Machine Learning Course Competition

'' Ten Problems, choose one to solve. ''

# Predicting Red Hat Business Value

## Description

Like most companies, Red Hat is able to gather a great deal of information over time about the behavior of individuals who interact with them. Theyʼre in search of better methods of using this behavioral data to predict which individuals they should approach—and even when and how to approach them.

In this competition, Kagglers are challenged to create a classification algorithm that accurately identifies which customers have the most potential business value for Red Hat based on their characteristics and activities.

With an improved prediction model in place, Red Hat will be able to more efficiently prioritize resources to generate more business and better serve their customers .

## Competition Link

[Link](https://www.kaggle.com/c/predicting-red-hat-business-value)

# Leaf Classification

## Description

There are estimated to be nearly half a million species of plant in the world. Classification of species has been historically problematic and

often results in duplicate identifications. Automating plant recognition might have many applications, including:

Species population tracking and preservation Plant-based medicinal research

Crop and food supply management



The objective of this playground competition is to use binary leaf images and extracted features, including shape, margin & texture, to accurately identify 99 species of plants. Leaves, due to their volume, prevalence, and unique characteristics, are an effective means of differentiating plant species. They also provide a fun introduction to applying techniques that involve image-based features.

As a first step, try building a classifier that uses the provided pre- extracted features. Next, try creating a set of your own features. Finally, examine the errors you're making and see what you can do to improve.

Acknowledgments

Kaggle is hosting this competition for the data science community to use

for fun and education. This dataset originates from leaf images collected by

James Cope, Thibaut Beghin, Paolo Remagnino, & Sarah Barman of the Royal Botanic Gardens, Kew, UK.

Charles Mallah, James Cope, James Orwell. Plant Leaf Classification Using Probabilistic Integration of Shape, Texture and Margin Features. Signal Processing, Pattern Recognition and Applications, in press. 2013.

We thank the UCI machine learning repository for hosting the dataset.

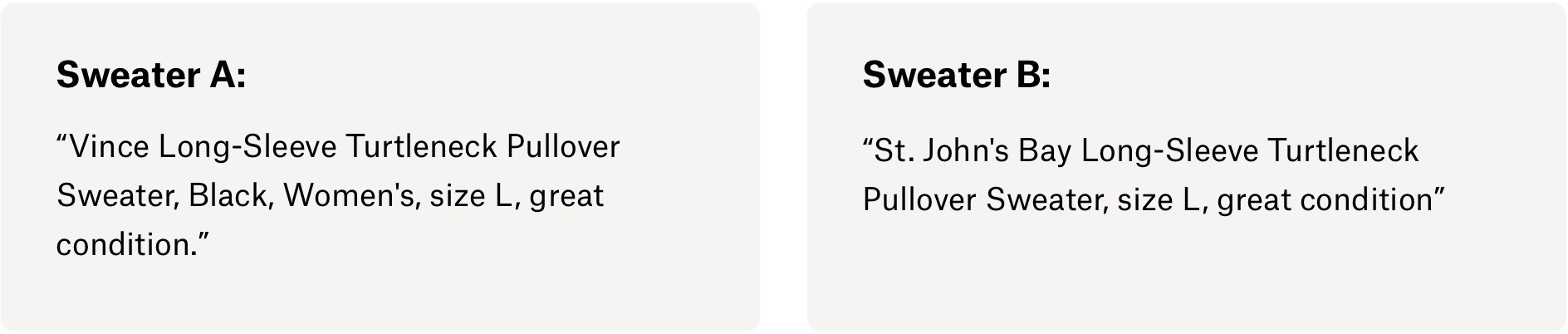
## Competition Link

[Link](https://www.kaggle.com/c/leaf-classification)

# Mercari Price Suggestion Challenge

## Description

It can be hard to know how much somethingʼs really worth. Small details can mean big differences in pricing. For example, one of these sweaters cost $335 and the other cost $9.99. Can you guess which oneʼs which?



Product pricing gets even harder at scale, considering just how many products are sold online. Clothing has strong seasonal pricing trends and is heavily influenced by brand names, while electronics have fluctuating prices based on product specs.

Mercari, Japanʼs biggest community-powered shopping app, knows this problem deeply. Theyʼd like to offer pricing suggestions to sellers, but

this is tough because their sellers are enabled to put just about anything, or any bundle of things, on Mercari's marketplace.

In this competition, Mercariʼs challenging you to build an algorithm that automatically suggests the right product prices. Youʼll be provided user- inputted text descriptions of their products, including details like product category name, brand name, and item condition.

