CASE STUDY

Feature Extraction from Unstructured Data at **Edmunds.com**

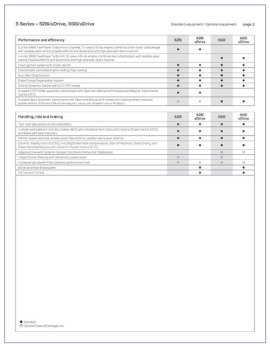
Edmunds wanted to reduce time-to-market by speeding creation of attribute data for new car models.

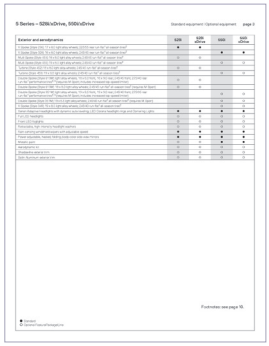
Silicon Valley Data Science designed and developed a new capability to automatically extract vehicle features from specification guides and categorize the features into appropriate vehicle classes.

Background and Business Problem

Edmunds.com is an online resource that allows users to research new and used cars including car prices, view incentives and dealer inventory listings, compare vehicles, get car buying advice and read reviews. They are headquartered in Santa Monica, CA.

Because shopping is the business focus, Edmunds needs to have real-time inventory and accurately described vehicle identification numbers (VINs).





In order to accomplish this, Edmunds relies on a large amount of third-party unstructured data, mostly from the original equipment manufacturers (OEMs). This data often arrives in the form of very long (hundreds of pages) PDFs.

Edmunds has a team of content editors who focus on creating and validating new vehicle configuration data. Core to Edmunds' competitive differentiation is the accuracy and timeliness of their product and inventory information. However, given the manual nature of moving data from unstructured PDFs into their online system, it could take up to two weeks to go from OEM data to live on the website: two valuable weeks that they were missing out on customer views while inventory was sitting around.

In order to keep pace with the rapidly growing data stream for new products, Edmunds needed to augment their existing manual processes with more automation for routine data entry and other tasks.

The Challenge

Couldn't keep pace with information from OEMs

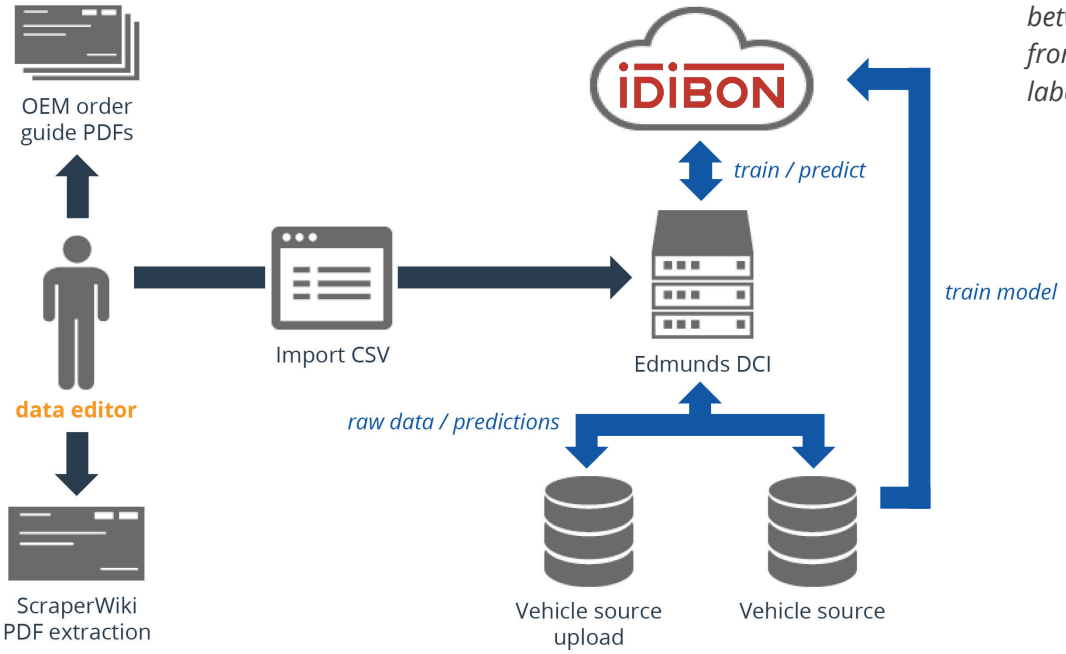
Data entry approach was largely manual, and backlogs would develop

~6.5% of VINs were being held back

Content operations team was a silo

Solution

Silicon Valley Data Science built an extraction service that reads PDF files with a complex structure—including data in tabular format—and converts them into raw text data. We then built an ontology-based attribute prediction engine using Idibon's cloud-based natural language processing services to automatically extract information about thousands of different car-related features from the raw text. This allowed us to automatically extract features and create vehicle attributes to define new vehicle models in Edmunds' database.



The structured database we built supports faceted search of models, searching available inventory, and other strategic uses. The NLP models can also be reused across other data, for mapping Edmunds' detailed ontology to a variety of unstructured data sources.

We selected the right tools and technologies for specific tasks, built the solution, validated the results, and refined the solution to meet business objectives using agile processes. The service we built for Edmunds created the largest Idibon ontology to date, supporting a hierarchical model of vehicle features and options with complex dependencies and relationships.

Our solution achieved remarkable results for Edmunds, including dramatically reducing their time to market by an order of magnitude—from two weeks to just a day or two—and allowing them to nearly eliminate their backlog.

Our Approach

Hierarchical classification was built to automatically predict group and individual attributes of new cars.

Idibon was trained using previous year's data to understand relationships between unstructured data from PDFs and associated labels.

New Capabilities

1–2 days to get a Style live online vs. 2 weeks (~85% reduction)

95% reduction in backlog

API for making predictions on unstructured vehicle data using Edmunds' ontology assets

Text classification models can be repurposed for other problems