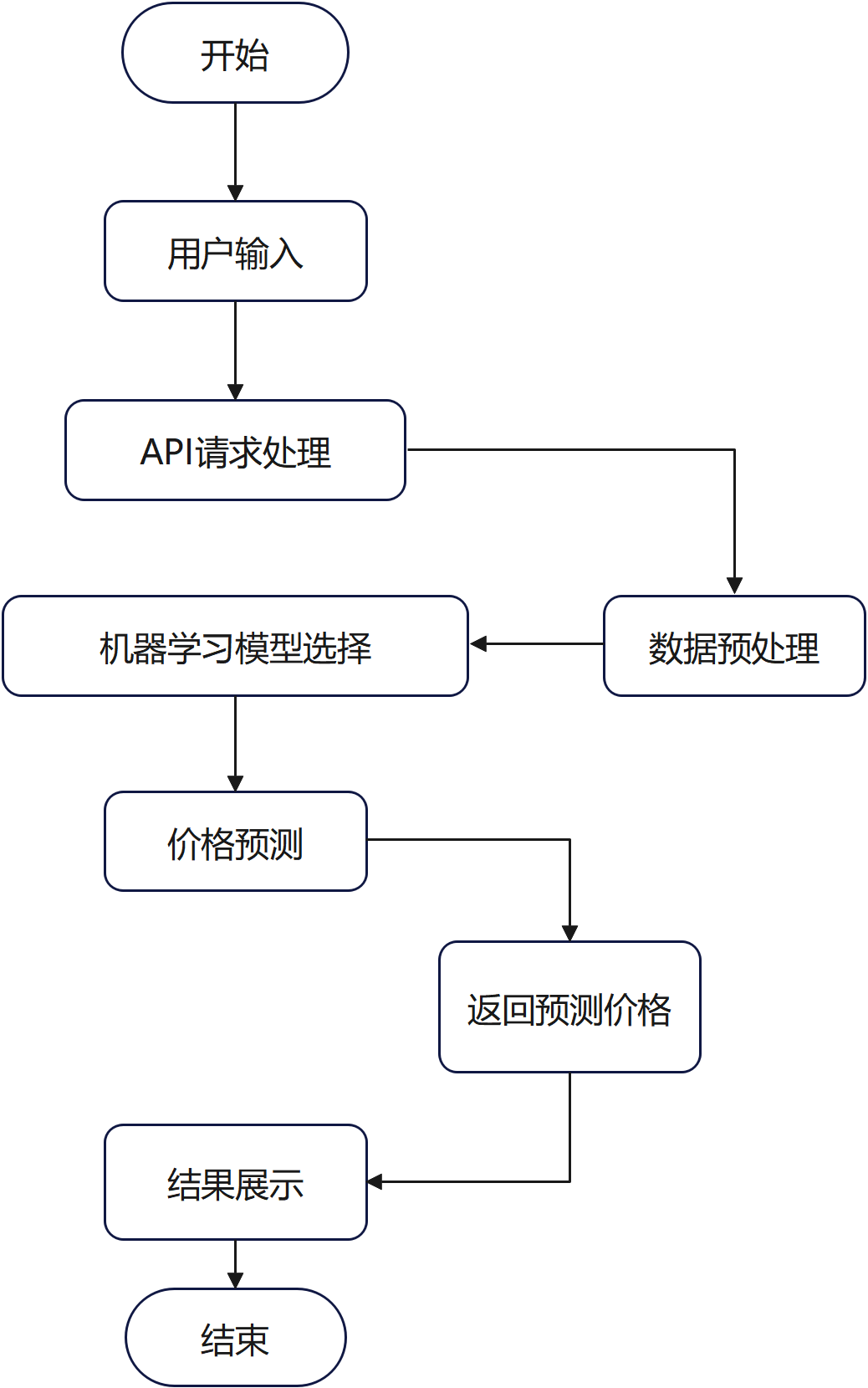
日志与监控模块用于记录系统在运行过程中的各类信息，包括用户请求、系统错误和模型运行状态等。这些信息可以帮助开发人员监控系统的运行状况，进行故障排查和系统性能优化。

8）数据流：

在整个系统中，数据从用户层开始，经过API层传输到后端逻辑层，再通过数据预处理模块清洗、标准化后输入到机器学习模型进行预测。最终的预测结果通过API返回给用户，形成一个完整的闭环。数据存储层和机器学习模型之间也有数据交互，用于存储模型参数和历史数据，支持模型的训练和更新。

通过这样的架构设计，系统能够有效地处理用户输入，利用机器学习模型进行二手车价格评估，并保证整个流程的安全性和可靠性。日志与监控模块的引入，也为系统的稳定运行和后期维护提供了有力的支持。