Note that, because of the public nature of this data, this competition is a “Kernels Only” competition. In the second stage of the challenge, files will only be available through Kernels and you will not be able to modify your approach in response to new data. Read more details in the data tab and Kernels FAQ page.

## Competition Link

[Link](https://www.kaggle.com/c/mercari-price-suggestion-challenge)

# Airbnb New User Bookings

## Description

Instead of waking to overlooked "Do not disturb" signs, Airbnb travelers find themselves rising with the birds in a whimsical treehouse, having their morning coffee on the deck of a houseboat, or cooking a shared regional breakfast with their hosts.

New users on Airbnb can book a place to stay in 34,000+ cities across 190+ countries. By accurately predicting where a new user will book their first travel experience, Airbnb can share more personalized content with their community, decrease the average time to first booking, and better forecast demand.

In this recruiting competition, Airbnb challenges you to predict in which country a new user will make his or her first booking. Kagglers who impress with their answer (and an explanation of how they got there) will

be considered for an interview for the opportunity to join Airbnb's Data Science and Analytics team.



Wondering if you're a good fit? Check out this article on how Airbnb scaled data science to all sides of their organization, and visit their careers page for more on Airbnb's mission to create a world that inspires human connection.

## Competition Link

[Link](https://www.kaggle.com/c/airbnb-recruiting-new-user-bookings)

# Mercedes-Benz Greener Manufacturing

## Description

Since the first automobile, the Benz Patent Motor Car in 1886, Mercedes- Benz has stood for important automotive innovations. These include, for example, the passenger safety cell with crumple zone, the airbag and intelligent assistance systems. Mercedes-Benz applies for nearly 2000 patents per year, making the brand the European leader among premium car makers. Daimlerʼs Mercedes-Benz cars are leaders in the premium car industry. With a huge selection of features and options, customers can choose the customized Mercedes-Benz of their dreams. .

To ensure the safety and reliability of each and every unique car configuration before they hit the road, Daimlerʼs engineers have developed a robust testing system. But, optimizing the speed of their testing system for so many possible feature combinations is complex and time-consuming without a powerful algorithmic approach. As one of

the worldʼs biggest manufacturers of premium cars, safety and efficiency are paramount on Daimlerʼs production lines.



In this competition, Daimler is challenging Kagglers to tackle the curse of dimensionality and reduce the time that cars spend on the test bench. Competitors will work with a dataset representing different permutations of Mercedes-Benz car features to predict the time it takes to pass testing. Winning algorithms will contribute to speedier testing, resulting in lower carbon dioxide emissions without reducing Daimlerʼs standards.

## Competition Link

[Link](https://www.kaggle.com/c/mercedes-benz-greener-manufacturing)

# Restaurant Revenue Prediction

## Description



With over 1,200 quick service restaurants across the globe, TFI is the company behind some of the world's most well-known brands: Burger King, Sbarro, Popeyes, Usta Donerci, and Arbyʼs. They employ over 20,000 people in Europe and Asia and make significant daily investments in developing new restaurant sites.

Right now, deciding when and where to open new restaurants is largely a subjective process based on the personal judgement and experience of development teams. This subjective data is difficult to accurately extrapolate across geographies and cultures.

New restaurant sites take large investments of time and capital to get up and running. When the wrong location for a restaurant brand is chosen, the site closes within 18 months and operating losses are incurred.

Finding a mathematical model to increase the effectiveness of investments in new restaurant sites would allow TFI to invest more in other important business areas, like sustainability, innovation, and training for new employees. Using demographic, real estate, and commercial data, this competition challenges you to predict the annual restaurant sales of 100,000 regional locations.

## Competition Link

[Link](https://www.kaggle.com/c/restaurant-revenue-prediction)

# Large Scale Hierarchical Text Classification

## Description

We are pleased to announce the 4th edition of the Large Scale Hierarchical Text Classification (LSHTC) Challenge. The LSHTC Challenge is a hierarchical text classification competition, using very large datasets.

Hierarchies are becoming ever more popular for the organization of text documents, particularly on the Web. Web directories and Wikipedia are two examples of such hierarchies. Along with their widespread use comes the need for automated classification of new documents to the categories in the hierarchy. As the size of the hierarchy grows and the number of documents to be classified increases, a number of interesting machine learning problems arise. In particular, it is one of the rare situations where data sparsity remains an issue, despite the vastness of available data: as more documents become available, more classes are also added to the hierarchy, and there is a very high imbalance between the classes at different levels of the hierarchy. Additionally, the statistical dependence of the classes poses challenges and opportunities for new learning methods.