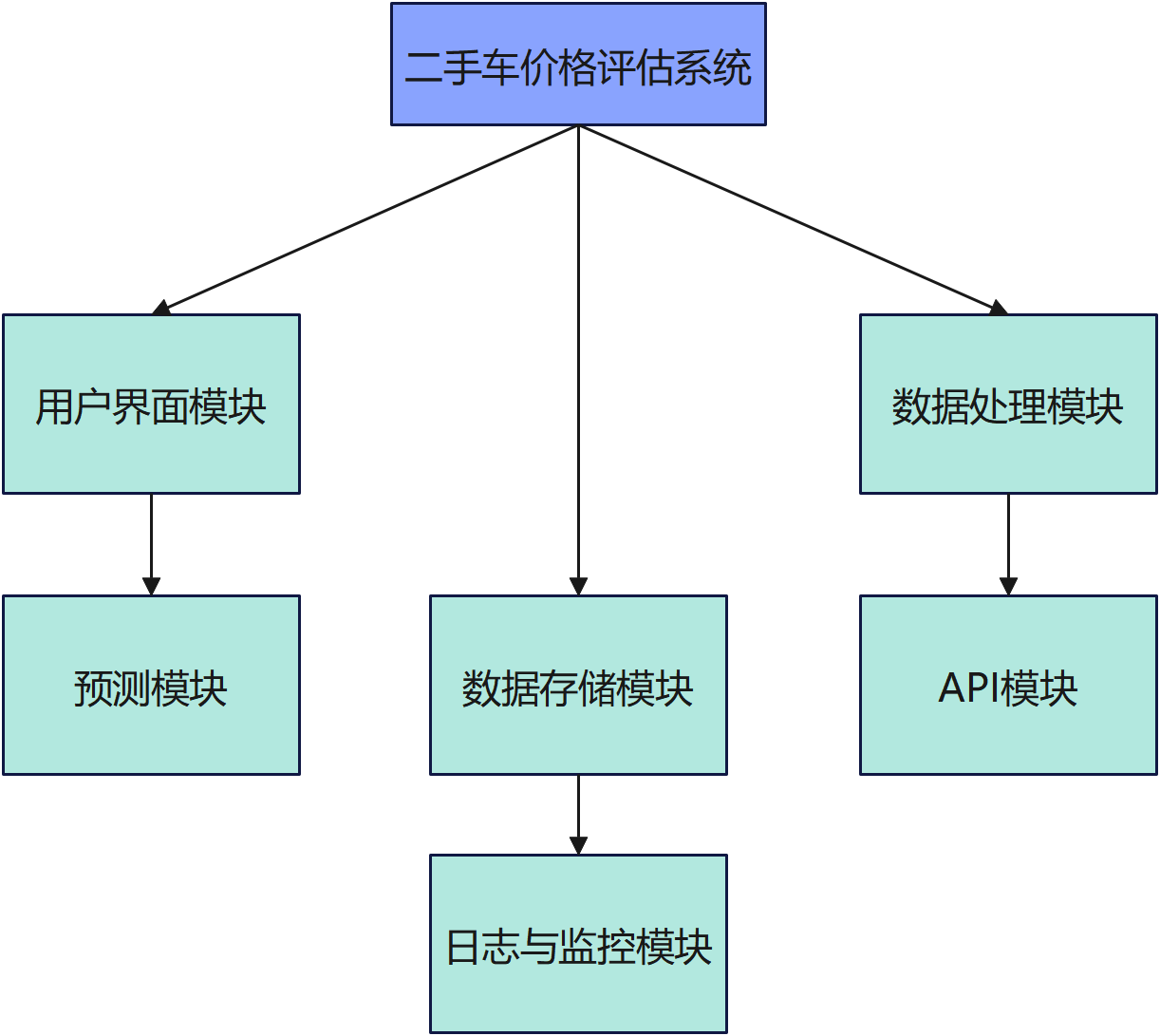
## 3.2 项目流程图



该项目流程图展示了二手车价格评估系统从用户输入到最终输出的完整流程。以下是各个步骤的详细解释：

用户首先通过系统的前端界面输入车辆的相关信息，包括品牌、型号、年份、行驶里程、车辆状况等。用户的输入是整个流程的起点，这些数据将被用于价格预测。输入的数据通过API层传递到系统后端，API负责管理用户请求与后端系统之间的通信。它确保用户输入的数据能够被后端逻辑模块正确地接收和处理。在后端接收到用户数据之后，首先对数据进行预处理。数据预处理是关键步骤，包括数据清洗（去除缺失值和异常值）、标准化（使数据具有相同量纲）和编码（将类别型数据转换为数值型数据）。这些步骤确保数据在进入模型之前具备高质量和一致性。数据预处理完成后，系统会根据之前的配置选择合适的机器学习模型。模型的选择取决于其在价格预测任务中的表现，通常会选择在训练过程中表现最优的模型（例如Lasso回归、随机森林等）。选定的机器学习模型会接收预处理后的数据，并根据输入的车辆参数进行价格预测。模型利用训练时学习到的模式，对输入数据进行推断，生成一个合理的价格区间。预测完成后，系统将生成的价格区间通过API返回给前端界面。API层的作用是将后端的结果转换为用户可以理解的格式，并将其发送给用户。最终，系统在用户界面上展示预测结果。用户可以查看该车辆的合理价格区间，从而更好地判断车辆的价值。这一结果可以帮助买家在谈判过程中有更理性的参考依据，也帮助卖家合理定价。通过以上各个步骤，整个系统形成了一个完整的闭环，从用户输入数据到输出结果，利用数据分析和机器学习技术帮助用户评估二手车的合理价格。每个步骤都相互衔接，确保系统的功能性和用户体验的顺畅性。

## 3.3 项目功能结构图



这张结构功能图展示了二手车价格评估系统的主要模块，以及各模块之间的功能和交互关系。以下是各个模块的功能结构详细解释：

### 1）二手车价格评估系统：

这是整个系统的核心，通过多个模块协同工作来处理用户的输入数据并给出二手车的价格评估。系统整合了用户界面、数据处理、预测、API、数据存储和日志监控等模块，实现了各模块的协同运作。

### 2）用户界面模块：

用户界面模块负责用户与系统之间的交互。用户可以在界面中输入车辆的参数信息（例如品牌、年份、行驶里程等），然后通过界面查看系统返回的价格评估结果。此模块注重用户体验，确保数据输入便捷，结果展示清晰明了。

### 3） 数据处理模块：

数据处理模块的作用是清理和转换用户输入的数据，以确保数据质量。在这里完成了数据清洗（处理缺失值和异常值）、数据编码（将类别数据转换为数值数据）以及数据标准化（确保特征一致性）。经过此模块处理后，数据被整理成机器学习模型可以接受的格式。

### 4）预测模块：

预测模块使用机器学习模型来对数据进行价格预测。模型根据历史数据和车辆的各项特征参数对输入数据生成合理的价格区间。该模块与数据存储模块紧密合作，从存储中调用已训练好的模型进行预测，并将预测结果传递回系统。

### 5） API模块：

API模块作为前端用户界面和后端逻辑之间的桥梁，负责数据传输和请求响应。用户的输入数据通过API模块传递到后端，API模块将处理后的预测结果再返回给用户。API模块保障了数据传输的安全性和系统响应的高效性。

### 6）数据存储模块：

数据存储模块用于存储系统数据和模型，包括历史输入、训练数据集以及模型文件。存储模块为预测模块提供支持，使系统能够利用历史数据进行持续的模型更新和优化。

### 7） 日志与监控模块：

日志与监控模块用于记录系统的运行状态和操作日志，包括用户请求、系统错误、预测过程中的模型状态等信息。它支持系统的维护和监控，确保系统稳定运行并在出现异常时提供调试依据。

### 8）模块间的交互关系：

模块之间通过箭头表示交互关系，展示了数据流动方向。各模块的数据交互如用户界面与API模块、数据处理模块与预测模块、API模块与数据存储模块等，形成了一个流畅的闭环，使得系统可以顺畅地完成从用户输入到结果输出的全过程。日志与监控模块则贯穿整个系统，实时记录和监控操作过程。

以上解释展示了系统如何通过各模块的协同作用来提供精确、可靠的二手车价格评估服务。

第4章 功能模块的分析和代码介绍

## 4.1 功能模块

在二手车价格评估系统中，每个模块都承担着特定的功能，通过模块间的紧密协作来完成整个系统的流程。以下是各个模块的详细分析，包括设计目的、功能实现、数据流程和模块间交互：

1） 用户界面模块

设计目的：用户界面模块是用户接触系统的入口，设计的目的是让用户能够轻松直观地输入车辆信息并查看系统的价格评估结果。界面需简洁友好，确保用户可以顺利完成操作而不会感到复杂或困惑。此模块还负责处理用户的反馈请求，例如修改输入参数或刷新评估结果等。

功能实现：用户界面包含输入表单、提交按钮和显示区等基本元素，便于用户输入车辆的详细信息，如品牌、型号、出厂年份、行驶里程、车辆状态等。输入验证功能，确保用户输入符合要求的格式（如年份为4位数字、里程为正整数等），避免不符合标准的数据进入系统。提交按钮将触发API请求，将数据传递到后端进行处理，并在结果区域返回预测的价格区间，用户可以快速获得信息反馈。

数据流程：数据从用户输入后通过提交按钮触发API请求，用户数据传送到系统的后端进行进一步处理。用户输入结果返回后显示在界面上，展示价格评估结果。

模块间的交互：用户界面与API模块直接相连，是用户和系统后端交互的桥梁。数据通过API模块传递到数据处理模块和预测模块，完成处理后再返回给用户界面。

2） API模块

设计目的：API模块的设计目的是实现用户界面和后端各模块的无缝连接，为用户请求提供安全、高效的传输通道，同时确保用户数据的完整性。API模块使前端用户请求能够被后端系统接受和处理，还可以对传输过程中的数据进行校验，保证数据安全。