The challenge is based on a large dataset created from Wikipedia. The dataset is multi-class, multi-label and hierarchical. The number of categories is roughly 325,000 and number of the documents is 2,400,000.

This challenge builds upon a series of successful challenges on large- scale hierarchical text classification. More information can be found at<http://lshtc.iit.demokritos.gr/>

Very Large Scale Supervised Learning Track

This track concerns multi-label classification based on the Wikipedia dataset. The hierarchy is a graph that can have cycles. The number of categories is roughly 325,000 and the number of documents is 2,400,000. A document can appear in multiple classes.

## Competition Link

[Link](https://www.kaggle.com/c/lshtc)

# NOAA Fisheries Steller Sea Lion Population Count

## Description

Steller sea lions in the western Aleutian Islands have declined 94 percent in the last 30 years. The endangered western population, found in the North Pacific, are the focus of conservation efforts which require annual population counts. Specially trained scientists at NOAA Fisheries Alaska Fisheries Science Center conduct these surveys using airplanes and unoccupied aircraft systems to collect aerial images. Having accurate population estimates enables us to better understand factors that may be contributing to lack of recovery of Stellers in this area.

Currently, it takes biologists up to four months to count sea lions from the thousands of images NOAA Fisheries collects each year. Once individual counts are conducted, the tallies must be reconciled to confirm their reliability. The results of these counts are time-sensitive.

In this competition, Kagglers are invited to develop algorithms which accurately count the number of sea lions in aerial photographs.

Automating the annual population count will free up critical resources allowing NOAA Fisheries to focus on ensuring we hear the sea lionʼs roar for many years to come. Plus, advancements in computer vision applied to aerial population counts may also greatly benefit other endangered species.

## Competition Link

[Link](https://www.kaggle.com/c/noaa-fisheries-steller-sea-lion-population-count)

# Statoil/C-CORE Iceberg Classifier Challenge

## Description

Drifting icebergs present threats to navigation and activities in areas such as offshore of the East Coast of Canada.

Currently, many institutions and companies use aerial reconnaissance and shore-based support to monitor environmental conditions and assess risks from icebergs. However, in remote areas with particularly harsh weather, these methods are not feasible, and the only viable monitoring option is via satellite.

Statoil, an international energy company operating worldwide, has worked closely with companies like C-CORE. C-CORE have been using satellite data for over 30 years and have built a computer vision based surveillance system. To keep operations safe and efficient, Statoil is interested in getting a fresh new perspective on how to use machine learning to more accurately detect and discriminate against threatening icebergs as early as possible.

In this competition, youʼre challenged to build an algorithm that automatically identifies if a remotely sensed target is a ship or iceberg. Improvements made will help drive the costs down for maintaining safe working conditions.

## Competition Link

[Link](https://www.kaggle.com/c/statoil-iceberg-classifier-challenge)

# Zillow Prize: Zillowʼs Home Value Prediction (Zestimate)

## Description

Zillowʼs Zestimate home valuation has shaken up the U.S. real estate industry since first released 11 years ago.

A home is often the largest and most expensive purchase a person makes in his or her lifetime. Ensuring homeowners have a trusted way to monitor this asset is incredibly important. The Zestimate was created to give consumers as much information as possible about homes and the housing market, marking the first time consumers had access to this type of home value information at no cost.

“Zestimates” are estimated home values based on 7.5 million statistical and machine learning models that analyze hundreds of data points on each property. And, by continually improving the median margin of error (from 14% at the onset to 5% today), Zillow has since become established as one of the largest, most trusted marketplaces for real estate information in the U.S. and a leading example of impactful machine learning.

Zillow Prize, a competition with a one million dollar grand prize, is challenging the data science community to help push the accuracy of the Zestimate even further. Winning algorithms stand to impact the home values of 110M homes across the U.S.

In this million-dollar competition, participants will develop an algorithm that makes predictions about the future sale prices of homes. The contest is structured into two rounds, the qualifying round which opens May 24, 2017 and the private round for the 100 top qualifying teams that opens on Feb 1st, 2018. In the qualifying round, youʼll be building a model to improve the Zestimate residual error. In the final round, youʼll build a home valuation algorithm from the ground up, using external data sources to help engineer new features that give your model an edge over the competition.

Because real estate transaction data is public information, there will be a three-month sales tracking period after each competition round closes where your predictions will be evaluated against the actual sale prices of the homes. The final leaderboard wonʼt be revealed until the close of the sales tracking period.

## Competition Link

[Link](https://www.kaggle.com/c/zillow-prize-1)