功能实现：接收用户界面传递的数据请求，将数据从前端传递到后端的处理模块。API模块会进行基本的数据校验，并在数据不符合预期时返回错误信息以供用户调整。负责响应和处理来自后端预测结果的返回，将结果数据格式化，发送到前端进行展示。

数据流程：用户界面中的数据通过API传递到后端的数据处理和预测模块，完成预测后API模块再次将结果数据传回前端。该流程保证了前端和后端之间的数据传输流畅性，提升了用户的交互体验。

模块间的交互：API模块作为前端用户界面模块和后端系统的接口，与数据处理模块、预测模块和数据存储模块进行数据交换。

3） 数据处理模块

设计目的：数据处理模块设计的目的是确保用户输入的数据具备一致性和可靠性，从而提高模型预测的准确性和稳定性。在数据进入预测模型之前进行预处理，确保模型能够有效理解和利用输入信息。

功能实现：数据清洗：对数据进行无效值、缺失值和异常值处理，如去除空值、修复或删除不合理的数值。

数据编码：对类别型数据（如车型、颜色等）进行数值化编码，以便模型能够接受并处理。

数据标准化：对数据进行尺度统一，将数据按相同的标准进行归一化或标准化，使不同特征在模型中具备一致性。

数据流程：数据处理模块从API模块接收用户的输入数据，对其进行预处理操作后，数据将被传递到预测模块以供模型使用。

模块间的交互：数据处理模块从API模块接收数据，处理后将清理后的数据传递给预测模块。与日志模块协作，记录数据预处理中的异常值和缺失数据修复过程。

4）预测模块

设计目的：预测模块利用机器学习模型对输入的数据进行价格评估，并输出合理的价格区间。通过应用机器学习算法，系统能够生成高精度的预测结果，以便用户参考车辆的市场价格。

功能实现：机器学习模型：选择并加载训练好的模型（如随机森林、Lasso回归），根据车辆参数进行价格预测。

输出价格区间：根据模型计算结果，生成合理的价格范围，便于用户理解和决策。

数据流程：预测模块接收处理后的数据，调用模型进行价格预测，生成结果并传回API模块以供用户界面展示。

模块间的交互：与数据处理模块协作，从中接收数据。与数据存储模块交互，从存储中读取模型文件进行预测。

5）数据存储模块

设计目的：数据存储模块的设计目的是为系统提供数据存储和读取支持，包括用户的历史输入、训练数据和已训练的模型。

数据库存储：存储用户的历史记录和系统生成的数据，以便在必要时用于进一步分析或模型更新。

模型存储：保存机器学习模型的文件，支持模型的动态调用与加载。

数据流程：数据存储模块与预测模块之间进行数据和模型文件的交互，支持数据的快速读取与写入。

模块间的交互：与API模块交互接收用户输入记录，并与预测模块协作读取模型文件。

6）日志与监控模块

设计目的：日志与监控模块用于记录系统运行过程中产生的日志信息，包括用户请求、系统异常和模型状态等。

功能实现：记录用户请求的时间和数据流，便于系统运行状况的分析。监控系统的状态，及时反馈错误并发送通知。

数据流程：系统中各模块的运行状态和请求过程将被实时记录到日志中。

模块间的交互：该模块与系统中的各个模块均有交互，帮助确保系统的稳定运行和错误排查。

## 4.2 功能模块的交互与协同

1）用户界面和API模块的交互：

用户界面模块收集用户输入的车辆参数并通过API模块将数据传递给系统后端。

API模块在传输时对数据进行验证，确保所有输入均符合系统要求，然后将有效数据传递给后续的数据处理模块。用户界面与API模块的交互为系统提供了数据的入口，确保用户数据能够安全、顺利地进入系统后端进行分析和评估。

2） API与数据处理模块的协同：

API模块接收来自用户界面的数据后，直接与数据处理模块进行交互，将数据传递给该模块。数据处理模块对数据进行清洗、编码和标准化等操作，确保数据的高质量和一致性。数据处理完成后，API模块会进一步将处理好的数据传递给预测模块，形成无缝的数据流。

3）数据处理模块和预测模块的交互：

数据处理模块和预测模块的协同作用在于数据的准备和使用：数据处理模块完成数据的预处理，确保输入数据符合模型的格式和需求。预测模块从数据处理模块接收高质量的数据输入，并利用机器学习模型对这些数据进行价格预测，从而生成合理的价格区间。

4）预测模块和数据存储模块的交互：

预测模块需要调用已训练的机器学习模型，这些模型文件保存在数据存储模块中。

预测模块通过访问数据存储模块，加载所需的模型，并利用模型对用户数据进行预测。预测结果通过API返回给用户界面。数据存储模块不仅存储历史数据，还保存训练好的模型，确保预测模块能够及时调用适合的模型进行分析和预测。

5）日志与监控模块的整体监控作用：

日志与监控模块遍布系统的各个流程，实时记录每一步的操作、系统异常和数据流状态，为系统的运行提供详实的记录。当系统发生错误或数据异常时，日志与监控模块会生成相应的记录，便于后期分析和调试。监控模块还可在重大故障时向管理员发出通知，确保系统稳定运行。通过与所有模块的交互，日志与监控模块为系统提供了全面的运行数据，增强了系统的可靠性和可维护性。

每个功能模块不仅具备明确的功能定位，还通过紧密的交互与协作确保数据流畅性和系统稳定性。模块之间的依赖性和独立性设计，使得系统在处理复杂的二手车价格评估任务时，能够保持高效、准确和灵活。

## 4.3 代码分析

1）数据来源概况

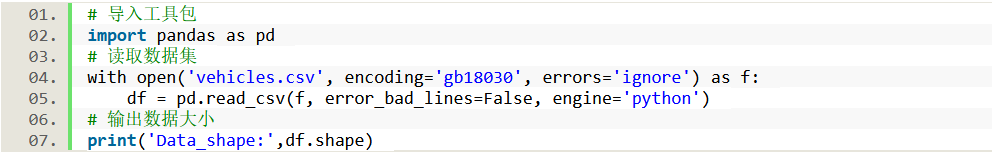
数据来源于外国某网站关于美国二手车交易记录的数据集，总数据量约为40万条，包括26列变量信息，数据集中的各字段如表1所示。

表1

|  |  |  |  |
| --- | --- | --- | --- |
| **Field** | **Description** | **Field** | **Description** |
| **id** | 用户名 | **transmission** | 自动挡/手动挡 |
| **url** | 网址 | **VIN** | 车辆识别号码 |
| **region** | 地区 | **drive** | 驱动 |
| **region\_url** | 地区网址 | **size** | 车辆大小 |
| **price** | 售价 | **type** | 车辆类型 |
| **year** | 出厂年份 | **paint\_color** | 漆色 |
| **manufacturer** | 制造商 | **image\_url** | 虚拟网址 |
| **model** | 型号 | **description** | 二手车描述 |
| **condition** | 车辆状况 | **county** | 县 |
| **cylinders** | 汽缸数目 | **state** | 州 |
| **fuel** | 燃料种类 | **lat** | 纬度 |
| **odometer** | 行驶里程数 | **long** | 经度 |
| **title\_status** | 车辆改造经历 | **posting\_date** | 挂售时间 |

2）数据概况

首先导入数据分析所用工具包，读取数据集，查看数据量大小。



输出结果：

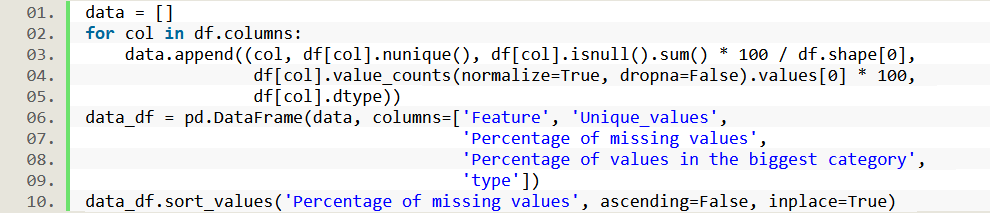
P6F3_LIUSN}Y($~`J9EL%UW

查看数据集前五行：

I63%2FZ(QI6]}T_{KO9D9CC



进一步查看数据概况：



得到数据概况如表2所示。

表2 二手车交易记录数据集数据概况

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **columns\_number** | **Feature** | **Unique\_values** | **Percentage of  missing values** | **Percentage of values in the biggest category** | **type** |
| 21 | **county** | 0 | 100 | 100 | float64 |
| 16 | **size** | 4 | 72.15472155 | 72.15472155 | object |
| 9 | **cylinders** | 8 | 40.15940159 | 40.15940159 | object |
| 8 | **condition** | 6 | 38.28938289 | 38.28938289 | object |
| 14 | **VIN** | 34773 | 36.03736037 | 36.03736037 | object |
| 15 | **drive** | 3 | 30.47430474 | 30.47430474 | object |
| 18 | **paint\_color** | 12 | 29.77629776 | 29.77629776 | object |
| 17 | **type** | 13 | 19.61119611 | 22.7402274 | object |
| 6 | **manufacturer** | 41 | 4.4200442 | 14.91914919 | object |
| 12 | **title\_status** | 6 | 2.27902279 | 95.03495035 | object |
| 11 | **odometer** | 38309 | 1.52101521 | 1.52101521 | float64 |
| 7 | **model** | 12613 | 1.28501285 | 1.4600146 | object |
| 10 | **fuel** | 5 | 0.60300603 | 82.91382914 | object |
| 23 | **lat** | 12791 | 0.53900539 | 2.03202032 | float64 |
| 24 | **long** | 12814 | 0.53900539 | 2.03202032 | float64 |
| 13 | **transmission** | 3 | 0.51400514 | 77.75577756 | object |
| 5 | **year** | 104 | 0.42800428 | 8.656086561 | float64 |
| 20 | **description** | 86284 | 0.03900039 | 0.08300083 | object |
| 19 | **image\_url** | 64480 | 0.03800038 | 1.53501535 | object |
| 25 | **posting\_date** | 94276 | 0.03800038 | 0.03800038 | object |
| 1 | **url** | 99999 | 0 | 0.00100001 | object |
| 4 | **price** | 6610 | 0 | 8.328083281 | int64 |
| 22 | **state** | 18 | 0 | 50.61450615 | object |
| 3 | **region\_url** | 84 | 0 | 2.98302983 | object |
| 2 | **region** | 84 | 0 | 2.98302983 | object |

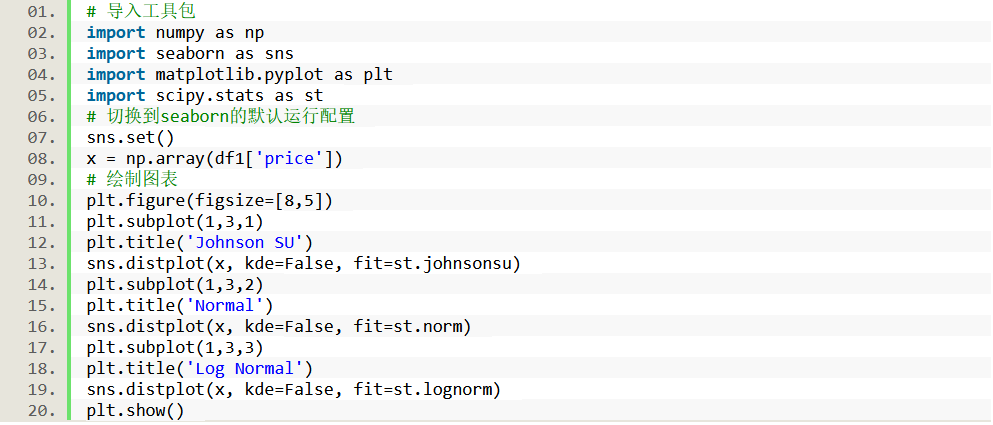
根据整体的统计信息得知，county为无效字段（信息量为空）；size、cylinders、condition、VIN、drive、paint\_color、type字段缺失量较大，在后续分析过程中要考虑是否将其纳入影响范围；大部分数据类型为object，应通过赋值将其转化成方便建立模型的int64或float64类型。

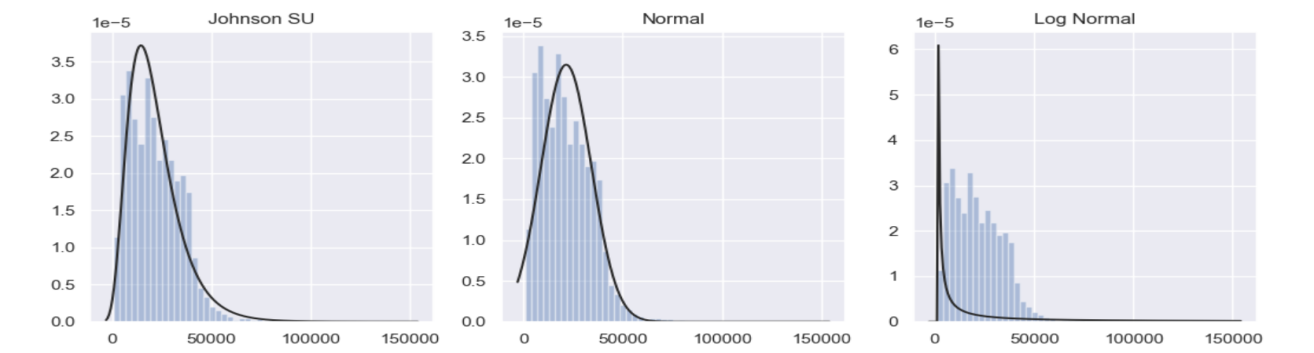
3）数据分析

在前面的数据信息统计表中发现有几个字段存在缺失值，接下来分析缺失值情况。

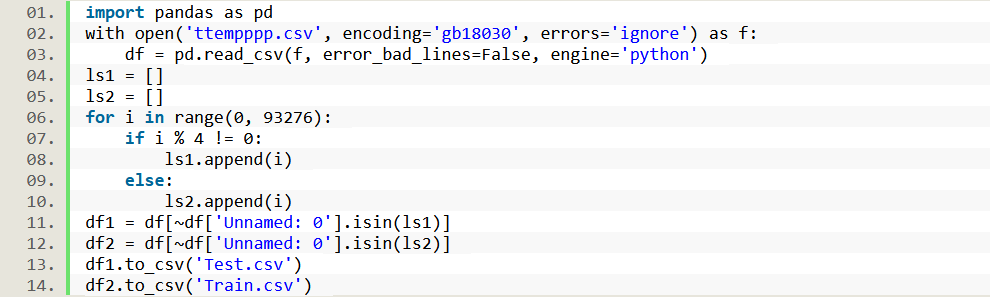


查看二手车售卖价格的大致分布：



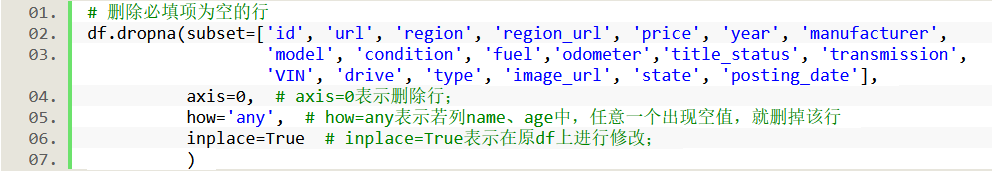


4）数据预处理

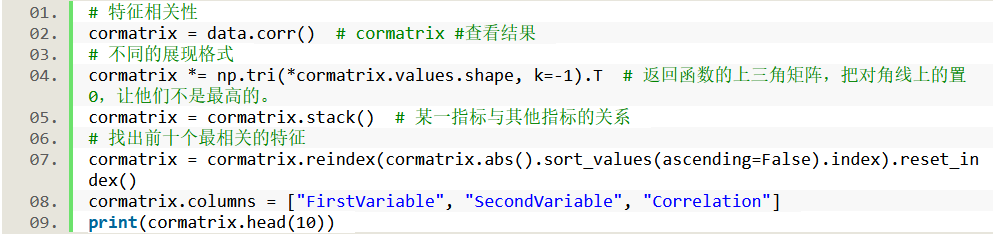


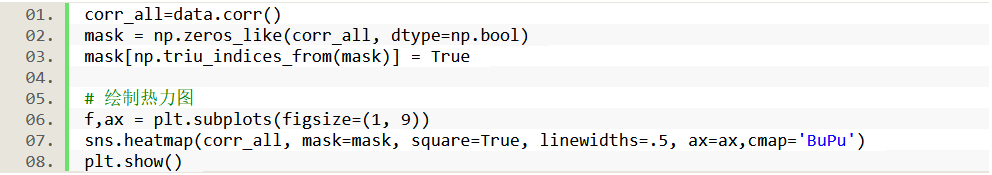
由于数据集中部分行存在缺失值，会对后续预估分析造成影响，还有部分列字段对模型作用不大（例如description字段，因为具有主观性且篇幅较大，难以将其作为车辆价格预估的参考，lat、long对本次分析无意义，county、size缺失量过大），因此将部分字段直接删除。

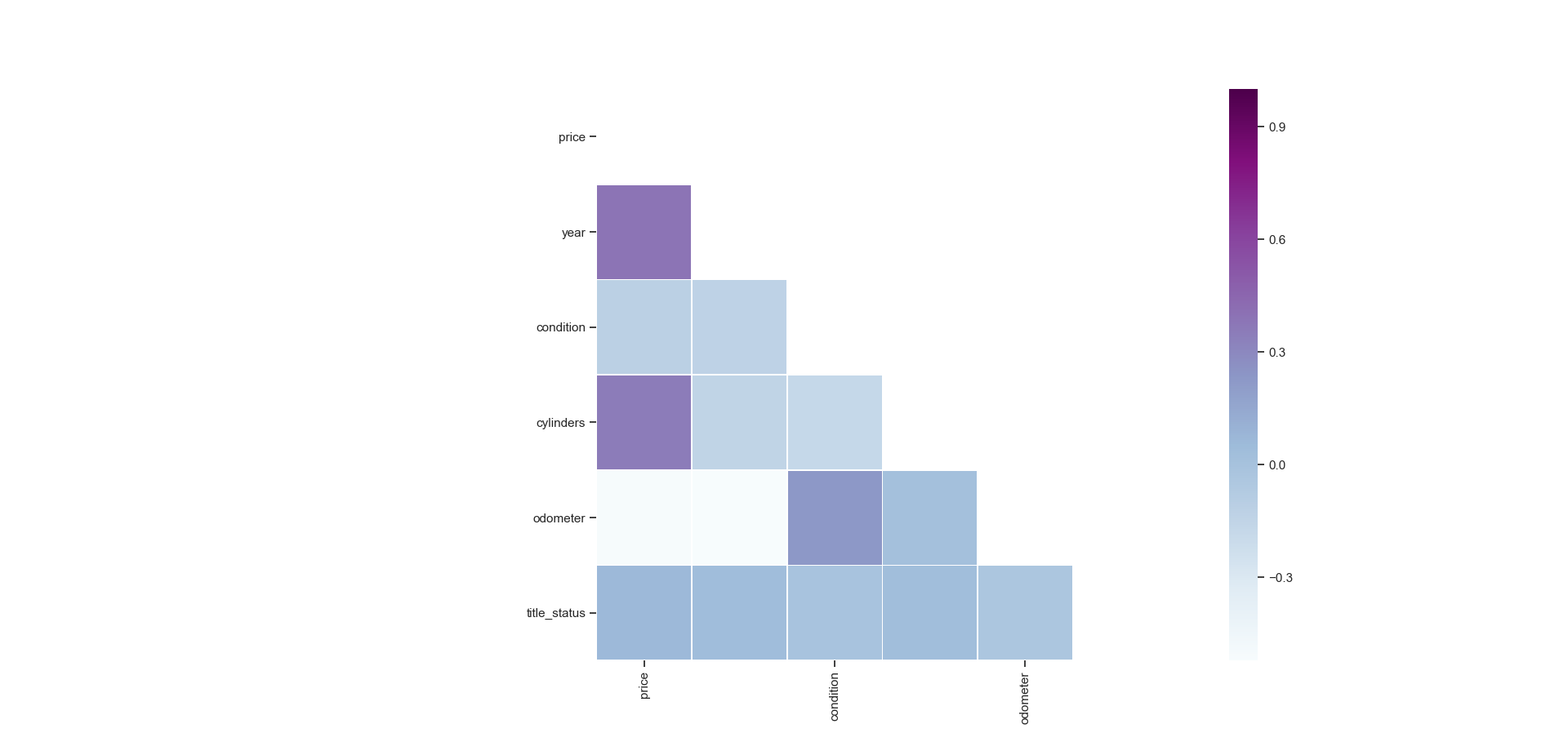
AE@QY$0N(XGC9NN$RN]NV0B



5）查看特征相关性

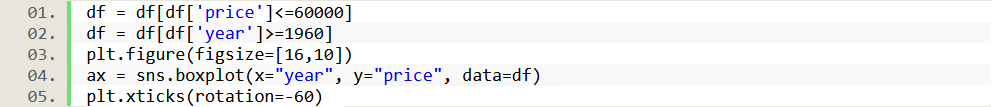


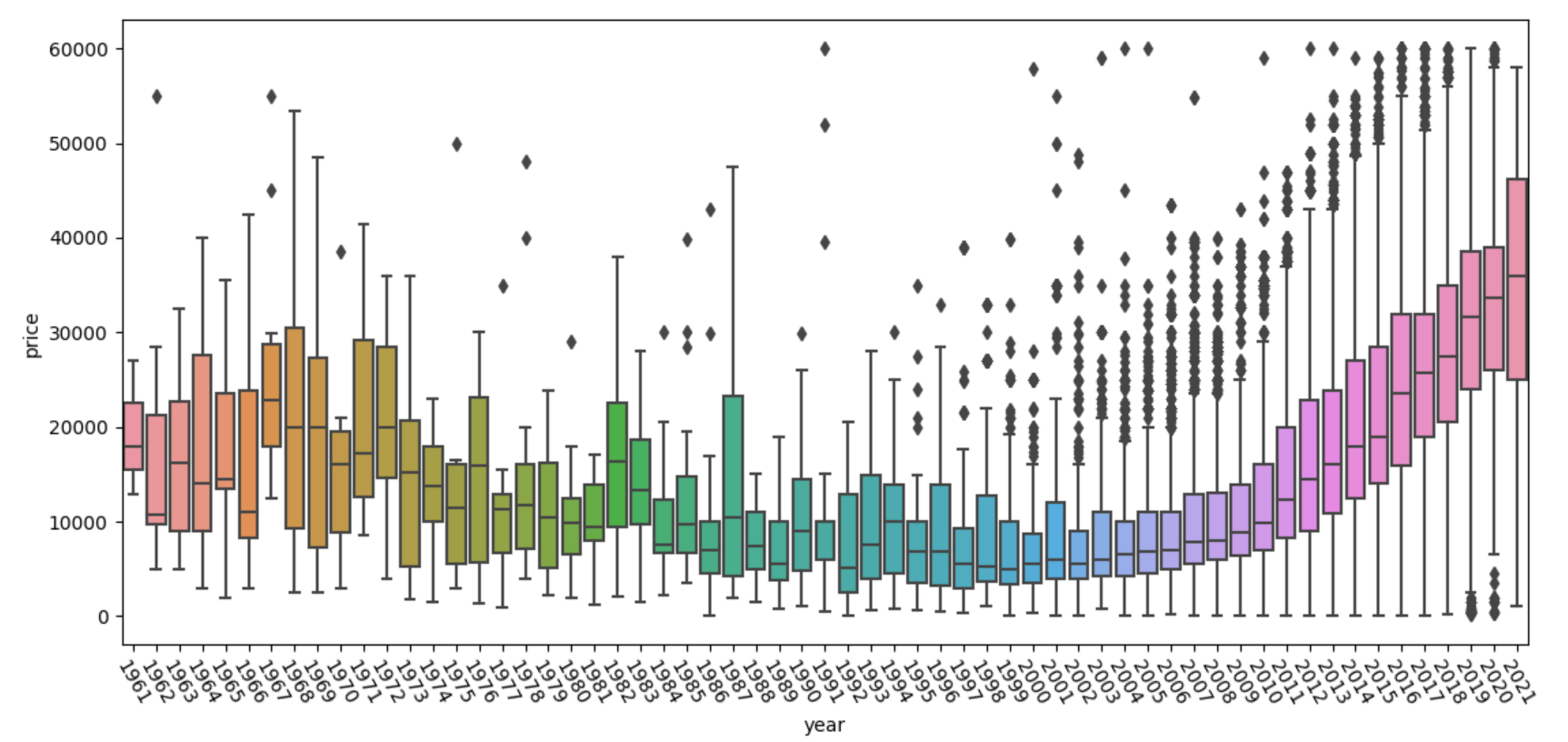




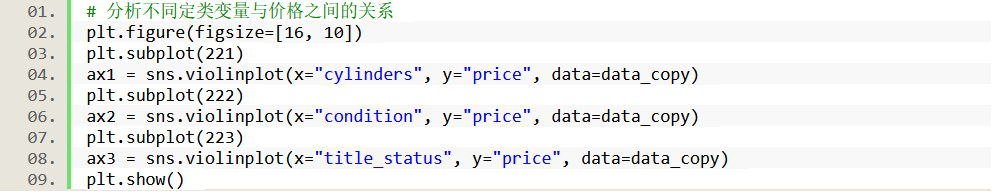
6）查看定类数据相关性

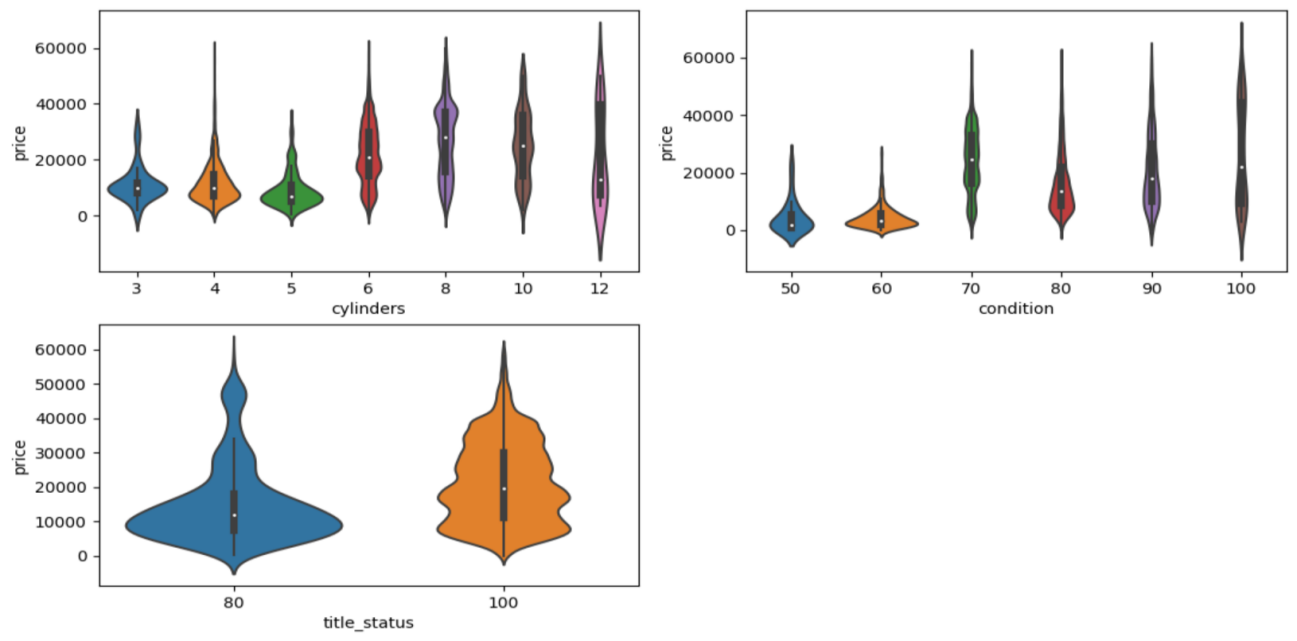
分析年份与价格的关系



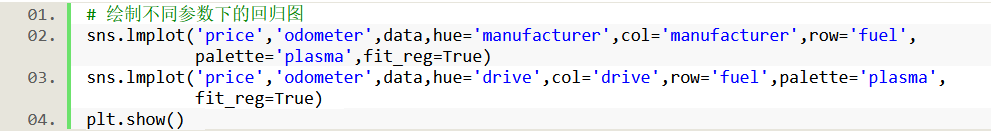


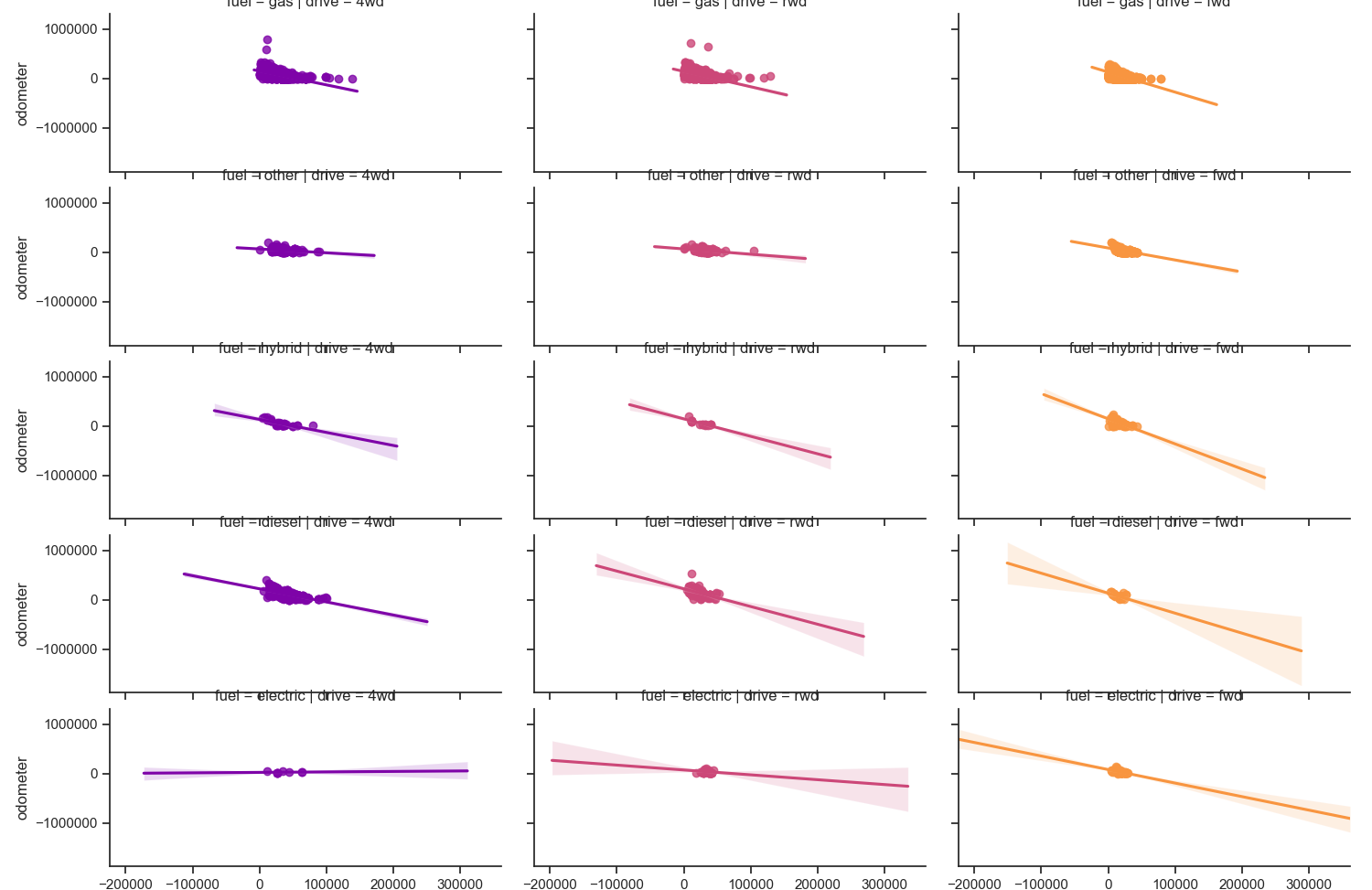
分析不同定类变量与价格之间的关系





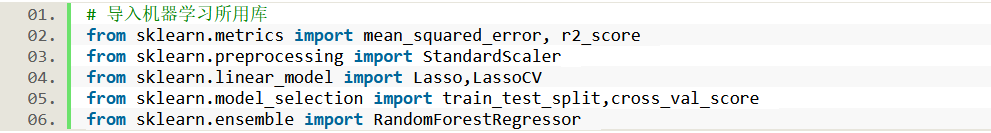
回归分析

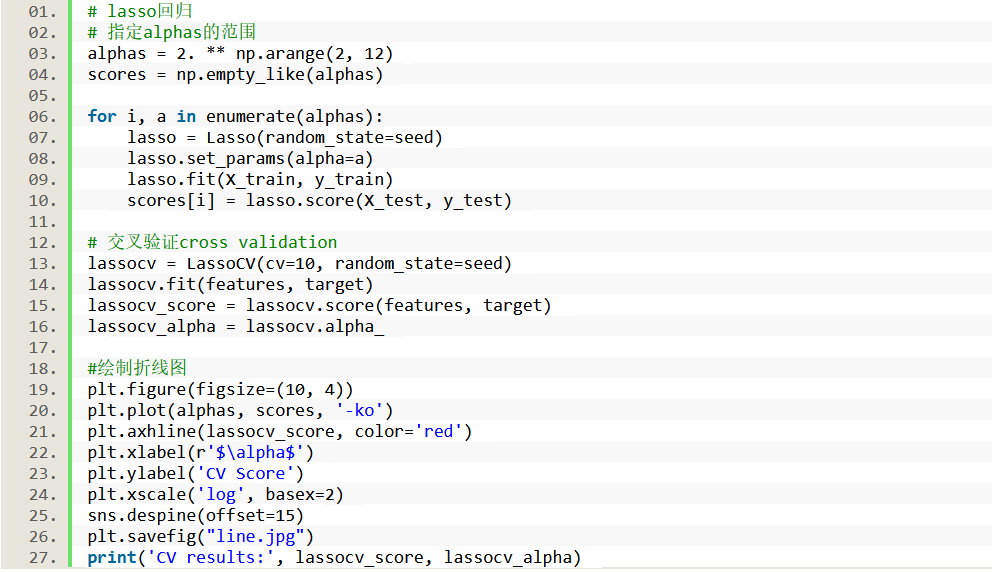


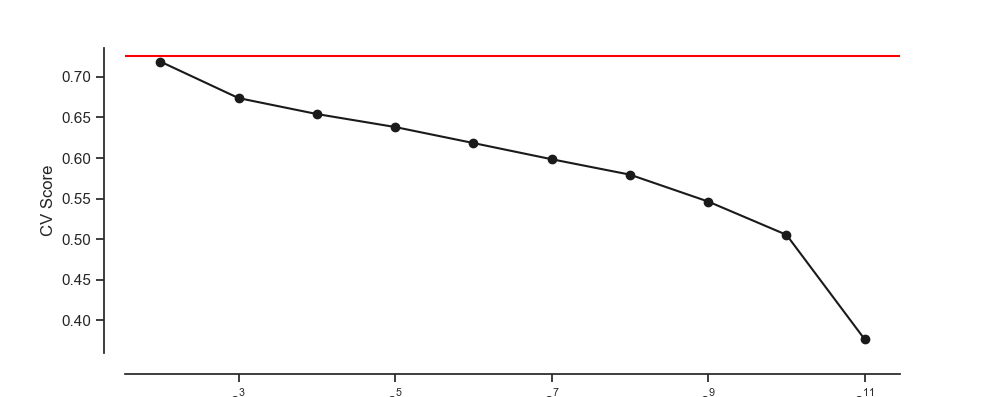


6）模型建立

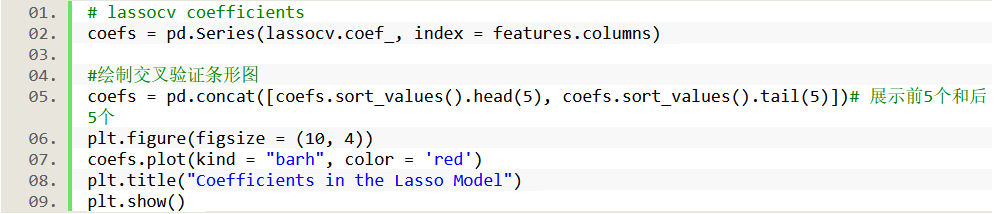
Lass回归

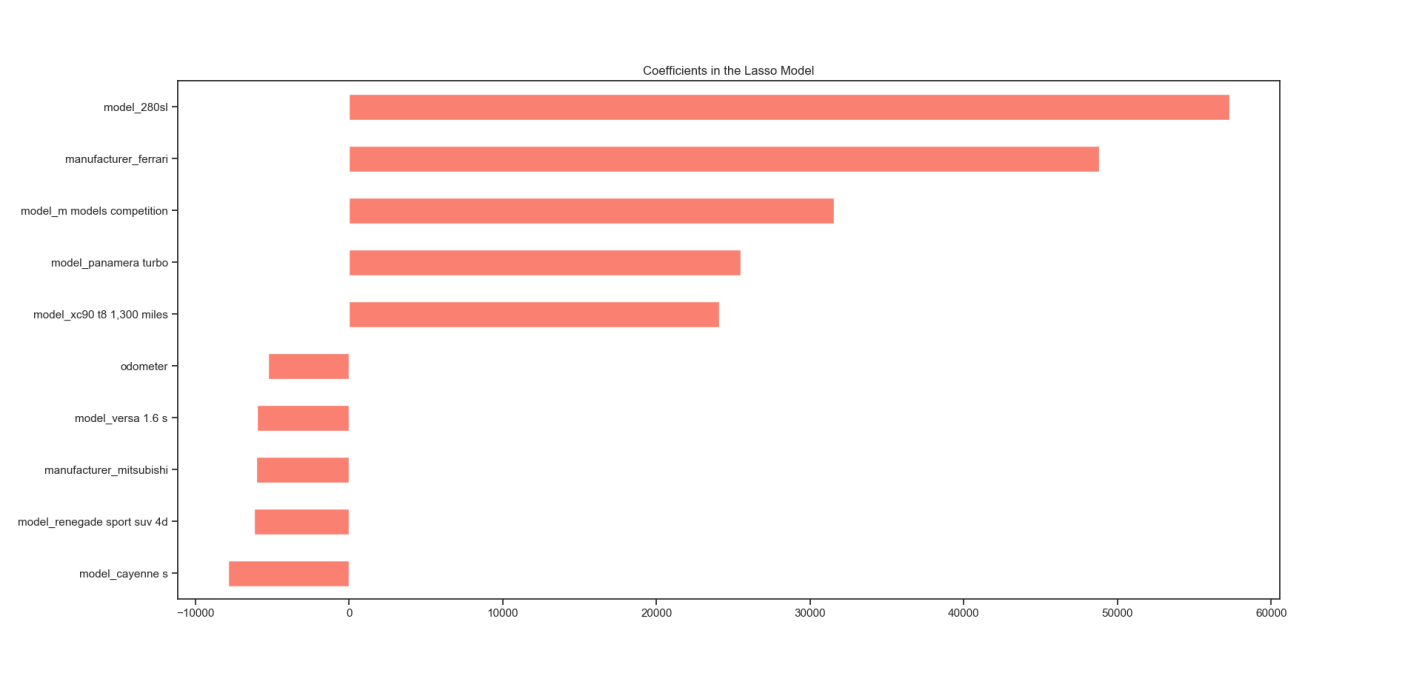


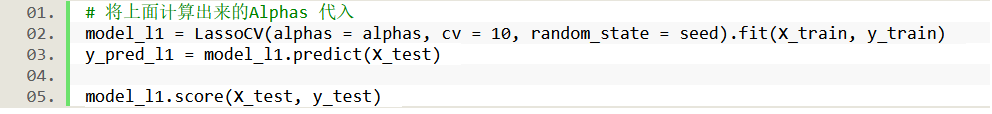




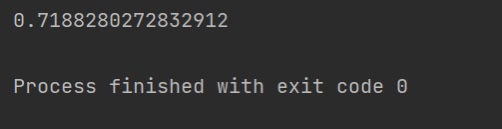
特征重要性分析





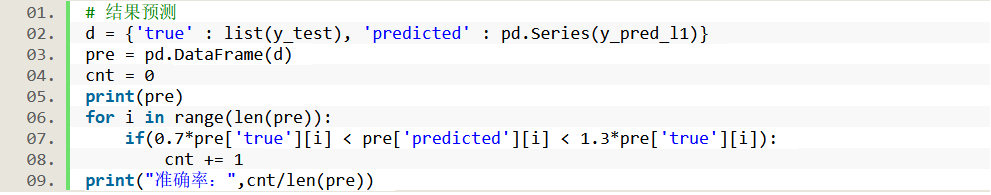


运行结果为：

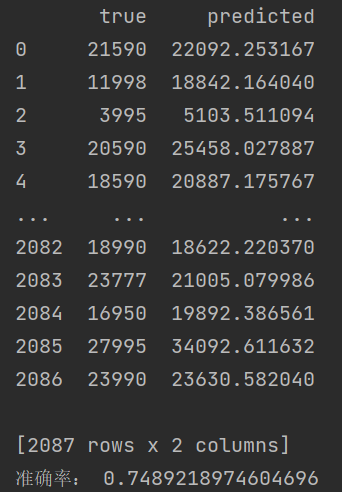


1. 软件功能演示

查看模型预测值与真实值的对比，并计算准确率



结果如下：



面向客户（买家/卖家）的窗口：

